frontiers in
EVOLUTIONARY NEUROSCIENCE

REVIEW ARTICLE
published: 13 April 2012
doi: 10.3389/fnevo.2012.00005

A bird's eye view of human language evolution

Robert C. Berwick[1,2]*, Gabriël J. L. Beckers[3], Kazuo Okanoya[4] and Johan J. Bolhuis[5]

[1] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA
[2] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
[3] Department of Behavioural Neurobiology, Max Planck Institute for Ornithology, Seewiesen, Germany
[4] Department of Cognitive and Behavioral Sciences, The University of Tokyo, Tokyo, Japan
[5] Behavioural Biology, Helmholtz Institute, University of Utrecht, Utrecht, The Netherlands

Edited by:
Angela Dorkas Friederici, Max Planck Institute for Human Cognitive and Brain Sciences, Germany

Reviewed by:
Christopher I. Petkov, Newcastle University, UK
W. Tecumseh Fitch, University of Vienna, Austria

***Correspondence:**
Robert C. Berwick, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 32D-728, 77 Massachusetts Avenue, Cambridge, MA 02139, USA; Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 32D-728, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.
e-mail: berwick@csail.mit.edu

Comparative studies of linguistic faculties in animals pose an evolutionary paradox: language involves certain perceptual and motor abilities, but it is not clear that this serves as more than an input–output channel for the externalization of language proper. Strikingly, the capability for auditory–vocal learning is not shared with our closest relatives, the apes, but is present in such remotely related groups as songbirds and marine mammals. There is increasing evidence for behavioral, neural, and genetic similarities between speech acquisition and birdsong learning. At the same time, researchers have applied formal linguistic analysis to the vocalizations of both primates and songbirds. What have all these studies taught us about the evolution of language? Is the comparative study of an apparently species-specific trait like language feasible? We argue that comparative analysis remains an important method for the evolutionary reconstruction and causal analysis of the mechanisms underlying language. On the one hand, common descent has been important in the evolution of the brain, such that avian and mammalian brains may be largely homologous, particularly in the case of brain regions involved in auditory perception, vocalization, and auditory memory. On the other hand, there has been convergent evolution of the capacity for auditory–vocal learning, and possibly for structuring of external vocalizations, such that apes lack the abilities that are shared between songbirds and humans. However, significant limitations to this comparative analysis remain. While all birdsong may be classified in terms of a particularly simple kind of concatenation system, the regular languages, there is no compelling evidence to date that birdsong matches the characteristic syntactic complexity of human language, arising from the composition of smaller forms like words and phrases into larger ones.

Keywords: birdsong, brain evolution, phonological syntax, speech

INTRODUCTION: BIRDSONG AND HUMAN LANGUAGE PERSPECTIVES

Over 2000 years ago, Aristotle in his *Historia Animalium* (Aristotle, 1984, c. 350 BCE) had already noted many striking parallels between birdsong and human speech – in remarkably modern terminology, he observed that some songbirds, like children, acquire sophisticated, patterned vocalizations, "articulated voice," sometimes learned, and sometimes not: "second only to man, some species of birds utter articulate phonemes"; and "some of the small birds do not utter the same voice as their parents when they sing, if they are reared away from home and hear other birds singing. A nightingale has already been observed teaching its chick, suggesting that [birdsong] ... is receptive to training" (*Hist. Anim.* 504a35–504b3; 536b, 14–20). In this passage, Aristotle uses the Greek word *dialektos* to refer to birdsong variation, paralleling the term he reserves for human speech, and anticipating even the most recent work on how the songs of isolated juvenile vocal learning finches might "drift" from that of their parents over successive generations (Feher et al., 2009). Given two millennia of research from neuroscience to genomics, our insights regarding the parallels between birdsong and human language have advanced since Aristotle's day. But how much have we learned? What can birdsong tell us today about the structure and evolution of human language?

In this article we consider this question from the perspective of modern linguistic theory, focusing on the connections between human language sound systems and syntax as compared to those of birdsong. We will maintain that while there are many striking parallels between speech and vocal production and learning in birds and humans, with both requiring similar, limited computational machinery, the same does not appear to hold when one compares language syntax and birdsong more generally. While there are many points at which birdsong and human syntax differ, summarized below in **Table 1** for reference, we highlight two here that seem especially prominent, returning to details and justification for this contrast in Section "Building Blocks for Human Language" below. First, human language syntax, but not birdsong, is organized into "chunks" – phrases – that are labeled by features of the elements from which the chunks are constructed (**Table 1**, row 7). For example, the word sequence *ate the starlings* has "verb-like" properties, inherited from the verb *ate*. In contrast, even though certain birdsong syllable sequences can be described as "chunks" (Suge and Okanoya, 2010), these do not have the properties of the

Table 1 | The major comparisons between birdsong syntactic structure and human syntactic structure.

	Birdsong	Human language syntax
Precedence-based dependencies (1st order Markov)	Yes	Yes, but in sound system only
Adjacency-based dependencies	Yes	Yes
Non-adjacent dependencies	In some cases	Yes
Unbounded non-adjacent dependencies	Not known	Yes
Describable by (restricted) finite-state transition network	Yes (*k*-reversible)	No
Grouping: elements combined into "chunks" (phrases)	Yes	Yes
Phrases "labeled" by element features	No	Yes (words)
Hierarchical phrases	Limited (in some species)	Yes, unlimited
Asymmetrical hierarchical phrases	No	Yes
Hierarchical self-embedding of phrases of the same type	No	Yes
Hierarchical embedding of phrases of different types	No	Yes
Phonologically null chunks	No	Yes
Displacement of phrases	No	Yes
Duality of phrase interpretation	No	Yes
Crossed-serial dependencies	No	Yes
Productive link to "concepts"	No	Yes

Most human language syntactic properties are not found in birdsong. The only exceptions relate to the properties of human language sound systems.

syllables out of which they are built; for example, the (hypothetical) chunk *warble-twitter* does not have the properties of either of the two syllables from which it is composed. Second, human language phrases are generally *asymmetrically hierarchical* (**Table 1**, row 9): the phrase *ate the starlings* is divisible into a small portion, the verb *ate*, and then a much larger portion, *the starlings*, which the larger portion might in turn contain further elaboration, as in *ate the starlings that sat on the wire*. Nothing like this syntactic complexity seems evident in birdsong.

Marler (1998) has advanced a very similar view in his contrast of "phonological syntax" or *phonocoding*, as opposed to "lexical syntax" or *lexicoding*. On Marler's account, songbirds exhibit only phonological syntax, that is, the stringing together of elements, sounds, according to some well-defined pattern, but without the meaning of the resulting sequence as a whole dependent on the meaning of its individual parts. In contrast, Marler argues that only human languages exhibit lexical syntax, that is, changes in meaning resulting from different combinations elements such as word parts, words, or phrases – *starling* means something different from *starlings*. Put another way, Marler notes that while both birdsong and human language are combinatorial, in the sense that they both assemble larger structures out of more basic parts, only human language is compositional, in the sense that the meaning of a word or sentence changes as we change its component parts.

In this article we have used Marler's distinction as the springboard for a more nuanced review of the differences between birdsong and human language, one that focuses on both details about computation and representation. From the standpoint of computation, the difference between birdsong and human language syntax has often been cast as a single, sharp formal difference in the computational machinery available to humans as opposed to birds (and other non-human species): all birdsongs can be described in terms what are technically called *regular languages* – languages that can be generated by a particularly simple kind of computational device called a finite-state automaton, while human languages are *non-regular* and fall outside this class, describable only by using more powerful computational devices. The distinction between regular and non-regular language is familiarly known as part of the *Chomsky hierarchy* (Chomsky, 1956), one formal way of partitioning the complexity of languages when viewed as a set of strings. However, we find that while the regular/non-regular distinction captures some of the differences between birdsong and human language, it is both too weak and too strong. As we describe in Section "Human Language and Birdsong: The Key Differences" below, this distinction is too weak, because it appears that all birdsong can be described by a far narrower class of regular languages, that turn out to be easily learned from examples, an important point if birdsong is to be learned from adult male tutors (Berwick et al., 2011a). But this distinction is also too strong, in the sense that several aspects of human language, such as the assignment of stress to words, or the way that prefixes or suffixes are assembled to form words, can be described by finite-state automata, while other aspects of human language seemingly go beyond the computational augmentations used to divide the regular from the non-regular languages (see, e.g., Huybregts, 1984).

In brief then, we find that from a computational perspective, the traditional Chomsky hierarchy does not draw the proper "bright line" separating human language from birdsong. (See Barton et al., 1987 for another view on the inadequacy of this hierarchy as a way to categorize human language.) Rather than impose an *a priori* classification on an inherently biological system such as language, drawn from the analysis of formal languages, the approach taken here turns the traditional classification on its head: we first attempt to characterize as best we can the minimally necessary computational components that empirically underpin language. Given this, we then characterize what class of sentences and structures this delimits. As **Table 1** indicates, human language must be analyzed at a finer grain than simply the regular/non-regular distinction.

Similarly, from a representational point of view, our characterization of how human language and birdsong differ in terms of asymmetrical, hierarchically arranged phrases does not fit neatly into any of the conventional categories fixed by the regular and non-regular languages, which do not typically address the question of what *structures* are assigned to particular strings. For example, as we shall see, the asymmetrical hierarchical structure associated with *ate the starlings* could just as readily be associated with a regular language as with a non-regular language. Finally, Marler's notion that it is "lexicoding" – words – that completely characterizes the division between human language and birdsong captures part, but not all, of the necessary distinctions. It does not account for the inherent asymmetry of human language structure, and falls short when it comes to describing human language structures that have no associated lexical meanings, such as the metrical or prosodic structure associated with human language.

Figure 1 sets out the gist of our account using a simple example sentence, where we have deliberately simplified linguistic details for expository purposes. **Figure 1A** displays the syntactic structure conventionally associated with the sentence *the rat chased the birds*. It exhibits two prominent properties. First, the representation is *hierarchical*. The sentence divides into two parts: on the left, the portion corresponding to *the rat*, ordinarily called a *Noun Phrase* (NP); and on the left, the portion corresponding to *chased the bird*, ordinarily called a *Verb Phrase* (VP). The VP itself then subdivides into two further parts, a verb *chased* on the left and a second Noun Phrase, *the birds*, on the right. Thus the first NP lies at one level *above* the second NP. This grouping of the verb and the second Noun Phrase together into a single unit, what linguists call a *phrase*, is not arbitrary. This analysis has been confirmed empirically for over a century, using established structuralist and generative linguistic techniques (see, e.g., Bloomfield, 1933; Chomsky, 1955; Jackendoff, 1977). For example, it is straightforward to show that the second Noun Phrase, *the birds*, conventionally called the Object, is bound together with the Verb as a single "chunk" or phrase, because the Verb plus its Object can be seen to be subject to syntactic rules that manipulate them single units, in the same sense that we identify particular combinations of atoms as specific molecules because they act identically in particular chemical reactions. Linguists have devised many standard tests to demonstrate the existence of such "chemical compounds" in language; we illustrate one of several here. Consider the example sentence (1a) below. Linguists note that the sequence *ate the birds* forms a single phrase, a verb phrase, because, as shown in (1b), one can remove the second occurrence of *ate the birds* in its entirety, substituting the word *did*, but retain the same meaning as in (1a), viz., that both the rat and the cat ate the birds. In contrast, if we delete any *part* of the "compound" Verb-plus-Object, and try to apply the same syntactic operation – the same "chemical reaction" – the result seems ill-formed, as evidenced by (1c):

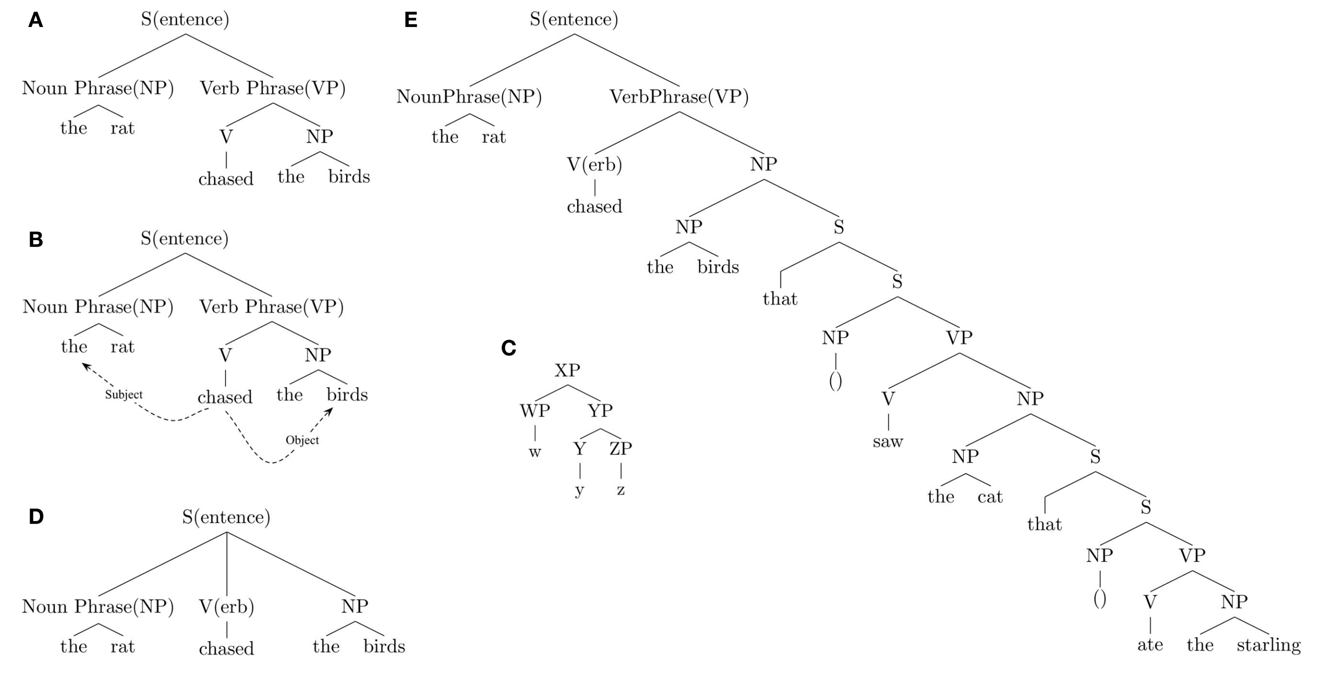

FIGURE 1 | The hierarchical and asymmetrical nature of sentence syntax. (A) The conventional syntactic structure associated with a simple English sentence. Note that the structure is asymmetrical, as highlighted in part **(C)** below. **(B)** The grammatical relationships "Subject" and "Object" are defined solely with respect to the hierarchical properties of the syntactic representation. **(C)** Abstract phrase structure representation for the sentence, highlighting the core asymmetry such structures, along with the "grounding" of a phrase of type *Y*, *YP*, on a particular word of type *Y*. Such structures comprise the core "molecular shapes" of syntax in language. **(D)** A symmetric syntactic structure associated with the same sentence, pointing out that there is no longer a distinction between "Subject" and "Object." **(E)** A more complex syntactic structure associated with a sentence that displays hierarchical, self-embedded containment relationships, with Sentences embedded within other Sentences, and Noun Phrases within Noun Phrases. Note that the basic asymmetric structure of **(C)** is replicated at each level.

(1a) *the rat ate the birds and the cat ate the birds*
(1b) *the rat ate the birds and the cats did too*
(1c) *?? the rat ate the birds and the cats did the birds too*
(1d) *the birds, the rat ate*

In this way, sensitivity to syntactic rules demonstrates that the Verb-plus-Object can be manipulated as if it were a single entity. Similarly, the sentence's Object, *the birds*, is itself a single unit, so it too can be manipulated as if it were a single syntactic "molecule": we can displace it to the front of a sentence, as in (1d). What about a hypothetical "compound" that would be formed by conjoining *the rat*, the so-called the Subject of a sentence, with the Verb, forming the unitary "molecule" *the rat ate*? Such a hypothetical unit is *never* observed to enter into distinguished syntactic operations – it is a "compound" that evidently does not participate in distinctive syntactic "chemical reactions." We may therefore conclude, along with the majority of linguists, that the "grouping" structure of words in English sentences like these may be portrayed in something like the form, Subject–Verb Phrase, where the Verb Phrase in turn is divided into a Verb plus its Object (if any). Because the Object itself forms a group, one is thereby licensed to represent the syntactic form of the entire word sequence as something like, (*the rat*) (*chased* (*the cat*)), where the Subject phrase is placed apart from the rest of the syntactic structure in the sentence, asymmetrically. It should be stressed that examples such as (1a–c) have also received confirmation from domains other than linguistic analysis, in this case, from psycholinguistic studies indicating that complete Verb Phrases, i.e., Verb–Object combinations, are "recycled" in human sentence processing, while there is no comparable evidence for this with respect to Subject–Verb combinations; see, e.g., Arregui et al. (2006), Mauner et al. (1995). For additional book-length treatment of the key role of asymmetric relations in language, see Kayne (1994), Moro (2000), Di Sciullo (2003).

In brief, language's syntactic structure is fundamentally asymmetric. **Figure 1A** illustrates this asymmetry graphically: the first NP, corresponding to *the rat*, lies off to the left side of the rest of the sentence, which is subsumed by the Verb Phrase. This fundamental asymmetry, cast in terms of a tree-structured representation as shown in **Figures 1A,B**, is central to how sentence structure drives sentence interpretation. The first NP directly dominated by the entire Sentence fixes what is the Subject, and this NP is typically, but not always, the "agent" of the action corresponding to the Verb. In contrast, the NP dominated by the VP and adjacent to the verb determines what is the Object, and this NP is typically the "affected object" of an action (Chomsky, 1965).

Importantly, such syntactic relationships do not depend on the temporal ordering of a sentence's words – the left-to-right way the words are orthographically transcribed, corresponding to their spoken (or manually signed) order. Rather, a considerable body of converging evidence, from linguistic, psycholinguistic, and more recently brain-imaging studies, has accumulated showing that this necessarily "linear" format is mapped to an internal representation that respects only hierarchical structure (see, e.g., Moro, 2008, 2011; for recent fMRI confirmation along these lines, see Pallier et al., 2011).

To take one additional example illustrating this point, consider the way that interrogative questions are formed in English, via the manipulation of the Subject and auxiliary verbs such as *is*. It was noted several decades ago by Chomsky (1965) that, given a declarative sentence such as, *the boy is sitting in the room*, the corresponding question form is given by, *is the boy sitting in the room*. Chomsky noted that the syntactic rule that forms such questions *cannot* be stated as, "displace the leftmost auxiliary verb to the front of the sentence." This is because, given a sentence where the Subject Noun Phrase contains another Sentence, such as *The boy who is sitting in the room is happy*, the corresponding question form works out as, *is the boy who is sitting in the room happy*; the corresponding question *cannot* be *is the boy sitting in the room is happy*. In other words, this syntactic rule does not pick out the *first* (as opposed to the *second* occurrence of *is*), but rather the *hierarchically most prominent* occurrence of *is*, the one that is part of the "main" sentence, *the boy is happy*.

More broadly, there is no known syntactic rule that operates on precisely the *third* element from the beginning of the sentence; that is, numerical predicates such as *third* or *fourth* are not part of the inventory of predicates in the human language syntactic system. Not only does this offer additional evidence on its own that human language syntactic structure is hierarchical, this hypothesis has been probed by psycholinguistic analysis. In a series of experiments, Musso et al. (2003) attempted to see whether there was a difference between the ability to acquire an artificial language rule that respected a numerical predicates, e.g., the formation of a question by placing a special word precisely *three* words from the start of a sentence, as opposed to a rule that respected more natural predicates for, e.g., question formation. The former type of rule they called a "counting rules." They found that such "counting rules" were indeed more difficult to acquire, being learned, if at all, as if they were "puzzles" as opposed to naturally occurring language patterns. In their later experiments, this finding was confirmed via brain-imaging: the "counting rules" activated distinctive brain regions that contrasted with those activated by "normal" linguistic rules. Unsurprisingly, the counting rules activated regions related to those also activated during non-linguistic puzzle solving. Similarly, Crain and Nakayama (1987) found that children acquired question formation rules that abided by hierarchical constraints, but never rules based on linear order.

Possibly, an even stronger position can be maintained. As far as can be determined, *all* syntactic relationships in human language syntax depend on the just the hierarchical properties of a sentence's structure, along with whether an item is simply adjacent to another item or not. Linear precedence is otherwise ignored. We present other evidence for this possibly surprising fact in Section "Human Language and Birdsong: The Key Differences" below. In contrast, in human speech (and in birdsong, as we suggest below), linear precedence *does* play a critical role; for example, in English, the past tense marker *ed* is placed at the end of a word, rather than the beginning, so that we say *chased* and not *edchase*.

The reason for decoupling human sound systems from human sentence syntax is that such key differences between spoken (or manually signed) language "output" and its internal representation bear critically on the comparison between birdsong and human language. While both birdsong and human language sound structures are linear, in the sense that left-to-right order, linear precedence, *does* matter, human language syntactic structure, drawing on hierarchical predicates, radically differs from birdsong. It is precisely here that one can pinpoint a "gap" between birdsong and human language. We return to this important point below, in Section "Birdsong Seems Analogous to Speech, Not Syntax."

Finally, as one more illustration of the hierarchical vs. linear contrast, note that the left-to-right order of the Subject, Verb, and Object in the example of **Figure 1A** is entirely particular to English. In other languages, for example, in Japanese, Bangla, and German, the Object would typically precede the verb. In this sense, the picture in **Figure 1A** might best be thought of as a mobile, with parts below the top, and at the two NP and hinge VP points, that can pivot around one another, interchanging, e.g., *the rat* with *chased the birds*. Such observed variation again underscores the fact that it is the hierarchical relationships that are central to syntax, rather than any left-to-right order.

If we now abstract away the details of the words and the names of the phrases, replacing them with labels like *XP* and *YP*, we arrive at **Figure 1C**, which highlights the basic asymmetry of human language syntactic structure. It displays a single asymmetric "molecule" structure virtually all current linguist theories posit at the heart of syntactic description. (This is true of even such otherwise divergent linguistic theories as Construction Grammar, Goldberg, 2006; Head-driven Phrase Structure Grammar, Sag et al., 2003; and modern generative grammar, Radford, 1997). Further note that the phrase *YP*, which in our example corresponds to a Verb Phrase, is partitioned into an element *Y*, corresponding to *chased*, plus another phrase, *ZP*, in our example, the Noun Phrase *the birds*. This reflects the important fact that a phrase of type *YP* is generally anchored on a word of the same sort *Y* in the way that a Verb Phrase is anchored on a Verb. We may contrast this kind of asymmetrical representation with a possible *symmetrical* structure assigned to the same sentence, depicted in **Figure 1D**, where the two Noun Phrases and the Verb are placed at one and the same level. While there is no difficulty with this representation in terms of separating out three components, NP, Verb, and NP, it is apparent that without additional information one cannot unambiguously determine which NP is the Subject, and which the Object, nor the demonstrable fact that the verb and the Object are more tightly bound together as if they were a single unit. In this sense, the symmetric representation is deficient. One could of course impose a linear ordering requirement on this triple of items to "solve" the problem of assigning the Subject and Object relations in this simple example, but this would not generalize to the full range of sentences, such as *the birds, the rat chased*. This is not to say that such structures are absent in language. For example, in conjunctions such as, *the starling ate fruit and insects*, the conjoined phrase *fruit and insects* can be reasonably construed as symmetrical – one can reverse the order to get *insects and fruit*, and obtain the same meaning. Nevertheless, asymmetrical structure remains the norm in human language. Indeed, there are evidently certain computational advantages to asymmetrical syntactic structure. For example, it has been observed since the work of Miller and Chomsky (1963), Chomsky (1963), Halle and Chomsky (1968), and Langendoen (1975), among others, that human language sentences are sometimes readjusted so as to render them asymmetric and easier to process. The classic example is the prosodic contour assigned to a sentence with several "embeddings" such as *this is the cat that bit the rat that chased the starlings*. The syntactic structure assigned to this sentence is deeply nested, as may be appreciated by its parenthetical syntactic representation, (*this* (*is* (*the cat* (*that chased* (*the rat* (*that* (*chased* (*the starlings*)))))))). However, interestingly, the sentence's prosodic contours do *not* follow the same syntactic format. Instead, there are strong intonational breaks that cut off after the asymmetrical first portion of each Subject is encountered, as may be indicated by vertical strokes: *the cat* | *that chased the rat* | *that chased the starlings* |. As emphasized by Langendoen (1975), it is as if the hierarchical structure has been "flattened," so rendering it easier to process by enabling a listener to process each chunk delimited by the vertical strokes before moving on to the next, rather than having to hold the entire Noun Phrase beginning with *the rat* in memory all at one time. Langendoen (1975) and Berwick and Weinberg (1985) suggest that this "chunking" is also partly semantic in character, in that the head word of each Noun Phrase (*cat*, *rat*, etc.) is seemingly interpreted semantically before "waiting" for the rest of the phrase (*that chased*. . .etc.) to be processed. In fact, Langendoen notes that this reflects part of a general processing strategy, what he calls "readjustment rules," that comprise some of externalization process referred to earlier. Further, there is an accumulating body of more recent results confirming the advantages of asymmetry in sentence processing; see, e.g., Fong and Di Sciullo (2005); and for a recent perspective from the perspective of neurolinguistics, confirming the basic asymmetry of language, see Friederici et al. (2011).

Though basic Subject/Object asymmetries have been confirmed by a substantial body of linguistic and psycholinguistic research, one line of experiment that has apparently not been attempted so far is in the area of artificial grammar learning. Here, the relevant questions have apparently yet to be pursued.

Why is this important for a comparison of human language and birdsong? It should also be evident that structures such as the one displayed in **Figure 1A**, accompanying the simplest of sentences, already carry us a long way from the domain of birdsong. As we describe in more detail below in Section "Human Language and Birdsong: The Key Differences," even the most complex birdsong does not use asymmetrical, hierarchical relations like that of "Subject" to fix its properties. Certain bird species such as nightingales apparently have quite complex songs which seem best described in terms of syllables linearly arranged into repeated "chunks," which are in turn arranged into song sections, then sections into packets, and finally packets into contexts (Todt and Hultsch, 1996). However, this kind of structure is neither asymmetrical nor built on combinations at one level that in turn constrain structure at one level above or below. We do not find that, say, the sequence of syllable chunks in a nightingale's song depend on the hierarchical structure of song sections or packets. This is in distinct contrast to the typical format of human syntactic structure illustrated above, where a verb that forms a Verb Phrase picks out a Noun Phrase one level *above* its structural level as the Subject. Rather, to reinforce the point made earlier, what (limited) hierarchical arrangements are found in birdsong seem fixed by a linear, left-to-right sequencing, unlike human syntax, but similar to human speech.

There is yet one more critical difference between birdsong and human language syntax, illustrated in **Figure 1E**. In human language, Sentences, Noun Phrases, and indeed phrases of any type, can be contained entirely within other Sentences, NPs, and phrases of other types *ad infinitum*. This was already illustrated by the example of question formation in an example such as, *the boy who is sitting in the room is happy*, where the phrase *the boy who is sitting in the room* is an example of a Noun Phrase *the boy*... that properly contains a sentence-like phrase, *who is sitting in the room*. Note that this kind of containment relation might be extended: we could have a sentence such as, *the boy who is sitting in the room that is on the top floor is happy*, where there are now *two* sentence-like objects contained within the Subject *the boy*. Since such sentence structures are asymmetrical, the basic asymmetrical "molecule" of **Figure 1B** is replicated at several different scales, in a self-similar, fractal-like way. Birdsong does not admit such extended self-nested structures, even in the nightingale: song chunks are not contained within other song chunks, or song packets within other song packets, or contexts within contexts. Moreover, there seems to be no evidence that distinguished structural containment relationships are manipulated in birdsong to yield distinct meanings in a way analogous to human language. In short, as **Figure 1A** indicates, such asymmetric containment relationships are basic to every sentence, the rule rather than the exception in human language.

In any case, the possibility of arbitrarily extended, labeled hierarchical structures in human language admits an open-ended number of internalized, distinct representations. In the remainder of this article, we will argue that birds seem to lack a comparable syntactic ability. This distinction remains even if one puts to one side the obvious fact that birds do not seem to have conceptual units like words, focusing purely on syntactic combinatorial abilities. While there is a single recent publication to the contrary suggesting that at least one species, Bengalese finches, might possess some facility at both learning and then perceiving open-ended hierarchical representations that fall into the class of so-called strictly context-free languages (Abe and Watanabe, 2011; see Section Birdsong Seems Analogous to Speech, Not Syntax below for a discussion of this terminology), the experimental design of this study is apparently flawed, as we discuss briefly below and as detailed in Beckers et al. (2012). This "gap" between human and avian syntactic abilities marks out a key difference between human language and birdsong, because an open-ended combinatorial syntax operating over atomic units (like words) has long been regarded as perhaps the hallmark of human language. Even though some have speculated otherwise (Petri and Scharff, 2011), there is no evidence that songbirds "name" and then re-use combinatorial units similar to *ate the birds* to arrive at an arbitrarily large number of combinatorial possibilities. **Table 1** in Section "Human Language and Birdsong: The Key Differences" brings together and summarizes all of these birdsong–human language comparisons.

In considering this summary comparison, we should emphasize that it would be a mistake to conclude that all birdsong–human differences result simply from the lack of words in birdsong, as we discuss further below. For example, even though birds lack words, there is nothing that logically blocks birdsong syntax from relying on syllable groupings or other features that could themselves be labeled by properties of their constitutive parts, which could then be assembled into more complex units in the same way that a Verb Phrase is labeled by the properties of the Verb it subsumes. Of course, this is a hypothetical example, since to the best of our knowledge no birdsong is in fact constructed in this manner. But examples like these illustrate that it is not the lack of words alone that blocks the possibility of more complex birdsong syntax. Rather, this gap is due to a fundamental deficiency in a very particular computational ability, namely, the lack of the combinatorial operation of the sort found in human language, as further described in Section "Human Language and Birdsong: The Key Differences."

Moreover, these distinctions between birdsong and human language do not entail that birdsong analysis can shed no light on human language. We conclude from our survey that birdsong currently serves best as our best animal model of language's "input–output" component, describing how language is externalized and to a certain extend acquired, along with associated auditory–vocal and motor learning behaviors, such as auditory–motor rehearsal and vocal learning by auditory feedback and reinforcement. While this certainly does not encompass full human sentence syntax, nevertheless such information seems quite valuable in focusing our understanding of how human language works, including important details as to how language is acquired and produced, in the same sense that an understanding of the input–output interfaces of a complex computer system constrains, at least in part, of the remainder of the system that lies beyond the input–output interfaces. For example, one currently fruitful line of research in child language acquisition has probed the nature of infants' abilities to acquire particular sound patterns and word boundaries in part via statistical regularities (e.g., Saffran et al., 1996; Shukla et al., 2011, among much other work). Since this acquisition process involves the "input–output" system of human language, it would seem that it is precisely here where the songbird animal model could prove most useful. Indeed, as emphasized by Yip (2006), there are many basic questions regarding the connection between human and animal sound systems that remain unanswered, such as the precise role of statistical and prosodic features in birdsong, and their possible connection to the human language case. In this way, a deeper understanding of birdsong might facilitate greater insight into the case of human language acquisition. Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-months-old infants. Finally, it seems equally misguided to reject out of hand the value of the songbird model because the "externalization" of human language can involve modalities other than sound, as in manually signed languages. In fact, the contrary seems to be true, as noted by Berwick and Chomsky (2011), and by Petitto et al. (2004); Petitto (2005): the sensory–motor sequencing involved in the human sound system can be carried over in large measure to the domain of manually signed languages. For instance, just as the physical constraints of the word limits the human sound system to the expression of dual predicates in a strictly linear, as opposed to a simultaneous fashion, e.g., *the cat chased the birds and ate the birds*, signed languages apparently operate under many of the same constraints, notwithstanding the different physical channel that logically admits such simultaneous expression.

The remainder of this article is organized as follows. We first review the basic evolutionary and neurobiological background comparing songbirds and humans with respect to auditory–vocal learning and sensory-guided motor learning, with a focus on homologous brain regions and genetic regulatory systems. Next, we situate both birdsong and human language within a common "system flow diagram" that delineates three major components: an "external interface," a sensory–motor-driven, input–output system providing proper articulatory output and perceptual analysis; a combinatorial rule system generating asymmetrically structured hierarchical sentence forms, incorporating words; and an "internal interface," a system mapping between the hierarchical structures of sentence syntax and a conceptual–intentional system of meaning and reasoning, loosely called semantics. This flow diagram will enable us to see more clearly what distinguishes birdsong and human language. We follow this system breakdown with a more detailed comparison of birdsong and human language syntax. We will see that *all* the special properties of human language syntax discussed earlier, along with others outlined in Section "Human Language and Birdsong: The Key Differences," can be directly accounted for if one assumes the existence of a single, simple combinatorial operation, anchored on words or more precisely, word features. It is this operation that is apparently absent in birds, so far as we know. However, even though birds seemingly lack words, it does not follow that the combinatorial operator is necessarily absent in birds. For example, the combinatorial operator could still work on other elements, for example, syllables, in this way yielding the distinctive metrical patterning of sound melodies, rhythmic patterns, as suggested in the domain of human language by Halle and Idsardi (1995). However, for whatever reason, the operator does not appear to have been exploited this way in birds. It remains an open question as to whether a similar analysis would apply to birdsong metrical patterns; this then is a possibly crucial open research question where a non-human model might (speculatively) provide insight into its counterpart in human language. If birdsong were found to operate in a similar way to human metrical structure, this might provide precisely the required evolutionary "bridge," in the sense that the combinatorial operator was present in the common ancestor of both species, but full-fledged language required in addition words and their features, an ability present in the human lineage, but not in any bird species. It follows that it is precisely here that one might look for key evolutionary innovations that distinguish humans from birds, a topic we briefly address in our conclusion.

AN EVOLUTIONARY PERSPECTIVE: CONVERGENT MECHANISMS AND SHARED COMMON DESCENT BETWEEN BIRDS AND HUMANS

The most recent common ancestor of birds and mammals, originating from the clade *Amniotes*, lived about 300 million years ago (Laurin and Reisz, 1995). Thus, at least 600 million years of evolution separate humans from *Aves*, a considerable stretch of time even in evolutionary terms. Given this length of time, is not surprising that birds and humans might share traits both in virtue of common descent, as well as a result of independent, convergent evolution. For example, evidently both birds and mammals share keratin genes derived from their common ancestor, giving rise to both feathers and hair, while wings and flight were developed independently by birds and bats or flying squirrels. Unsurprisingly, some traits are also a complex blend resulting both from common descent and convergent evolution. For example, birds (and their ancestors) retain a superior color vision system that was apparently lost in mammals, and then only recently recovered by certain mammals, in part by multiple recent gene duplications or other tinkering of certain retinal photoreceptive opsin control regions that differ in important details even from primate to primate – one reason people, but not birds, can be colorblind (Dulai et al., 1999). Even more recently it has been shown that bats possess "superfast" laryngeal muscles for echolocation that can work at frequencies greater than 100 Hz; such muscles are also found in songbirds (Elemans et al., 2011). Note that while such laryngeal muscles are apparently not found in humans, there is other evidence for adaptations to speech; see Fitch (2010) for a comprehensive review. Such complexity of evolutionary patterning is worth bearing in mind when discussing the comparative evolution of sophisticated behavioral traits like birdsong and human language.

A complex interplay between convergent evolution and common descent even arises within the class *Aves* itself. From the most recent genomic evidence (Suh et al., 2011) it has been proposed that the capacity for vocal learning in passerine (oscine) birds such as the zebra finch and the non-passerine vocal learning birds such as parrots is more likely to have evolved in a common Psittacopasseran ancestor as a unique evolutionary event, leading to shared genetic/neural components enabling vocal learning, such as an anterior–medial vocal pathway as delineated by standard genome expressions studies (e.g., transcription factor expression studies; Jarvis and Mello, 2000; Jarvis et al., 2000). While this phylogenetic analysis remains controversial, on this account, hummingbirds developed their vocal learning abilities separately, as a result of convergent evolution. A similar comparative evolutionary analysis is not possible for humans, since no extant primates exhibit human vocal learning abilities. Consequently, absent evidence to the contrary, for the present it seems more secure to assume that, much like hummingbirds, vocal learning in humans is a convergent evolutionary trait, with clear specializations for both auditory/motor sequencing and vocal learning and imitation. Earlier hypotheses that certain components of the vocal tract have been uniquely adapted for human speech, such as a descended larynx, now seem questionable (Fitch, 2005). More recently it has been argued that the convergent specializations for human vocalization and speech seem to lie at a deeper neural level and involve, among other components, a capacity for vocal imitation (Fitch, 2005). The recent findings regarding the role of a specific regulatory protein, Foxp2, in motor sequencing, addressed below, reinforce this view.

Turning to the interplay between common descent and convergent evolution, over the past decade many studies have confirmed that songbirds and humans possess homologous brain regions for auditory–vocal and motor-driven learning (Jarvis et al., 2005). There are several neural and genetic parallels between birdsong and speech (Bolhuis et al., 2010). The songbird brain has two interconnected neural networks, involved in song production, perception, and learning, as depicted in **Figure 2** (Bolhuis and Eda-Fujiwara, 2003, 2010; Bolhuis and Gahr, 2006; Jarvis, 2007; Bolhuis et al., 2010). First, secondary auditory regions, including the caudomedial nidopallium (NCM) and caudomedial mesopallium (CMM; **Figure 2A**), are involved in song perception and are important for the recognition of tutor song (Bolhuis and Eda-Fujiwara, 2003, 2010; Moorman et al., 2011). Second, the "song

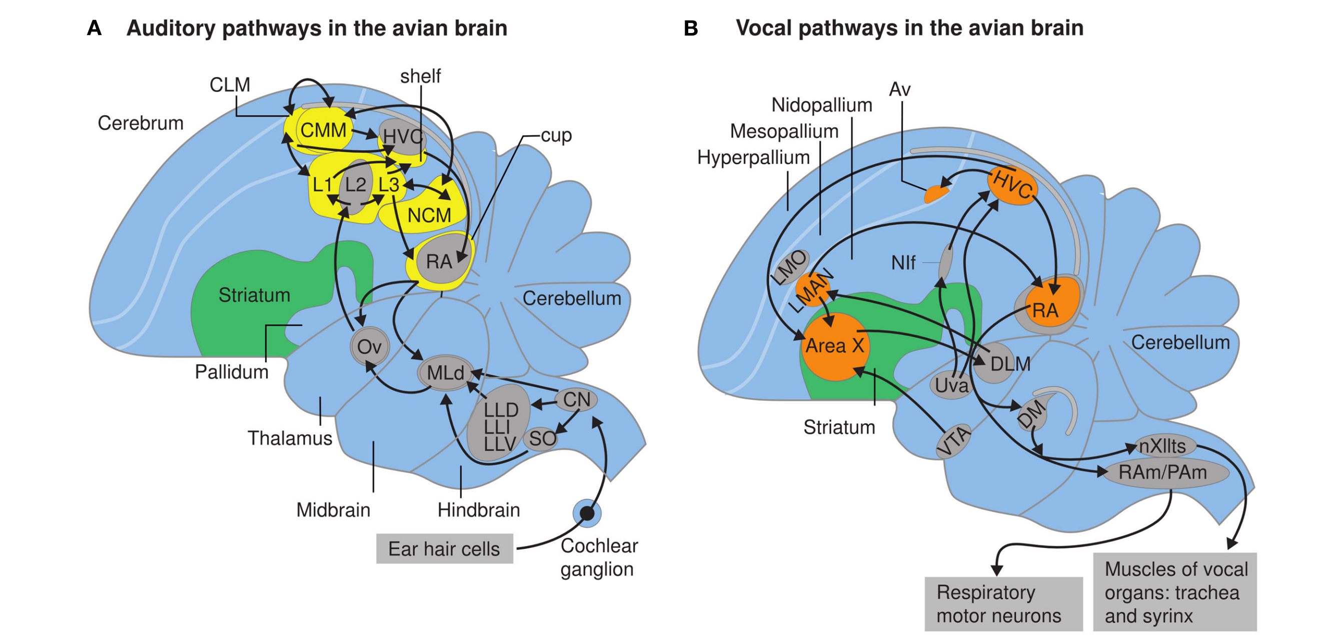

FIGURE 2 | Schematic diagrams of composite views of parasagittal sections of the songbird brain. (A) Diagram of a songbird brain giving approximate positions of nuclei and brain regions involved in auditory perception and memory. Yellow areas represent brain regions that show increased neuronal activation when the bird hears song. **(B)** Diagram of a songbird brain giving approximate positions of nuclei and brain regions involved in vocal production and sensorimotor learning. The orange nuclei in the song system show increased neuronal activation when the bird is singing (see text for details). Abbreviations: Cb, cerebellum; CLM, caudal lateral mesopallium; CMM, caudal medial mesopallium; DLM, nucleus dorsolateralis anterior, pars medialis; HP, hippocampus; HVC, acronym used as a proper name; L1, L2, L3, subdivisions of Field L; LaM, lamina mesopallialis; LMAN, lateral magnocellular nucleus of the anterior nidopallium; mMAN, medial magnocellular nucleus of the anterior nidopallium; NCM, caudal medial nidopallium; nXIIts, tracheosyringeal portion of the nucleus hypoglossus; RA, robust nucleus of the arcopallium; V, ventricle. Modified and reproduced, with permission, from Bolhuis et al. (2010) and Moorman et al. (2011), copyright 2010 Nature Publishing Group. All rights reserved.

system" is involved in song production and certain aspects of song learning (**Figure 2B**). The song system is subdivided into two major pathways, the song motor pathway (SMP; Mooney, 2009) and the anterior forebrain pathway (AFP; Brainard and Doupe, 2000; Doupe et al., 2005). The SMP is a posterior motor pathway connecting the HVC (acronym used as a proper name), the robust nucleus of the arcopallium (RA) and the tracheosyringeal portion of the nucleus hypoglossus (nXIIts), and is important for song production. The AFP is an anterior cortical–basal ganglia–thalamic loop that originates in HVC and passes through Area X, the thalamic nucleus dorsolateralis anterior, pars medialis (DLM) and the lateral magnocellular nucleus of the anterior nidopallium (LMAN), and eventually connects with the motor pathway at the nucleus RA. The AFP is essential for sensorimotor learning and adult song plasticity (Brainard and Doupe, 2002; Mooney, 2009).

In humans, conventionally the neural substrate of motor representations of speech is thought to involve Broca's area in the inferior frontal cortex, while perception and memory of speech is considered to involve Wernicke's area and surrounding regions in the superior temporal cortex. Although there are considerable differences between avian and mammalian brains, there are many analogies and homologies that have recently prompted a complete revision of the nomenclature of the avian brain (Jarvis et al., 2005). Similarities in connectivity and function would suggest at least analogies between the human neocortex and the avian pallium (including the hyperpallium, mesopallium, nidopallium, and arcopallium; see **Figure 2A** (Bolhuis and Gahr, 2006; Bolhuis et al., 2010). Specifically, Bolhuis and Gahr (2006) have suggested that the NCM and CMM regions in the songbird brain may be analogous with the mammalian auditory association cortex. In addition, Doupe et al. (2005) have argued that the AFP loop in the song system (**Figure 2B**) bears strong similarities in connectivity, neurochemistry and neuron types to the mammalian basal ganglia, while both LMAN and HVC have been suggested to be functionally similar to Broca's area (see Bolhuis et al., 2010 for further discussion).

In addition to these neuroanatomical parallels, there is increasing evidence for a similar neural dissociation between auditory recognition and vocal production regions in the brains of songbirds and humans (Gobes and Bolhuis, 2007; Bolhuis et al., 2010). Regions in the songbird caudomedial pallium (including the NCM) contain the neural representation of tutor song memory that is formed in juvenile males (Bolhuis and Gahr, 2006), whereas nuclei in the song system are required for sensorimotor learning and song production (Brainard and Doupe, 2000). Lesions to the NCM of adult zebra finch males impaired recognition of the tutor song, but did not affect song production, while lesions to the HVC in songbirds disrupted song production, but lesions to the nidopallium and mesopallium did not (Gobes and Bolhuis, 2007; Bolhuis et al., 2010). These and other findings suggest that in songbirds there is a neural dissociation between song recognition and song production that is already apparent in juveniles (Gobes and Bolhuis, 2007; Gobes et al., 2010). In human speech there is a comparable dissociation between brain regions involved in auditory perception and memory on the one hand, and vocal production on the other. Human newborns show increased neural activity in the superior temporal lobe, but not in the inferior frontal cortex, in response to human speech (Imada et al., 2006), while 3- to 12-month-old infants showed activation in both Wernicke's and Broca's areas in response to hearing speech (Dehaene-Lambertz et al., 2006; Imada et al., 2006). Taken together, these studies suggest that Wernicke's area is (part of) the neural substrate for speech perception in neonates and that Broca's area becomes active at a later stage, when infants start babbling; see Bolhuis et al. (2010), Brauer et al. (2011).

It is not yet completely clear whether these neural structures and information processing pathways are the result of shared ancestry, and so represent instances of homology, as opposed to convergent evolution and so analogy. Much remains to be understood about the detailed genetic, developmental, and neural underpinnings of vocal learning and language in both species. One key genetic parallel between birdsong and speech involves FOXP2, the first gene specifically implicated in speech and language (Fisher and Scharff, 2009). This is an ancient gene that codes for the transcription factor FoxP2, a protein that regulates DNA expression. Mutations in this gene in a large three-generation family and in some unrelated individuals were found to correlate with a speech disorder (Fisher et al., 1998). FOXP2 sequences are highly conserved between birds and mammals, and FOXP2 mRNA is expressed in song nuclei in the three known orders of song learning birds. FOXP2 is developmentally and seasonally regulated in songbirds and intact FOXP2 levels are required for normal song learning (Fisher and Scharff, 2009). As noted by Scharff and Petri (2011), this system may be part of a "molecular toolkit that is essential for sensory-guided motor learning" in the relevant brain regions of both songbirds and humans. Depressed vocal learning in songbirds that has been attributed to FoxP2's role in regulating other genes involved guiding neuronal development (Haesler et al., 2004; Vernes et al., 2011). In this sense, FoxP2 serves as an example of "deep homology" – a shared trait involved as part of both human speech and songbird vocal learning (Bolker and Raff, 1996; Shubin et al., 1997; Fitch, 2011; Scharff and Petri, 2011). However, the scope of this homology must be considered with some care. Since both vocal learning and non-vocal learning birds possess identical FoxP2 genes (Webb and Zhang, 2005), and the birds' FoxP2 genes are distinct from those of humans, differences in this gene alone cannot be what accounts for the vocal learning/non-learning distinction in birdsong. Rather, this difference seems to reflect differential gene expression as part of some larger overall gene network, as Haesler et al. (2004, p. 3174) note, "*FoxP2* has a characteristic expression pattern in a brain structure uniquely associated with learned vocal communication, Area X in songbirds." From this point of view, FoxP2 comprises one of probably many necessary ingredients in a complex recipe for vocal learning and production, rather than a single "master gene" that sits at the top of a regulatory cascade as in the case of the well-known regulatory *Pax*-6 *eyeless* homeobox gene (Halder et al., 1995).

BUILDING BLOCKS FOR HUMAN LANGUAGE

To better frame a comparison between birdsong and human language, it is helpful to partition language's fundamental relationship between sound and meaning into three distinct components: (1) an input–output system encompassing how language is produced, either acoustically, by vocal production, or manually, by signed gestures, as well as how language is perceived, by the auditory or visual system; (2) an internal rule system generating legitimate organism-internal structured representations, including, but not limited to, the kinds of structures depicted in **Figures 1A,E**, and (3) a system interfacing to cognitive processes such as meaning and inference, often glossed as "semantics." The first component includes representations such as the placement of stress that are not strictly sensory–motor in nature. In current linguistics, this component includes both acoustic phonetics and phonology. The second, rule-governed component feeds the other two, both the input–output interface as well as the semantic interface. This division is by no means universally accepted. For example, some linguistic accounts reduce or eliminate the role of a distinctive syntactic component, instead assuming a more direct relationship between sound and meaning (e.g., Culicover and Jackendoff, 2005; Goldberg, 2006; Jackendoff, 2010).

For example, Jackendoff (2010) argues that both components (1) and (3) have additional, separate interfaces to the mental repository of information about words, the lexicon, bypassing the syntactic component (2). Such additional links are quite plausible, because words – lexical items – have both phonological and semantic aspects, their particular sounds and meanings. In Jackendoff's view, such a division lends itself to a more natural evolutionary account where sounds and meanings might similarly be directly connected, without the intervention of syntax, this possibly serving as a kind of "protolanguage" stage. On the other hand, this position requires that there be an independent generative component for semantic representation, one that, according to Jackendoff, antedated human language syntax. At the moment, there seems to be little hard evolutionary evidence to distinguish between such alternatives, and in any case, the three-way division suffices for the bird–human comparison. This three-way dissection does factor apart the distinct knowledge types and representations generally recognized as central to language, in one way that enables a fruitful comparison.

BIRDSONG SEEMS ANALOGOUS TO SPEECH, NOT SYNTAX

Referring then to these three components, it is important to respect both the similarities and the differences between human speech and the totality of human language on the one hand, and birdsong on the other, which can and have led to some common misunderstandings. While speech is one prominent component of human language, it is neither necessary (as manually signed languages illustrate) nor sufficient. Rather, human speech, or more precisely, the sequenced motor commands involving a small number of vocal articulators such as the tongue, lips, velum, and larynx, comprises the end product of more sophisticated cognitive computations that engage at least two additional components: first, an internal combinatorial syntax; and second, a mental representation of both individual words and their meanings as determined by a particular syntactic combinations.

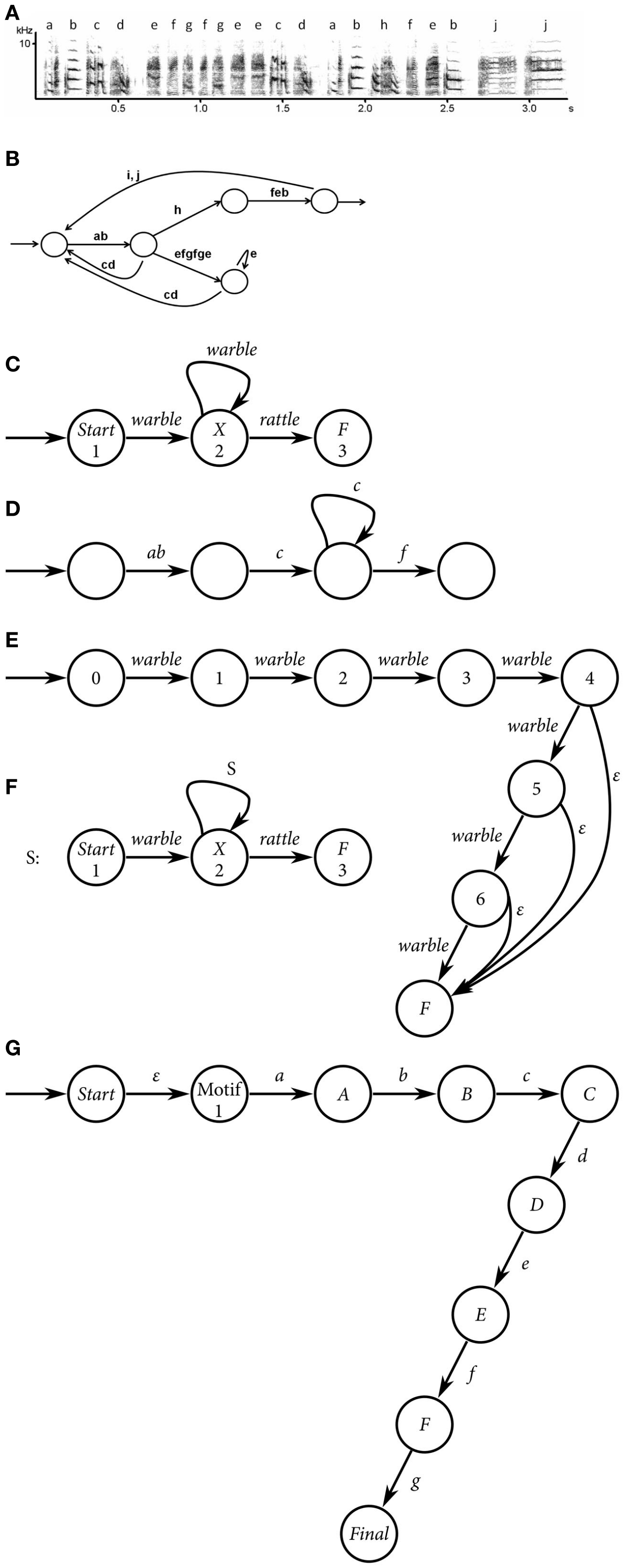

FIGURE 3 | Continued

FIGURE 3 | Continued

(A) Sonogram of an adult male Bengalese finch. *X*-axis is in seconds, *Y*-axis in kilohertz. Song syllables are demarcated by alphabet letters. **(B)** Finite-state transition network corresponding to the song syllable sequence in **(A)**. The network begins at the extreme left. Open circles correspond to states in the network, with transitions on arcs labeled with the syllables corresponding to those identified from the sonogram in **(A)**. Note that loops in the network can go back to previous states. **(C)** A finite-state transition network that generates syllable sequences containing at least one or more *warbles*, ending with a *rattle*. **(D)** A finite-state transition network encoding an "unbounded" dependency, in the sense that a syllable sequence beginning with *ab* must always end with exactly a single *f*. Note that syllable sequences may be arbitrarily long, due to the loop labeled with a *c* from one state back to itself. Thus, even though the *ab* and *f* may be arbitrarily far apart, a finite-state network can still determine whether this constraint holds. **(E)** A finite-state transition network that "counts" any number of *warbles* between four through seven, inclusively. The transitions labeled with ε denote so-called "epsilon transitions" where an output syllable is not produced when moving between states. **(F)** A recursive transition network labeled *S* that uses *S* itself on the transition looping from state 2 back to state 2 as a subroutine to generate an indefinitely large number of properly nested *warble-rattle* pairs. States are numbered for convenience. **(G)** A finite-state transition network that describes a hypothetical zebra finch song motif, as represented by a sequence of seven syllables, *a* through *g*. Note that if there are no nested dependencies, then the state *Motif* could be reached from any other state as part of a larger network describing the overall song.

In order to meet the demands of real-time speech/signed language production, in some way the human language system must map structured syntactic word combinations onto a sequence of motor commands, feeding a sensory–motor articulatory/gestural system for vocal or signed output, "flattening" the structure onto the output channel so that vocal output is sequentially ordered; see Stevens (2000). Conversely, the human processor recovers hierarchical structures from a time-ordered sound sequence. We might call this output projection *externalization*. It is typically here that *linear precedence* relationships hold among word elements in regards to their output as articulatory sequences, as was noted in the Introduction. Importantly, the detailed study of human sound systems has established that only linear precedence relations are required for the description of such systems; see Heinz and Idsardi (2011) and Wohlgemuth et al. (2010) for further discussion. To consider another simple language example here, the plural marker for *apple*, the so-called *z* morpheme in English, is placed at the end of *apple*, rather than the front, yielding *apples* (pronounced *applez*), rather than *zapple*. Conversely, if one regards the perception of language as mapping the time stream of acoustic signals into an internal representation, one must invert this process, recovering the hierarchical structures associated with sentences from the "flattened" signal.

From this standpoint, it is misleading to equate birdsong vocal production with the totality of human language. As we will now argue in some detail, birdsong seems more comparable to human language sound systems, not human language syntax. As we will argue, both human and bird sound systems are describable solely in terms of a network of what basic sound elements can come before or after one another – either syllable chunks in the case of birdsong, or so-called phonemes in the case of human language.

We will formalize this intuition below as the notion of a *finite-state transition network.*

What does this difference amount to descriptively? For birds, songs may consist of individual notes arranged in order as syllable sequences, where a syllable is defined, contrary to its usual meaning in linguistic theory, as a sound preceded and followed entirely by silence. Birdsong syllables, in turn, may be organized into recognizable sequences of so-called "motifs," and motifs into complete song "bouts." In some cases, the description seems to require more complexity than this, a matter considered in detail in what follows. Depending on the songbird species, the motif arrangements and ordering vary greatly, with the transitions between motifs probabilistic. For example, starling song bouts may be composed of many distinctive syllabic motifs lasting 0.5–1.5 s, up to a total length of 1 min (Gentner and Hulse, 1998), while nightingale songs consist of fixed 4-s note sequences, but arranged into a number of distinctive "chunks" with up to 200 distinctive song types. Supporting this view, Gentner and Hulse (1998) found that a first-order Markov model is sufficiently complex to describe possible starling song sequences.

Figure 3A displays a representative sonogram of a Bengalese finch song, with distinguishable syllables labeled as *a*, *b*, *c*, and so forth. By assembling a large sample of this bird's songs, one can extract a corresponding state diagram description as exhibited by **Figure 3B**. This picture consists of a finite, ordered sequence of states, the open circles, with transitions between the states labeled either by certain single syllable sequences, such as *h*, or multiple syllable units such as *ab* or *efgfge*. There are also loops that can carry one back to previous states, such as the syllables *i* or *j* that return to the leftmost open-circle state. By tracing out a syllable sequence starting from the entering arrow at the leftmost circle in the transition network, through to the exit arrow on the right, the network spells out or *generate* the entire set of legitimate syllable sequences for this bird's song repertoire, e.g., *ab efgffge cd ab h feb.* Note that even though there are only a finite number of states in this network, because of loops between some states, there can be a countably infinite number of valid possible paths from the starting arrow to the end state. To capture a bird's behavioral repertoire, typically these transitions are defined probabilistically, so that a transition between states occurs only with some positive probability corresponding to the likelihood of observing such a transition in the behaviorally observed data (Kakishita et al., 2009).

Such descriptions are conventionally called *finite-state transition networks* (Kleene, 1956); see **Figure 3C**. We now situate these within the standard framework of formal language theory (Hopcroft and Ullman, 1979). Here, a *language* is defined as any set of strings, equivalently, sentences, defined over a (fixed) alphabet, where the alphabet consists for example of the distinct syllables in a birdsong, or the distinct words in a human language. So for example, we might describe a particular birdsong "language" as consisting of "sentences" with any number of *warble* syllables *w* followed by an ending coda syllable, *rattle*, *r*. Such a birdsong language would contain an infinite number of "sentences," or songs, *wr*, *wwr*, *wwwr*, and so forth. Formally, languages are said to be *generated* by *transition networks*, where a finite-state transition network is a directed, labeled graph, consisting of a (finite) set of states, the nodes in the graph, connected by directed, labeled arcs, the edges of the graph. The notion of *generation* means that one can traverse the graph, beginning at a single designated *Start* state (denoted by a single incoming, unlabeled arrow in **Figures 3B,C**), and ultimately arriving at one or more designated *final* states. Generated sentences correspond to the sequence of labels on the edges arising during graph traversal. The set of all such possible label sequences from the *Start* state to a *final* state constitutes the *language generated* by the transition network. For present purposes, we need consider only two distinct types of networks: first, the *finite-state transition networks*; and second, a more powerful type of network, the *recursive transition networks* (Woods, 1970). (There is an equivalent approach that can be framed in terms of rule systems called *grammars*, either *regular* grammars, corresponding to the finite-state transition networks; or *context-free* grammars, corresponding to the recursive transition networks.)

We first consider finite-state transition networks and the languages they can generate. Finite-state transition networks can enforce the constraint that all syllable strings begin and end with one *warble*, or, following our earlier example, that a song contains any positive number of warbles, and end with a special final syllable *rattle.* The finite-transition network displayed in **Figure 3C** displays a finite-transition network obeying this second constraint. Let us see how. Generation begins at the *Start* state, also numbered 1. If we make a transition along the directed edge labeled *warble* to the state *X* (numbered 2), the system generates the first syllable in a possible output string, a *warble.* From state *X* there are two possible directed edges in the graph: one that leads back to state *X*, labeled with *warble*, and the other leading to the (single) distinguished final state *F* (numbered 3), labeled with *rattle.* If we take the transition labeled with *warble* back to state *X*, the generated sequence includes a second warble, and we can clearly continue in this way to output any number of *warbles* by traversing this loop any number of times. As soon as the system makes the transition from state *X* to state *F*, the syllable sequence ends with *rattle*, as desired. Note that the language so generated contains an infinite number of legitimate syllable strings, even though the network itself is entirely finite. It is in this sense that a finitely represented object can compute an extensionally infinite set of possible sentences.

More generally, the set of all finite-state transition networks generate the (syllable) stringsets called the *regular languages*, equivalently, stringsets defined by *regular expressions* (McNaughton and Yamada, 1960). Dependencies encoded by the regular languages can appear quite complex, including dependencies between items that are arbitrarily far apart from each other, what are sometimes called "unbounded dependencies." For example, the set of strings that begin with the syllable chunk *ab*, and then are followed by any positive number of *c* syllables, ending with an *f* that matches up with the beginning *ab*, can be described with via the regular expression $abc^{+}f$, where the $^{+}$ symbol denotes "1 or more occurrences." This language thus expresses an "agreement" constraint between the first and last syllables of any legitimate syllable sequence, even though there can be an indefinite number of *c*'s between the leading *ab* and the final *f.* Such "unbounded dependency" sequences can be generated by a very simple finite-state transition network with just four states, as displayed in **Figure 3D**. Petersson et al. (2012) are thus correct to point out that "the phenomenon of non-adjacent dependencies… can not simply be reduced to a choice between regular [i.e., finite-state transition network] or non-regular grammars [i.e., recursive transition networks].” However, as described in the introduction and as we pursue in more detail below, the phenomenon of *containment* of one type of phrase within a phrase of another type, when carefully articulated, *can* adjudicate between these two types of rule systems.

It appears that finite-state transition networks suffice to describe all birdsong. Indeed, it remains unclear whether birdsong even contains unbounded dependencies of the sort described in the previous paragraph, if we follow the results of Gentner and Hulse (1998) and others that first-order Markov processes, a more restricted network system, suffices to describe birdsong. (For a more recent confirmation of this claim, see Katahira et al., 2011.)

There are some apparent exceptions that merit additional discussion. Researchers have observed that the songs of certain bird species, such as chaffinches, consist of sections that must contain a particular *number* of iterated syllables of a certain sort, e.g., between 4 and 11 *warbles* (Riebel and Slater, 2003). Consequently, Hurford (2011) proposes adding a numerical *counter* to finite-state transition networks to accommodate such patterns, suggesting that this amounts to a “significant increase in the power of the processing mechanism” (p. 54).

However, “counting” up to a fixed bound or counting within finite interval is well within the descriptive power of ordinary finite-state transition networks. One simply grafts on a sequence of states that spells out the possible integers from 4 to 11. **Figure 3E** displays a simple illustrative example that captures the “4–11” chaffinch syllable patterns, though it saves space by only displaying a network that counts out four through seven *warble* syllables. The network uses transition arcs labeled with *warbles*, as well as a second kind of transition, labeled with an epsilon, which means that one can move between the indicated states without a corresponding output syllable. In this way, the network can count out four *warbles* and then move to its final state; or five *warbles* and move to the final state, and so forth). This is not the only way to implement finite “counting” bounds of this sort, while remaining within a finite-transition network framework. As is familiar from the literature on finite-state machines, bounded arithmetic operations are straightforward to implement in finite-state devices. Minsky (1967) has many examples illustrating how finite-state adders and counters of this sort may be implemented. In any case, as we describe in more detail just below, such patterns, even of this iterative sort, still form a highly restricted subset of the entire set of patterns that the finite-state transition networks can describe, crucially, one that is easily learned from positive exemplars of adult tutors’ songs to juveniles.

What sorts of constraints *cannot* be described by finite-state transition networks? Roughly, such systems cannot describe containment constraints that can be arbitrarily nested, in the sense that the state transitions generate syllable sequences in form, (*warble*$_1$ (*warble*$_2$ (*warble*$_3$ … *rattle*$_3$) *rattle*$_2$) *rattle*$_1$). Here we have indicated that *particular warbles* and *rattles* must be paired with each other by the use of subscripts, matching from the inside-out, so that the innermost *warble* must be associated with the innermost *rattle*, the next innermost *warble* with the next innermost *rattle*, and so forth. The ellipses indicate that a song might have, at least in principle, an indefinite number of such nestings, to any depth. We have artificially introduced parentheses to more clearly indicate the grouping structure, which is not actually part of the string sequence. Long-standing results (Chomsky, 1956; Rabin and Scott, 1959) demonstrate that such patterns cannot be generated by any finite-state transition network, because, for example, in order ensure that each *warble*$_i$ on the left is matched with its corresponding *rattle*$_i$ on the right one must in effect be able to match up *warbles* and *rattles*, working from the innermost *warble*$_i$ *rattle*$_i$ pair outward. To do this matching requires the machine to use one state to “remember” that an *warble*$_i$ has been seen, until the corresponding *rattle*$_i$ has been seen, one state for each possible *warble*$_i$. But this means that to check a candidate string *warble*$_1$ *warble*$_2$ *warble*$_3$…*warble*$_n$ *rattle*$_n$ *rattle*$_n$*—1*…*rattle*$_2$ *rattle*$_1$ for validity, one must have at least *n* states in the corresponding transition network. If *n* can be arbitrarily large, no machine with a finite number of states will be able to do the job correctly; an indefinitely large memory is required. At a minimum, one must augment a finite-state network with a single counter that is increased by 1 each time a *warble* is seen, and decremented by 1 each time a *rattle* is seen, and the counter must be able to “count” arbitrarily high.

To handle such examples, one must move to a more powerful computational device, such as recursive transition networks (Woods, 1970); equivalently, context-free grammars. For networks, the augmentation involves some means of invoking subportions as though they were subroutines in a computer program. This can be done by expanding the domain of labels on transition arcs to include the names of whole networks, rather than just output symbols such as *warble* or *rattle*. **Figure 3F** illustrates one way to build such a network, where we have numbered the states for convenience. Referring to this figure, we name this entire three-state network with the label *S* and then add a transition from the second state of that network back to that same second state via a transition labeled *S* (the name of the entire network itself). Such a network machine can be organized to use itself as a subroutine, to spell-out all and only the legitimately paired *warble-rattle* sequences.

To see how such an augmented network can generate the syntactically valid string *warble-warble-rattle-rattle* we can again trace through an imagined traversal from the *Start* state to the *Final* state of the network. Again referring to **Figure 3F**, the machine begins in the *Start* state 1, and then travels to state 2, corresponding to *warble*. It can now traverse the network *S* again, by following the loop labeled *S* that goes from state 2 back to state 2, rather than making a transition to the *Final* state (and outputting a *rattle*). This means moving to state 1 again, with the proviso that the network implementation must “remember” that it must return to state 2 when it has traversed the *S* network successfully, by arriving at the final state. We now suppose that during this second passage through the *S* network the machine moves from state 1 to state 2, and outputs another *warble* as before, so that so far the sequence generated is *warble-warble*. If we now have the machine make a transition to state 3, the final state of the network, it adds a *rattle*, which in this case is paired up with the immediately preceding *warble*, as required. However, instead of simply ending its computation at this point, the network has only completed its second traversal of the entire *S* network. It thus must remember that it is required to return to the state where the *S* network was invoked for the second time, namely state 2, and can finish by making a transition from state 2 to state 3, outputting a second *rattle*. In this way the network generates (alternatively, verifies) the desired, legal syllable sequence *warble-warble-rattle-rattle*.

To organize a transition network this way so as to be able to use its parts as if they were subroutines is typically implemented by means of an additional, special memory structure, what is called a *pushdown stack*. As is familiar, a pushdown stack stores information in a first-in, last-out order, like a stack of dinner plates: if items *x*, *y*, and finally *z* are placed on the stack in that order, then the order in which they are removed must be the reverse of this, namely, *z*, *y*, *x*, in this way obeying the characteristic "nested" structure in our example. So for example, traversing the network *S* for the first time, the number of the state to return to, say, 2, would be placed on the pushdown stack. After traversing the *S* network the second time and moving to the final state, the machine would examine the top symbol on its stack, remove it, and returning to the state indicated, in this case state 2, and continue. In this way, a sequence of $n-1$ *warbles* would result in a corresponding sequence of $n-1$ invocations of the network *S* and $n-1$ instances of state symbol 2 being placed on the pushdown stack. Returning from this sequence of invocations in turn and traversing from state 2 to 3 each time will output $n-1$ *rattles*, leaving the machine in state 2 with a single final *rattle* transition to make to reach the end of its very first full traversal through the *S* network, generating the proper sequence of *n warbles* followed by *n rattles*. (As suggested above, since one need only put a single fixed state symbol 2 on the pushdown stack, one could also implement this particular network with a single *counter* that simply indicates the number of 2's that have been placed on the stack, decrementing this counter as each transition to the final state is made.)

Adapting this approach to human language requires more. If we have at least two networks with different labels, say S (corresponding to a Sentence), and NP (corresponding to a Noun Phrase), then the resulting system can be set up to generate Noun Phrases properly containing Sentences, and vice-versa, in the manner suggested by our *the rat chased the birds*... example cited in the Introduction. Such a system would place at least two distinct symbols on its stack, corresponding to the two different types of phrases. This seems to be the minimum augmentation required to describe human language syntax, and goes beyond augmentation of a finite-state transition network with a single counter. One can choose to augment a finite-state device with two counters, but this makes such a machine as powerful as any general-purpose computer (Hopcroft and Ullman, 1979), which would seem to be quite powerful indeed. Below we suggest that human language may be more restricted than this.

It has also sometimes been suggested (see, e.g., Hurford, 2011; Scharff and Petri, 2011) that the shallow hierarchical structure of birdsong, with syllables organized into motifs, and then into some linear motif sequence, could be interpreted as representative of a *general* hierarchical structure-building competence in birds. This conclusion seems too strong. Note that the hierarchical structure here is quite limited. It is comparable to how linguists have described the sound structure of human words in terms of linear syllable sequence "chunks." For example, the word *starling* can be broken down into two consonant (C) vowel (V) combinations, with the first consisting of two consonants, *st-ar* and *l-ing*, that is, CCV–CV. Here the same CCV combination shows up in other words, such as *startle*, so it is similar to a birdsong chunk or motif. We may call this kind of re-use of a linear sequence *linear grouping*. In any particular language such as English, only certain linear groupings are possible. For example, the sequence, *st-ar* is possible in English, while *st-xa* is not. In this way, legitimate CV sequences can be spelled out as allowed linear grouping sequences. This also appears to be true of birdsong.

In both birdsong and human language, this kind of linear grouping has also been shown to have psychologically verifiable correlates. For example, Suge and Okanoya (2010) demonstrated that Bengalese finches perceive songs in terms of syllable "chunks" that can be detected by placing a brief noise either at the boundary of chunks or in the middle of chunks, while training birds under operant conditions to react to the noise as quickly as possible. The birds' reaction time differed in these two conditions, with a longer reaction time for noise introduced into the middle of a chunk, indicating that birds reacted to "chunks" as categorical units for production. In humans, syllable chunks have been found to be an integral part of perception, even in very young infants as early as 4 days old (Bijeljac-Babic et al., 1993).

Some researchers have suggested that linear grouping implies that the underlying birdsong *must* be modeled by a recursive transition network system, but this conclusion too seems unwarranted. For example, Hurford (2011) posits that nightingale song necessitates description in terms of context-free rules (what Hurford calls, "phrase structure rules," equivalent to what augmented transition networks can describe). Hurford further grounds his claim on certain neurophysiological evidence from Fee et al. (2004) regarding the interaction between HVC–RA nuclei in zebra finches' brains during song production. Hurford advances the hypothesis that there is a putative rule expanding a finch birdsong motif as a particular set of seven syllables, *a* through *g*, that is literally represented in a finch's brain by means of HVC–RA interaction, where this rule may be invoked any number of times:

(2) $Motif1 \rightarrow a\ b\ c\ d\ e\ f\ g$

However, it is extremely difficult, if not impossible, to distinguish this possibility from one that simply encodes this sequence as a small finite-state transition network, as displayed in **Figure 3G**. Note that the finite-state transition network, as usual, uses only a small finite amount of memory; it seems entirely possible that a bird could store dozens of such network snippets. No stack-like augmentation is necessary, since, as Hurford himself notes, the entire system in such cases remains a first-order Markov network. By the definition of a first-order Markov system, a finch does *not* have to "remember" whether a motif of one type is "embedded" within another of the same type; it simply has to branch to the part of the network shown in **Figure 3G** at any one of a number of distinct points within a larger, overall song sequence. The sequence would remain entirely linear. Counterfactually, if it were the case that finch song incorporated *nested* dependencies of the "warble-rattle" sort that we described above, then one would be forced to use a more powerful network. But as Hurford himself states, this does not appear to be true of the finch's song. Further, in a recent study, Katahira et al. (2011) demonstrate that very simple first-order Markov processes, even simpler than the "long-distance" finite-state transition networks described above, along with interaction between the HVC–RA bird brain nuclei, can yield apparently "higher order" syllable constraints of precisely sort that Hurford describes.

Currently, then, there is no compelling evidence that recursive transitions networks *must* be literally encoded in finch's brains. To distinguish between the finite-state and non-finite-state possibilities demands artificial language learning experiments that are carefully crafted to distinguish between these two possibilities, along the lines of the experiments carried out with human subjects by Uddén et al. (2011). There is one recent, controversial artificial language learning experiment in Bengalese finches (Abe and Watanabe, 2011) that superficially appears to run counter to this conclusion. However, as demonstrated by Beckers et al. (2012), and as we touch on briefly below, the experimental design here seems to be flawed because the training and testing materials confound acoustic familiarity with syntactic well-formedness. In fact, Uddén and colleagues show that even in the human case, it can be extremely difficult to distinguish experimentally between the use of adjacent dependencies, requiring only a first-order Markov description, and non-adjacent dependencies that might tap the power of a pushdown stack. Absent such careful experimentation, which has to date not been carried out in birds, all current evidence suggests that only finite-state transition networks are required to describe a bird's "knowledge of birdsong." Indeed, it would be surprising if this were not true, since this is in line with what is also known about the acquisition and use of human sound systems as well (Heinz and Idsardi, 2011).

As mentioned earlier, birdsong appears to be much more constrained than this, however. It appears to be describable by a narrowly constrained subset of the regular languages (Berwick et al., 2011a), namely, those that are learnable in a computationally tractable way from examples sung to juvenile males by adult tutors. Here "computationally tractable" adopts its usual meaning in computer science, namely, computable in a length of time proportional to *kn*, where *n* is number of states in the to-be-acquired network and *k* is a small "window size" of one to three syllables. This is an extremely favorable result from the standpoint of both perceptual processing and learning, since in general, learning finite-state transition networks is not possible even given a large number of positive examples, possibly exponential with respect to the number of states in the final, to-be-acquired network (Gold, 1978). Intuitively, this is true of general finite-state transition networks because if all we know is that a target automaton is a finite-state automaton with *n* states, then it could take a very long string to distinguish that automaton from all other possible *n*-state machines. More precisely, it appears that one can characterize the formal complexity of birdsong sound systems as a so-called *k*-reversible finite-state transition network (Angluin, 1982; Berwick and Pilato, 1987; Berwick et al., 2011a). Sasahara et al. (2006) have shown that one can in fact apply the same computer algorithms described by Berwick and Pilato to the problem of automatically inducing *k*-reversible transition networks from birdsongs. For instance, the finite-state transition network described in **Figure 3B** is *k*-reversible.

There is no comparable learnability result for human language sentence syntax. However, if one restricts one's domain to human language sound systems, as Heinz (2010) among others have shown, one can obtain a comparable positive learnability result. In this respect then, birdsong and human sound systems again seem alike in terms of ease of learnability (Heinz, 2010). In this context, it should be noted that it is sometimes suggested that the difficulty of learning human syntax as established by Gold (1978) and others can be overcome by adopting another learnability framework. For example, one might adopt a statistical approach, such as rules that apply probabilistically; or a learning system that selects rule systems according to a size criterion (where a smaller rule system is better; equivalently, a Bayesian formulation); While a detailed analysis of such proposals like these lies outside the scope of this paper, in fact while these methods might eventually turn out to be successful, none of them solve the problem of human language acquisition. Such alternatives were originally advanced by Solomonoff (1964), Horning (1969), and later pursued by Berwick (1982, 1985), Stolcke (1994), De Marcken (1995, 1996), and, more recently, Chater and Christiansen (2010) and Hsu et al. (2011), among several others). However, these results have yet led to provably efficient algorithms that cover substantial linguistic knowledge beyond sound systems. Simply making rules probabilistic actually does not work, particularly for sentence syntax that is not describable by means of a finite-transition network. One this point see, e.g., Stolcke (1994), De Marcken (1995), and Niyogi (2006) for further discussion as to why this is so. Intuitively, it is actually more difficult to estimate probability *distributions* over some function that learns a rule system than simply learning the learnability function itself. In particular, current alternative approaches either advance a method that has *no* corresponding constructive algorithm, let alone an efficient one (Solomonoff, 1964; Hsu et al., 2011); or rely on the hand-construction of an initial grammar that is in any case covers but a small fraction of the human language system (Perfors et al., 2010). (See De Marcken, 1996; Niyogi, 2006, for further discussion of why moving to a probabilistic setting does not solve the difficult question of language learnability; and Berwick et al., 2011b for a detailed analysis of recent approaches.)

Formally, a finite-state transition network is *k*-reversible if, when we exchange the *Start* and final states, and then reverse all the directed transitions from one state to the next, then the resulting new network can be traversed deterministically, that is, without choice points. More intuitively, what this means whenever two prefixes of a song whose last *k* words match have an end-sequence in common, then they have *all* end-sequences in common. A juvenile learner can acquire such a language by considering examples of an adult male's song, incrementally. For example, if it is the case that a valid song consists of sequences such as *warble-rattle*; *warble-rattle*; *warble-rattle-rattle*; *twitter-rattle*; and *twitter-rattle-rattle*, then all the sequences following *warble* or *twitter* are shared, and the language is 0-reversible. If this hypothetical birdsong language contained in addition the sequence *warble-rattle-rattle-rattle*, since the end-sequence *rattle-rattle-rattle* does not follow *twitter*, then the language is not 0-reversible, unless the bird "generalized" its language to include *twitter-rattle-rattle-rattle*, thereby maintaining 0 reversibility. The 1-reversibility constraint is similar, buts adds an additional syllable of "lookahead," a predictive window 1-syllable long: it asserts that if some syllable *plus* 1 additional syllable – so a chunk of two syllables in all – has *one* suffix in common with another two syllable chunk with the same second syllable, then such a pair of two syllable chunks must have *all* suffixes in common. We can illustrate the difference between 0 and 1-learnability with another caricatured birdsong example. Suppose the set of possible songs consisted of the following five syllable sequences: (1) *warble-rattle-twitter*; (2) *twitter-warble-twitter*; (3) *warble-rattle-tweet*; (4) *warble-trill-tweet*; and (5) *twitter-rattle-tweet*. First, note that *warble* and *twitter* do *not* share all suffixes in common, since in sequence (4) *warble* can be followed by *trill tweet*, but there is no similar suffix for *twitter* – the sequence *twitter-trill-tweet* is *not* part of the song repertoire. Thus, the song language is *not* 0-reversible. However, the language *is* 1-reversible. To test this, we observe which single syllables are held in common between *warble* and *twitter*. There is one such case, for the syllable *rattle*, in sequences (3) and (5), where we have *warble-rattle-tweet* and *twitter-rattle-tweet*. Since in both such sequences (3) and (5) share all suffixes past *rattle* in common, namely, *tweet*, the 1-syllable "window" test is met, and the language is 1-reversible. The extra syllable *warble* makes all the difference. From a learnability stand point, if such a constraint holds for some relatively small value of *k*, then the resulting song is easy to learn just by listening to song examples, as Sasahara et al. (2006) have shown by a direct computer implementation, with *k* at most three.

Taken together then, all these results so far point a single conclusion: birdsong is more closely analogous to human speech than human language syntax. Even so, one must be cautious here as well, because even human speech and birdsong are different from one another in certain respects – unsurprisingly, birdsong is *song*, and human speech does not have all the aspects of song; Fitch (2006) has a thorough review of this comparison. In particular, both birdsong and human songs include as essential aspects both explicit *pauses* and *repetition* – as was noted in the discussion of chaffinch song. One need only bring to mind any Mozart aria to recognize that in human song, even when accompanied by words, pauses, and repetition play a key role in the music itself. This is not typically the case in human speech or language. In language, pauses can indeed be found as part of the descriptive prosodics of an utterance, as in the brief pause after a stressed focal item, as indicated by commas. But pauses are not integrated into the acoustic speech stream in the same essential way as in music, where specific numbers of pauses and pauses of particular lengths *must* occur in certain places, as is clear from musical notation. Repetition is also found in human language, but also strictly delimited, for example, the type that linguists call "reduplication," the repeated occurrence of particular words or morphemes, often indicating some manner of intensification, as in, *very*, *very*, *cold* (see, e.g., Marantz, 1982). Like pauses, the role of particular repetitions in human language is much more limited than in song, where entire phrasal units are deliberately repeated. Putting aside the lack of words, the analogy between human song and birdsong seems in fact extremely close. All things considered, birdsong might serve best as a comparative model for human song, and secondarily for human speech, encompassing vocal learning and vocal production.

HUMAN LANGUAGE AND BIRDSONG: THE KEY DIFFERENCES

As outlined in the Introduction, in human language, hierarchical grouping is also accompanied by additional properties not found in birdsong or human sound systems. Let us revisit these, and then see in Section "A Model for Human Language Syntax" how they might be modeled by a single, very simple combinatorial operation. For reference, **Table 1** brings together in one place the birdsong–human language comparative results described in this section and the article as a whole.

First, human language admits indefinitely extendible, asymmetric containment relationships with at least *two* (and generally more) distinct types of labels. A sentence-like *John knows the starlings* can contain another sentence, as in *John knows the starlings will eat the apples*. Even this possibility does not seem to arise in sound systems, where legal motifs (in the case of birdsong) or legal consonant–vowel possibilities (in the case of language) do not form whole units that are in turn further contained within one other, e.g., we do not find human consonant–vowel structures in the form, (CV(CV(CV))), with the parentheses demarcating the containment of the leftmost consonant–vowel component inside two others.

The multiple types of phrases derive from a second property of natural language structure not found in birdsong, and that is *labeling* dependent on *word features*. The phrase *ate the apples* has the properties of a particular component based on the features a just one lexical item, the verb *eat* (surfacing as *ate* in its past tense form). Note that while it is logically possible to fix the properties of a Verb Phrase in some other way – say, by using the properties of the Noun Phrase *the apples*, or by somehow combining the properties of *ate* and *the apples*, that is not the way human syntactic machinery seems to operate. For this reason, the phrase *ate the apples* is conventionally called a Verb Phrase (VP; rather than a noun-like phrase or something in between). We can say informally that the phrase is *labeled* by *selecting* the verb and certain of the verb's features, and this how the phrase inherits "verb-like" properties. Similarly, a phrase like *the apples* is conventionally called a Noun Phrase (NP). Here we will simply assume informally that the features for the label of this kind of phrase are drawn from some properties of the noun *apples*, namely, that is a noun.

Using the conventional notational system devised by linguists, we can write out the hierarchical structural description for *eat the apples* in a bracketed notation in lieu of the graphical description of **Figure 1A**, where the opening and closing square brackets with labels indicate the extent of a phrase:

(3) $[_{\mathrm{VP}}$ *eat* $[_{\mathrm{NP}}$ *the apples*$]_{\mathrm{NP}}]_{\mathrm{VP}}$

Since the VP label is a simply an arbitrary gloss for particular properties of *eat*, we may replace it with the label *eat** where *eat** denotes these verbal features, whatever they might be. We can do the same for the Noun Phrase, or NP. We further suppress the label on the closing right brackets for readability, arriving at this representation for the syntactic structure corresponding to the sentence:

(4) [$_{eat}$* *eat* [$_{apples}$* *the apples*]]

We now recall that this bracketing structure is different from a linear word or sequence pattern, as in a consonant–vowel combination or a birdsong motif. The key difference is the use of a verb or noun's features to label an *entire* word sequence with a single label, in our gloss, *eat**, or *apples**. As we described in the Introduction, the selection of a privileged element in this way renders the underlying structure fundamentally asymmetric. Note that there is no analog to this in birdsong, a second key difference with human language. Consider as an example the birdsong motif described earlier, consisting of seven particular syllables. This motif is not "labeled" by selecting just one of these syllables and its properties to name the entire motif; none of the syllables takes priority in the same way that *eat* does in the human language example. Neither is the resulting structure asymmetric as it is in human language. This is true precisely because birds apparently do not have words or manipulate word features at all. This is one difference between the human language syntactic system and birdsong. We noted earlier that this does not in principle bar the possibility of birdsong making use of features of song elements, for example, syllables and their acoustic features, and assembling them in a similar hierarchical fashion. However, current evidence suggests that this does not occur in birdsong. Rather, the combinatorial operator itself is absent.

A third difference between human language and birdsong also follows. Once a labeled phrase is available to the human language syntactic engine, it can enter into additional syntactic manipulations as a new, single unit, as if it were a single word. So for example, once having established *eat the apples* as "chunk" *eat**, the human language system uses *eat** as a single verb-like object to build forms such as, *the starlings will eat**, i.e., *the starlings will eat the apples.* More interestingly, even more complex examples with *eat** can be constructed, such as, *the starlings will eat the apples and eat the apples the starlings did*, where *eat the apples* is understood as occurring in at least three different places: (1) at the start of the sentence; (2) after *and*; and, more interestingly, (3) in an unpronounced (phonologically null) "understood" form after *did* that is interpreted in exactly the same way as if *eat the apples* was actually present after *did.* More precisely, one can say that *eat the apples* is in fact present in the *syntactic structure* following *did*, does not surface phonologically – that is, it is not spoken or signed. This happens when the internalized syntactic form must be externalized; the third occurrence of *eat the apples* is suppressed and remains unrealized as part of the sound/manually signed stream.

This last example is quite characteristic of human language as we shall see with additional examples. However, it is absent from birdsong, where there are no "unpronounced" song components, virtually by definition. If we let *eat** denote the label of the entire phrase, *eat the apples,* then we can write out the syntactic structure of this last example as follows, where *S* denotes a sentence, and we have suppressed irrelevant details, like the analysis of *will* and *did*, that carry tense:

(5) [$_S$ [$_{starlings}$* *the starlings*] [*will eat**] *and* [$_S$ *eat** [$_{starlings}$* *the starlings*] [*did eat**]]]

We can see in this example that the syntactic structure has encoded a dependency between these three occurrences of *eat**: they are in effect linked copies, in the sense that they refer to the same syntactic object, *eat the apples,* but the copies appear in several different positions in the sentence. In the same way, given the sentence, *the starlings ate the apples*, if we label the phrase *the apples* as, *apples**, then one can form a sentences such as, *the apples the starlings ate*, which is interpreted as, *apples** *the starlings ate apples**. In this case, *the* apples is interpreted in two positions. The first position is at the front of the sentence, corresponding to its role as the so-called "topic" or "focus" of the sentence (which carries a special intonation peak, as the comma indicates). The second position is as the Noun Phrase adjacent to the verb *ate*, corresponding to the status of *the apples* as the Object of the verb, just as in the simple declarative sentence, *the starlings ate the apples.* The fact that one and the same phrase can be, indeed must be, interpreted in two distinct places in a sentence, one associated with discourse factors, and the second, with semantic interpretation as the argument of a predicate, is yet another wide-spread phenomenon in human language, absent in birdsong. This kind of "displacement" of phrases, no matter how it is described, seems nearly ubiquitous in human language, as most standard linguistic accounts note (see, e.g., Radford, 1997).

Birdsong, even when described via "chunks" that might correspond to phrases, does not seem to have *any* of these additional distinctive properties of human language. Let us see in detail why not. If a birdsong motif is made up of, say, two syllable sounds, we do not find that the features of one of these syllables is differentially selected to characterize the motif as whole. This would amount to a representation something like the following, where "warble" and "rattle" are presume to be two distinct birdsong motifs, along the lines of *eat the apples*:

(6) [$_{warble}$* *warble-rattle*]

However, nothing like this seems to be found in birdsong. Nor do we find the embedding of one motif inside another, or the embedding of two *different* kinds of phrases within one another, like a Sentence within a Noun Phrase. Finally, we do not find examples like *eat the starlings did eat* [*the apples*], with unpronounced, but elliptically understood syllabic chunks. In short, *none* of these distinctive properties of human language that move it beyond the domain of simple linear sequencing seem to be found in birdsong.

To indicate how crucial and wide-spread this difference is, we will describe several more such examples of what we might call "copy dependencies," all absent in birdsong, but present in human language syntax. First consider the example below, where there are *two* positions that seem to contain unpronounced "copies" of a Noun Phrase:

(7) *this is the bird that the starlings saw without leaving*

This is a rather more complex sentence. In this example, *the bird* serves as a phonologically suppressed copy in two places: it is the Object of *saw* and it is the Object of *leave.* We can "reconstruct" the unpronounced form *the bird* in these two positions to recover the required structure for proper semantic interpretation (though the sentence sounds more unnatural after this reconstruction):

(8) *this is the bird that the starlings saw the bird without leaving the bird*

Second, one can observe that there are examples where multiple dependencies can be "nested," precisely as in our example of **Figure 1E**, corresponding to the sentence, *the rat chased the birds that saw the cat that ate the starling*. Referring back to that figure, note that we had left empty parentheses for the Noun Phrases that serve as the Subject of both *saw* and *ate*. It is clearly the case that *the birds* is the Subject of *saw*, (e.g., it is the birds that saw the cat) and *the cat* is the Subject of *ate*. We can now clarify that these positions actually represent the same sort of unpronounced, phonologically null instances as in our earlier examples, so that the corresponding Noun Phrase in each case may be replaced in the now reconstructed syntactic form, again so that proper semantic interpretation can proceed. That is, the reconstructed sentence is something like the following, using now the labeled bracket notation:

(9) [$_{S}$ [$_{NP}$ *the rat*] [$_{VP}$ [$_{V}$ *chased*] [$_{NP}$ [$_{NP}$ *the birds*] [$_{S}$ *that* [$_{S}$ [$_{NP}$ *the birds*] [$_{VP}$ [$_{V}$ *saw*][$_{NP}$ [$_{NP}$ *the cat*] [$_{S}$ *that* [$_{S}$ [$_{NP}$ *the cat*] [$_{VP}$ [$_{V}$ *ate*] [$_{NP}$ *the starlings*]]]]]]]]]]]

Again referring to **Figure 1E**, it is evident that we now have at least *two* sets of dependencies: between *the cat* and its unpronounced position; and between *the rat* and its unpronounced position. Furthermore, crucially these dependencies are "nested" and could be arbitrarily extended in the manner discussed in Section "Birdsong Seems Analogous to Speech, Not Syntax." As we have seen, these kinds of patterns cannot be captured by any finite-state transition network.

As a final example of a more complex dependency found in human language syntax, there are examples that involve what are called *crossed-serial dependencies* as opposed to *nested* dependencies. Such dependencies are called *crossed* because the relationship between the elements overlap rather than nest (see **Figure 4** for an example). These are evidently less common in human language sentence syntax. Among the first examples were described by Huybregts (1984) in certain Dutch and Germanic dialects. But even in English, examples of such dependencies can be found in circumscribed contexts. The classic example was provided by Chomsky (1957), to account for the sequence of English auxiliary verbs and their morphological endings indicating aspects of tense, such as the passive or perfective endings of a verb. Chomsky noted that the apparent underlying syntactic form of the auxiliary verb sequence that is pronounced as, e.g., *will have been being eaten* is best described by the following context-free rule. (See Lasnik, 2000 for additional discussion of this point; we roughly follow Lasnik's discussion and notation below).

(10) Verbal "motif" → Tense-element Modal-verb-∅ *have en be ing be en eat*

That is, to generate English verb auxiliary sequences, the *en* suffix that follows *eat*, to form *eaten*, indicating a passive inflection, is actually attached to the preceding auxiliary verb *be*. Similarly, the suffix *ing* following the last *be*, to form *being*, indicating progressive tense, is actually attached to the *be* that precedes it. The pattern continues all the way through: the suffix *en* that follows the first occurrence of *be*, forming *being*, is actually attached to *have*; and, finally, the "zero" suffix after *have* is actually attached to the modal-verb *will*. If we then draw out these relationships, as shown in **Figure 4**, it is clear that the dependencies between the elements follow a crossed-serial pattern.

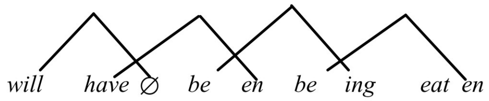

FIGURE 4 | The "crossed-serial" dependency structure of the English auxiliary verb system. Each auxiliary verb, *will, have, be,* and *be*, is linked to the suffix of the verb that follows. The zero element associated with *will* is a so-called "zero morpheme" – the suffix is present, but not pronounced.

Such examples have important computational implications: they require even more sophisticated networks (or grammars) than those we have yet described. Informally, this additional power amounts to having the individual memory locations in a pushdown stack themselves act like separate stacks, alternatively, to have a second, additional pushdown stack. A variety of related formalisms for processing such patterns have been proposed (see, e.g., Weir, 1988), under the name of *mildly context sensitive languages* (and their corresponding grammars). The second stack-like memory behavior is required because in the case of overlapping or crossing dependencies, one must be able to retrieve and insert whole phrases at positions other than those that occur in a last-in, first-out order.

Labeling-plus-grouping also imbues human language syntax with two final characteristic properties that do not appear to be found in birdsong. The infiltration of word features into language's syntactic system serves as a key "hook" between the conceptual atoms underpinning individual words and indefinitely large sentence structures, yielding the open-ended conceptual character of human language generally. This follows from the principle of compositionality grounded on syntactic structure, originally formulated by Frege, as noted by Fodor (1996): if one has separately acquired the words associated with *apples, bananas*, etc., along with the verbs *eat* and *want*, then the algebraic closure of the grouping and labeling operation implicit in forming *eat the apples* applied to this miniature lexicon yields the cross-product of the two possibilities, *eat apples*, . . ., *eat bananas, want apples*, . . .,*want bananas*. In short, we immediately obtain an account of the open-ended productivity of human language as a side-effect of syntactic combination, along with a link to conceptual productivity.

Summarizing, **Table 1** lays out all the comparisons between birdsong and human language that we have surveyed in this article. There are just two areas where birdsong and human language align; this is between birdsong and human language sound systems. All other areas differ. Considering these differences, along with the current evidence from analogous brain regions, to genomics, to vocal learning and production, to the extent that birdsong and, human sound systems are comparable, they align at one particular formal level, that of "input–output" externalization systems, including that of sensory–motor-driven vocal learning. While this is by no means an insignificant part of the entire language system, it is the part devoted only to externalization, and even then this does not address the modality-independence of human language, which can be non-acoustic in the case of manual signing. In this sense, a large gap remains between human language syntax proper and birdsong.

In the next section we outline how a human-capable syntactic system, evidently then quite different from that of birdsong, can be modeled.

A MODEL FOR HUMAN LANGUAGE SYNTAX

All the distinguishing properties of human language listed in **Table 1** can be modeled within a simple computational system, first developed in Chomsky (1995) and in many subsequent places by others. The model described below is not intended to cover all aspects of human language syntax, but rather delimit a minimum set of assumptions, shared by many different linguistic theories, that account for the birdsong–human language syntax distinctions of **Table 1**.

First, whatever other characteristics such an algorithm must have, above all it must be able to associate unboundedly many strings of words with structured expressions, of the sort noted in **Table 1**. Familiar results from the study of computational systems since the latter part of the twentieth century have shown that any such system requires some kind of primitive combinatory operation that can construct larger objects from smaller ones, where the smaller objects may themselves be complex, but ultimately reduce to some set of atomic items, in our case, essentially words (Kleene, 1953). The required combinatory operation has been cast in many forms both in abstract computational systems and in specific generative proposals for describing human language syntax. For example, one such system is the Lambek (1958) calculus, in which individual words have properties corresponding to the "valences" of chemical theory, dictating how they allowably combine with other words or structures. For example, in this system, *ate* has the property (NP§)/NP, meaning that it requires an NP Subject to its left, and an NP Object to its right. Further, there is a single rule of combination that "glues together" two words or previously assembled structures into larger wholes, ultimately an entire sentence. For instance, given *ate* with its properties as above, and a corresponding NP to its right, associated with, say, *the starlings*, the Lambek combinatory operator takes as input these two items, (NP§)/NP and NP, and output a new structure, NP§, corresponding to a traditional Verb Phrase, with the "NP§" notation indicating that a Subject NP is still required to the left. See, e.g., Steedman (2000) for a broader and more recent treatment of human language syntax within an extended version of this framework.

In a similar spirit, here we will also assume a combinatorial operator that associates strings of words with structures, along with a rule of combination, though of a different sort, along the lines described by Chomsky (1995), where this combinatorial operator is called "merge." We follow Chomsky's presentation closely in what follows below, because along many dimensions it is relatively theory neutral, in the sense that it makes the fewest possible assumptions about the syntactic machinery needed to generate possible sentence structures – many current linguistic theories contain at their heart some sort of combinatorial operation similar to the one described here. Also like the Lambek calculus, the inputs X and Y are either individual lexical items, what we will also call *atomic* units, or else more complex syntactic objects previously constructed by application of the operator from such atomic units. Finally, also in accord with the Lambek system, we assume that the combinatorial operator can apply to its own output, that is, a previous application of the combinatory operation. Following the Church-Turing thesis, this is a requirement for yielding a system that can associate indefinitely many words with structured syntactic objects.

At this point, our model diverges from the Lambek system. Unlike the Lambek calculus, for computational simplicity, following Chomsky (1995), we will assume that X and Y are unchanged by the combinatorial operation, so we can represent the output simply as the *set* $\{X, Y\}$. Assuming otherwise would take additional computational effort. For example, if $X =$ the single lexical item *ate*, and $Y =$ the more complex syntactic object corresponding to the phrase *the starlings*, then the output from the operator given this X, Y input would simply be the set, $\{ate, Y\}$. Referring to more traditional notation, this particular set would correspond to what we earlier called a Verb Phrase, with X equal to the atomic item *ate*, and Y equal to the Noun Phrase Object associated with *the starlings*.

In further contrast with the Lambek system, note that this output set is by definition unordered, in this way reflecting the apparent lack of any syntactic predicates based on linear precedence. Any such order is assumed to be imposed the sound system of the language, which, for example, must determine whether verbs precede or follow Objects. (English chooses to externalize syntactic structure so that Objects follow Verbs, while in German or Japanese the choice might be otherwise.) Crucially, the operator can apply again to its own output, so generating a countable, discrete infinity of possible syntactic structures (Berwick, 2011).

We will introduce one more bit of machinery to describe this system, and that is the notion of *labeling*. Here too, as in the Lambek system, after the combinatorial operation has been applied, the newly created syntactic object has properties that are based on, but not exactly the same as, the properties of the objects out of which it has been constructed. In fact, in the Lambek system, the new composite object is some subset of just one of the features of the two objects that were joined together; in our example above for instance, the structure corresponding to a Verb Phrase, (NP§), obtains its properties from that of the verb, (NP§/NP). We will follow something along these lines, though somewhat distinct. With this notion in hand, now proceed to show how the syntactic structure for an entire sentence, *the birds ate the starlings*, might be generated.

Note that where we have Y identified with an entire phrase, *the starlings*, it must be the case that this syntactic object Y was itself constructed by some previous application of the operator. In particular, we must have applied it to the two lexical items, *the* and *starlings*, so obtaining Y. Recalling our earlier discussion where we described Verb Phrases as inheriting their properties from the properties of verbs, we need in addition a way to identify and label such newly minted sets. To do this we will assume that when applying the operator to two sets X and Y, we must always select the properties of just *one* of these to serve as the *label* for the set combination that results. So for example, when we apply the operator to two sets consisting of simply lexical items, say {*the*} and {*starlings*}, then we select one of these, here *starlings*, and write the actual combination as, {*label*, {*X*, *Y*}}. Following the lead of our example from the previous selection, we gloss these label features as starlings*. The output result is the more complex syntactic object in (11) below:

(11) Y = {starlings*, {{the}, {starlings}}}

It is this set that corresponds to the conventional notion of a Noun Phrase, though it is important to recall again the crucial difference that unlike a traditional Noun Phrase, there is no linear order imposed between *the* and *starlings*, which is of no concern to the internal syntactic system. The ordering between these two words is left for the phonological sound system to spell-out, when the phrase is actually pronounced.

With this elaboration in mind, when the operator is applied to inputs *X* = {*ate*}, and *Y* is as defined just above, the syntactic object corresponding to *the starlings*, one must again select a new label for the output of the combinatorial operation. In this case, we assume to be the label drawn from *X*, namely, ate*. (We leave to one side the question of why it is *X* rather than *Y* that is selected for fixing the label.) Following through with our example then, the operator applies to the two sets *X* and *Y*, yielding the more complex structure in (12):

(12) {ate*, {{ate}, {starlings*, {{the}, {starlings}}}}}

This set corresponds to a conventional Verb Phrase, though again without any linear precedence ordering between what would conventionally be called the Verb and the Object Noun Phrase. Finally, by using this set-structure along with the set-structure corresponding to the Subject Noun Phrase, e.g., *the birds*, we may apply the operator once again, outputting a final Sentence structure along these lines:

(13) {ate*, {birds*, {{the}, {birds}}}, {ate*, {{ate}, {starlings*, {{the}, {starlings}}}}}}

While this notation appears complex, it in fact contains all the hierarchical information needed to recover the Subject and Object relations, the adjacency of the Object NP with the verb, and in fact any other required syntactic relationships associated with the sentence. Let us see how this works out. Consider the required adjacency relationships. First, the Object must be adjacent to the verb. This is true in our output structure, because in the form:

(14) {{ate}, {starlings*, {{the}, {starlings}}}}

we can see that {*ate*} and {starlings*...} correspond to pairs {*X*, *Y*} at the same level in structure (13), and thus meet the correct notion of "adjacent to" required. Note that this property crucially follows because we have (tacitly) assumed that composition always takes two arguments. This is not a necessary property, but one that seems empirically sufficient, as noted in Chomsky (1995). Similarly, the Subject must be adjacent to the syntactic object that denotes the conventional Verb Phrase, and here too we find that the set construction properly describes this relationship:

(15) {birds*, {{the}, {birds}}}, {ate*, ...}

Here, {birds*, ...} and {ate*, ...} are pairs *X*, *Y* at the same level, and so adjacent to one another.

Turning to hierarchical relationships, the Verb–Object combination is set off as a phrase distinct from the Subject, in virtue of its *containment* within a subset of its own, apart from the one that contains the set associated with *the birds*:

(16) {ate*, {{ate}, {starlings*, {{the}, {starlings}}}}}

Further, the asymmetry of the set-structure is fixed by the very definition of labeling, since only one lexical item participates in determining a label's features.

What is the advantage of this alternative system? Importantly, such a system automatically admits the possibility of examples such as *the birds will eat the starlings and eat the starlings the birds did* [*eat the starlings*], because the combinatorial operator applies to *any* two sets *X*, *Y*, even when *Y* happens to be a subset of *X*. Suppose for instance that we have already constructed a (slightly different) Sentence along the lines of our other example sentence above, corresponding to the syntactic structure for *the birds will eat the starlings*, where we have suppressed certain of the linguistic details for expository purposes:

(18) {will*, {birds*, {{the}, {birds}}}, {will, {{will}, {eat*, {{eat}, {starlings*, {{the}, {starlings}}}}}}}}

Given the general combinatorial operator, one of its choices is to freely select to combine the *entire* set object above as its choice for *X*, along with any proper *subset* of this set as its second choice for *Y*, for example, {starlings*, {{the}, {starlings}}} (=11), corresponding to the Noun Phrase *the starlings.* Given this choice for *Y*, the output from the combinatorial operator acting on the pair *X*, *Y*, and selecting the label associated with *X* for the output structure, yields an apparent *copy* of the set representation for *the starlings* as suggested in Section "A Model for Human Language Syntax," shown below in (19). In other words, set-structure for *the starlings* now appears in two places: the first position, that of the "discourse focus"; and the second position as the argument of the verb predicate *eat*. In (19) we have highlighted these two occurrences in bold font.

(19) {will*, {{{**starlings*, {{the}, {starlings}}}**}, {will*, {birds*, {{the}, {birds}}}{will*, {{will}, {eat*, {{eat}, {**starlings*, {{the}, {starlings}}}**}}}}}}}}

When this syntactic structure is sent to the phonological system for output, the second occurrence is suppressed, which we indicate below by striking a line through it:

(20) {will*, {{{**starlings*, {{the}, {starlings}}}**}, {will*, {birds*, {{the}, {birds}}}{will*, {{will}, {eat*, {{eat}, ~~**starlings*, {{the}, {starlings}}}**~~}}}}}}}}

The sound system that "externalizes" this internal syntactic structure will as usual output only the actual words in brackets, not the labels or the rest of the syntactic form, imposing precedence relations, so that the output from the sound system surfaces as (21):

(21) *the starlings the birds will eat*

In this way, a single combinatorial operator, without any additional assumptions, automatically generates the syntactic structures described above in Section "A Model for Human Language Syntax," with copies that are present in at least two places, but that remain unpronounced when the internal syntactic form is mapped into its phonological counterpart and spoken (or manually signed). Furthermore, this way of constructing the form automatically ensures that the discourse prominent copy is hierarchically superior to the copy that serves as the verb's argument, as required. We do not have to specify some new, separate operation apart from the single combinatorial operator in order to generate structures with copies. This is part of the way the operator works with all syntactic objects.

A second major advantage of this system is that it can account for a wide range of syntactic phenomena within both English and across many dozens of other languages where there are apparent "pairings" between forms such as the following:

(22a) The starlings will eat the birds
(22b) Will the starlings eat the birds

Here, as discussed in Berwick and Chomsky (2011), the auxiliary verb *will* in the question form (22b) must be interpreted in the same position as it is the declarative form (22a), in order that the close link between the sense of (22a) and (22b) can be maintained. (The second is the interrogative form of the first.) This property is automatically accommodated under a model where (22b) is formed by the combinatorial operation acting on *will* as a subset of the larger set-structure corresponding to *will eat the birds.* Just as before, an apparent copy of *will* is placed at the end of the sentence, with *will* remaining in its "original" position, where we have inserted brackets to highlight the phrasal and subset-set relationships:

(23) [Will] the starlings [[will] eat the birds]

Once again, when pronounced, the second occurrence of *will* is suppressed, and the sound system outputs the form (22b).

(24a) [Will] the starlings [~~[will]~~ eat the birds]
(24b) Will the starlings [eat the birds]

There is a large range of similar cases that have been investigated by linguists over the past 60 years covering many dozens of languages, all of which can be accounted for by the combinatorial operator posited above. This provides substantial empirical support for the particular assumptions we have made; see, e.g., Radford (1997) among other recent texts for details.

THE EVOLUTIONARY PICTURE

We have framed the "gap" between birdsong and human language in Sections "Human Language and Birdsong: The Key Differences" and "A Model for Human Language Syntax" in way that lends itself to two main evolutionary questions. The first concerns the combinatorial operator itself. Is this computational competence present in other species? If not, how did it arise? Does it have antecedents in terms of older or related competences? Can we break down the operator into smaller components, and use these to envision an evolutionary scenario such that the operator might have been acquired in distinct stages? The second question concerns the stock of atomic elements, the lexical items or words that feed the combinatorial operator. Several possible evolutionary scenarios have been envisioned regarding these puzzles, for the most part difficult to verify, given the absence of the relevant evidence.

We review just one position here: that in fact there is no such gap, and that songbirds (and other non-human species) actually possess the same syntactic combinatorial ability as humans, though lacking lexical items. To determine whether this is so, in recent years researchers have attempted to determine whether songbirds can succeed at artificial language learning tasks. Following the lead of experiments carried out with non-human primates (Fitch and Hauser, 2004), these approaches have most often attempted to probe whether songbirds can learn to discriminate the strings of languages that are not describable by any finite-state transition network. In particular, researchers have focused on artificial languages of the form $a_i f_i$, with any number of matching *a*'s and *f*s. Some experiments have added a distinguished center marker, *c*, yielding languages in the form, $a_i c f_i$. As noted earlier, such a language can only be successfully generated or recognized by a finite-state automaton if it is augmented with a single counter, or, equivalently, adding a pushdown stack with a single symbol. In this sense, it is perhaps the "simplest" example of a language that cannot be recognized by an unadorned finite-state transition network, as noted by Rogers and Pullum (2011). The general experimental methodology is to train subjects on a set of familiarization strings drawn from a language known to be non-regular, and then test the subjects to see if they correctly accept syntactically well-formed examples of the language, and reject syntactically ill-formed examples.

Given success in this task, the implication is that the subjects have acquired and then used a system of rules that go beyond the power of finite-state transition networks. Some researchers have in addition suggested that success in this task implies that the subjects have acquired and then used a particular *kind* of finite-state augmentation, either a rule system equivalent to a fully recursive transition network with a pushdown stack as described above, or, a rule system equivalent to this. It can be difficult to test such details about implementation, even in human subjects, as attested by the recent work by Uddén et al. (2011). Using an artificial language learning paradigm, they found experimental support for pushdown stack storage in human subjects to be lacking. However, they did find experimental evidence that crossed-serial dependencies required additional computational effort, in line with a full two-pushdown stack model mentioned earlier. However, it should be noted that as soon as one imputes a full two-pushdown stack system to a computational device, then this computational machinery is as powerful as any general-purpose computer, i.e., it is as powerful as a Turing machine. It remains unclear how such a device is actually implemented in the brain.

In any case, the experiments attempting to demonstrate that non-human species have a cognitive competence that could be emulated by even a single pushdown stack have so far proved inconclusive. In the first experiment to report an apparent success with this kind of protocol in any bird species, Gentner et al. (2006) used operant conditioning to train and then test starlings on an artificial language defined over an alphabet of acoustically distinct *whistle*, *warble*, *rattle*, and *high-frequency* motifs, drawn from the song of one male starling. Eight distinct *warble* and *rattle* motifs were used to formulate a training language consisting of four-syllable strings in the form, *rattle*$_i$-*rattle*$_j$-*warble*$_l$-*warble*$_j$, with i, j, k, l ranging from 1 to 8, corresponding to a sample from the "correct" target language in the form, $a^i f^i$. This was used for positive-reinforcement operant conditioning. The starlings were also trained to avoid syntactically ill-formed strings of the form, *rattle*$_i$-*warble*$_j$-*rattle*$_l$-*warble*$_j$, corresponding to the language $(af)^i$, a language that can be generated by a finite-state transition network. After many thousands of positive and negative reinforcement trials, the birds were then probed with different novel correct and incorrect sequences, including longer length-6 and length 8-strings, and responded positively to the correct strings while also properly avoiding the incorrect ones. Can one therefore conclude that starlings can acquire and use the rules for hierarchical structures along the lines of human languages?

The answer seems to be no, for at least two reasons. First, the language that was used to exemplify the use of a finite-state transition network with recursive subroutines, alternatively a context-free grammar, was not in fact of the right type to *unambiguously* demonstrate the conclusion that was sought, as noted by Corballis (2007), Friederici and Brauer (2009), and Friederici et al. (2011), among others. Above and in **Figure 3D** we indicated that in such a language the *a*'s and *f*'s must be nested and paired with each other from the inside-out. But this was not true of the artificial language in the Gentner et al. (2006) experiment, where the *warbles* and *rattles* could be of *different* types, not necessarily paired with each other. That is, instead of the language containing strings such as, $a_1 a_2 f_1 f_2$, the starlings were trained on strings including, $a_1 a_2 f_2 f_3$, with the critical nesting property violated. As a result, all that is required for the starlings to succeed on novel, well-formed probe stimuli is that the birds be able to count that the number of *warbles* is followed by the same number of *rattles.* This can be done by a finite-state network with a single, limited counter – that is, all the birds are required to do is to count – subitize – up to this numerical limit, an ability that has already been attested in this species by Dehaene (1997). It is therefore more parsimonious to assume that the birds are simply drawing on abilities that have already been demonstrated, rather than some novel cognitive ability. Second, as suggested by Rogers and Pullum (2011), testing that the starlings reject the "illicit" strings of length six, e.g., *warble-rattle-warble-rattle-warble-rattle* is confounded with the possibility that such strings can also be generated by a non-finite-state transition network, one that tests, in general, whether the length of all the *a*'s is the same as the length of all the *f*s; this is not a language that can be generated by a finite-state transition network.

In part to remedy the first problem, more recently, Abe and Watanabe (2011) carried out an artificial language learning experiment with Bengalese finches, *Lonchura striata* var. *domestica*, concluding that the finches acquired and then used context-free grammar rules for language discrimination. Watanabe and Abe exposed finches to training examples of distinctive song syllables. Birds were exposed to two sets of familiarization strings (denoted FAM), $a_x\ c_z\ f_x$ ("non-embedded strings") and $a_x\ a_y\ c_z\ f_y\ f_x$ ("center-embedded strings," or "CES"), where the letters denote syllables, and matching subscript letters denote matching syllables that always co-occur in a string. These sequences were designed to follow the possible patterns generated by a context-free grammar with the syllables of similar types properly paired. One other difference from the Gentner training language was that in each pattern, the c_k syllable type marked the middle of a legitimate pattern. We can write out an example CES string $a_1\ a_2\ c_3\ f_2\ f_1$ to display the implied containment relationships using bracketing notation as follows, where we have arbitrarily labeled the left-hand square brackets with S_1, S_2, and S_3.

$$[S_1 a_1\ [S_2 a_2\ [S_3 c_1]\ f_2]\ f_1]$$

Watanabe and Abe improved on the Gentner protocol in at least one other respect: no operant conditioning was needed, as the birds' natural calling behavior was used as a response measure.

The finches were then tested to see whether they would reject ill-formed examples such as $a_2 a_1 c_1 f_2 f_1$ (where the order of syllables does not follow the proper nested containment pattern); reject examples like $a_1 f_2 a_2 c_1 f_2 f_1$, where an *f* precedes the *c* marker; and accept as well-formed novel examples such as $a_2 a_1 c_3 f_1 f_2$, where the particular pattern with the center marker c_3 was not part of their training set. The litmus test for recognition (conversely, rejection or non-recognition) was a measurable increase (conversely, a decrease) in calling rate response to the test examples. The finches did vary their calling rates as predicted: calling rates were higher for syntactically correct syllable strings, as opposed to syntactically incorrect syllable strings. At first glance then, this result would seem to confirm that the finches had acquired the syntactic rules for generating nested hierarchical structure, since both the recognition and rejection tasks that would seem to require the grouping of syllables in a nested, hierarchical way.

However, the conclusion that the birds were actually constructing hierarchical representations remains arguable (Beckers et al., 2012). The training and test stimuli were not balanced for acoustic similarity. For example, the correctly center-embedded syllable test strings (COR), e.g., $a_1 a_2 c_1 f_2 f_1$, were largely similar to the familiarization strings, e.g., $a_1 a_2 c_2 f_2 f_1$, $a_1 a_2 c_3 f_2 f_1$, and $a_1 a_2 c_4 f_2 f_1$, both in terms of syllable positions and acoustically, mismatching on just a single syllable, the distinguished center syllable c_i. Thus, even though all COR strings are novel, four out of five syllable positions match in the sense that they contain the same sounds at the same positions. The other group of syntactically correct test strings had a similar problem. This means that the birds could have treated these novel test strings as familiar simply on the basis of their phonetic characteristics alone, without every analyzing their syntactic structure. Since it is already known that Bengalese finches can distinguish a set of familiarization strings as in the Watanabe and Abe experiment as distinct in terms of memorization alone (Ikebuchi and Okanoya, 2000), by this argument we do not need to posit any novel syntactic ability for finches, a more parsimonious explanation for the finches' behavior since it requires fewer assumptions.

Even if similarity matching had been controlled for by using a different experimental design that eliminated the overlap between familiarization and test strings, the sound sequences and presumed underlying structures in this experiment are unnatural in the sense that they are perfectly *symmetric*: there are an equal number of *a*'s, *f*'s, etc. to be matched up on either side of a distinguished center marker. This kind of structure quite unlike the *asymmetric* structures found in natural language, illustrated in **Figures 1A,B**. In fact, even humans have great difficulty mastering artificial languages whose underlying structures are symmetric. At least since the work of Miller and Isard (1964) it has been known that people have great difficulty parsing both naturally occurring self-embedded sentences as well as center-embedded sentences constructed in artificial language learning experiments (Miller and Chomsky, 1963). Confirming this, as briefly described in the introduction it has also long been known that people restructure sentences so as to avoid producing complex center-embedded structures, as well spontaneously using alternative strategies for solving tasks that would otherwise provide evidence for the processing of such structures (Langendoen, 1975, Perruchet and Rey, 2005).

Given the experimental evidence that there is a computational gap in processing ability that reflects a difference between songbirds and humans, then one way to express this distinction is in the way memory is organized. The ability to assemble sequences into groups with distinguishable labels like "VP" (or more carefully, *eat**) and then set them aside for additional later processing suggests the existence of memory locations where these newly assembled units like "VP" might be located and then re-used. At this point, how such memory might be organized remains challenging to discern, given current information about how neuronal structure might "implement" the computational architectures computer scientists are familiar with, as we review briefly below; for a recent discussion and speculations as to several possible "implementations," see Uddén et al. (2011). One problem is that once one moves to a machine with two stacks, one can easily show (Minsky, 1967) that an equivalent computational power can be attained by a machine with just two counters (see also Sipser, 1997 for a good introduction to these issues and automata theory). Such an abstract device could have many possible physical realizations, and at present the empirical evidence under-constrains these.

Pushdown stack storage is sometimes assumed to be implemented by, for example, a set of decaying neural networks, with each offset in a net's decay time corresponding to a different stack location (for more details see, e.g., Pulvermüller, 1993; Pulvermüller and Knoblauch, 2009). Alternatively, Uddén et al. (2011) suggest that arithmetical operations could be used to simulate stack-like operations – one could use a number that grows or shrinks in size, which as they observe might have some more straightforward realization in neural "wetware" than decaying or reverberating circuits. But all such statements should be treated with some caution, because there are many ways of implementing computational devices, particularly if memory access is carried out not according to some symbolic addressing scheme, as in conventional digital computers, but in terms of so-called content-addressable memory. Although algorithms for content-addressable memory are less well-known, even here, hierarchical representations can be readily developed, as described by, e.g., Oldfield et al. (1987). Thus it would be simply incorrect to state by fiat that a content-addressable memory would not be compatible with the efficient storage and manipulation of hierarchical or "tree" structures, even of a fairly complex sort. In any case, from the earliest studies carried by Bever (1970) and Chomsky and Miller (1963), and as further described by Berwick and Weinberg (1985), syntactic structures seem to form locally coherent trees that are then rapidly dispatched for semantic interpretation, so lessening any short-term, local memory load. It remains to explore how a system more faithful to what we know about neuronal structure and operation actually would work to implement this kind of abstract computation.

Earlier we noted that Hurford (2011) suggests that in songbirds, some arrangement of links between the HVC and RA nuclei encode phrase structure rules, but that this analysis is flawed and cannot actually distinguish between finite-state transition network and augmented transition networks. As for non-human antecedents or possible pre-adaptations for pushdown stack storage, it has been proposed that the requirement for this kind of auxiliary storage may have been driven by the requirements for animal navigation such as map following or foraging (Bartlett and Kazakov, 2005; Okanoya, 2007; Shettleworth, 2010). According to these proposals, if animals must remember the particular places where food has been cached, in some particular order, then this might require landmarks to be traversed in the manner of a pushdown stack to retrieve food left at these locations, starting with the most recently visited location first, then the second-to-last visited location, and so forth. As stated, this would amount to a symmetrical visiting pattern, like the embedded strings tested in the Gentner et al. (2006) experiment, in the pattern $a_2a_1c_3f_1f_2$. While a pushdown stack memory would seem of obvious benefits here, such suggestions have again remained at the level of simulation.

CONCLUSION

What can we conclude about birdsong, human language, and evolution? Birdsong seems most comparable to the sound system of human languages, that is, the externalization of human language proper, encompassing both speech and the manual modality of signed language. This comparison seems to hold both at a formal level of analysis, that is, systems compared as a sequence of acoustic (or gestured) strings, as well as at many distinct levels of neurological analysis, from brain regions down to the genomic level. Given the long divergence time between the *Aves* and human last common ancestor, some of these similarities may well be analogous, that is, the result of convergent function, rather than homologous, that is, the result of shared ancestry. It remains to be seen whether this remains a sensible conclusion given ongoing research uncovering so-called "deep homologies" among distant ancestors; however, the patchy existence of vocal learning, imitation, and production abilities in both birds and the (far sparser set of) primates suggests that many birdsong-speech commonalities fall more into the class of convergent evolution, while similarities in brain regions, categorical and prosodic perception, and the like may prove to be homologous. While birdsong syntax is more complex than simple bigrams, unlike human language syntax, it does not appear to go beyond languages describable by a narrow class of easily learnable finite-state transition networks. More broadly still, birdsong lacks nearly all the chief attributes of human language: it does not match the syntactic complexity of human language, without the multiple-label, and often asymmetric containment relationships of human language; it does not consist of phrases grounded on the conceptual atoms we call words; and, without words, it does not possess a compositional semantics driven by a full-fledged combinatorial syntax. Nevertheless, as far as a model for human speech acquisition and production, birdsong remains a useful model for the analyzing the evolution of a still-complex interplay between brain and behavior.

ACKNOWLEDGMENTS

Johan J. Bolhuis is funded by Netherlands Organization for Scientific Research (NWO) grants (ALW Open Competition and GW Horizon Programme). Publication of this paper was supported by an NWO Incentive Fund Open Access grant.

REFERENCES

Abe, K., and Watanabe, D. (2011). Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nat. Neurosci.* 14, 1067–1074.

Angluin, D. (1982). Inference of reversible languages. *J. Assoc. Comput. Mach.* 29, 741–765.

Aristotle. (1984). *Historia Animalium*, trans. J. Peck. Cambridge, MA: Harvard University Press [Loeb Classical Library].

Arregui, A., Clifton, C., Frazier, L., and Moulton, K. (2006). Processing elided verb phrases with flawed antecedents. *J. Mem. Lang.* 55, 232–246.

Bartlett, M., and Kazakov, D. (2005). The origins of syntax: from navigation to language. *Conn. Sci.* 17, 271–288.

Barton, E., Ristad, E., and Berwick, R. (1987). *Computational Complexity and Natural Language.* Cambridge, MA: MIT Press.

Beckers, G. J. L., Bolhuis, J. J., Okanoyoa, K., and Berwick, R. (2012). Birdsong neurolinguistics: songbird context-free grammar claim is premature. *Neuroreport* 23, 139–145.

Berwick, R. (1982). *Locality Principles and the Acquisition of Syntactic Knowledge.* Cambridge: MIT Doctoral Dissertation, MIT.

Berwick, R. (1985). *Locality Principles and the Acquisition of Syntactic Knowledge.* Cambridge: MIT Press.

Berwick, R. (2011). "All you need is merge: biology, computation, and language from the bottom-up," in *The Biolinguistic Enterprise*, eds A. M. Di Sciullo and C. Boecxk (Oxford: Oxford University Press), 706–825.

Berwick, R., Beckers, G. J. L., Okanoyoa, K., and Bolhuis, J. J. (2011a). Songs to syntax: the linguistics of birdsong. *Trends Cogn. Sci.* 15, 113–121.

Berwick, R., and Chomsky, N. (2011). "Biolinguistics: the current state of its development," in *The Biolinguistic Enterprise*, eds A. M. Di Sciullo and C. Boeckx (Oxford: Oxford University Press).

Berwick, R., Pietroski, P., Yankama, B., and Chomsky, N. (2011b). Poverty of the stimulus revisited. *Cogn. Sci.* 35, 1207–1242.

Berwick, R., and Pilato, S. (1987). Learning syntax by automata induction. *J. Mach. Learn. Res.* 2, 9–38.

Berwick, R., and Weinberg, A. (1985). *The Grammatical Basis of Linguistic Performance.* Cambridge, MA: MIT Press.

Bever, T. (1970). "The cognitive basis of linguistic structures," in *Cognition and the Development of Language*, ed. J. R. Hayes (New York: Wiley), 279–362.

Bijeljac-Babic, R., Bertoncini, J., and Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Dev. Psychol.* 29, 711–721.

Bloomfield, L. (1933). *Language.* New York: Holt.

Bolhuis, J. J., and Eda-Fujiwara, H. (2003). Bird brains and songs: neural mechanisms of birdsong perception and memory. *Anim. Biol.* 53, 129–145.

Bolhuis, J. J., and Eda-Fujiwara, H. (2010). Birdsong and the brain: the syntax of memory. *Neuroreport* 21, 395–398.

Bolhuis, J. J., and Gahr, M. (2006). Neural mechanisms of birdsong memory. *Nat. Rev. Neurosci.* 7, 347–357.

Bolhuis, J. J., Okanoya, K., and Scharff, C. (2010). Twitter evolution: converging mechanisms in birdsong and human speech. *Nat. Rev. Neurosci.* 11, 747–759.

Bolker, J., and Raff, R. (1996). Developmental genetics and traditional homology. *Bioessays* 18, 489–494.

Brainard, M. S., and Doupe, A. J. (2000). Interruption of a basal ganglia-forebrain circuit prevents plasticity of learned vocalizations. *Nature* 404, 762–766.

Brainard, M. S., and Doupe, A. J. (2002). What songbirds teach us about learning. *Nature* 417, 351–358.

Brauer, J., Anwander, A., and Friederici, A. (2011). Neuroanatomical prerequisites for language functions in the maturing brain. *Cereb. Cortex* 21, 459–466.

Chater, N., and Christiansen, M. H. (2010). Language acquisition meets language evolution. *Cogn. Sci.* 34, 1131–1157.

Chomsky, N. (1955). *The Logical Structure of Linguistic Theory.* Chicago: University of Chicago Press.

Chomsky, N. (1956). Three models for the description of language. *IEEE Transactions on Information Theory* 2, 113–124.

Chomsky, N. (1957). *Syntactic Structures.* The Hague: Mouton.

Chomsky, N. (1963). "Formal properties of grammar," in *Handbook of Mathematical Psychology*, eds D. Luce, R. Bush, and E. Galanter (New York: Wiley), 323–418.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Chomsky, N. (1995). *The Minimalist Program.* Cambridge, MA: MIT Press.

Chomsky, N., and Miller, G. (1963). "Finitary models of language users," in *Handbook of Mathematical Psychology*, eds R. Luce, R. Bush, and E. Galanter (New York: Wiley), 419–491.

Corballis, M. C. (2007). Recursion, language and starlings. *Cogn. Sci.* 31, 697–704.

Crain, S., and Nakayama, M. (1987). Structure dependence in question formation. *Language* 62, 522–543.

Culicover, P., and Jackendoff, R. (2005). *Simpler Syntax.* Oxford: Oxford University Press.

De Marcken, C. (1995). "Lexical heads, phrase structure and the induction of grammar," in *Proceedings of the Third Workshop on Very Large Corpora*, eds D. Yarowsky and K. Church (Cambridge: Association for Computational Linguistics), 14–26.

De Marcken, C. (1996). *Unsupervised Language Acquisition.* Ph.D. dissertation, MIT, Cambridge.

Dehaene, S. (1997). *The Number Sense.* Oxford: Oxford University Press.

Dehaene-Lambertz, G., Hertz-Pannier, L., Dubois, J., Mériaux, S., Roche, A., Sigman, M., and Dehaene, S. (2006). Functional organization of perisylvian activation during presentation of sentences in preverbal infants. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14240–14245.

Di Sciullo, A. M. (2003). *Asymmetry in Grammar: Syntax and Semantics.* New York: John Benjamins.

Doupe, A. J., Perkel, D. J., Reiner, A., and Stern, E. A. (2005). Birdbrains could teach basal ganglia research a new song. *Trends Neurosci.* 28, 353–363.

Dulai, K., von Dornum, M., Mollon, J. D., and Hunt, D. M. (1999). The evolution of trichromatic colour vision by opsin gene duplication. *Genome Res.* 9, 629–638.

Elemans, C. P. H., Mead, A. F., Jakobson, L., and Ratcliffe, J. M. (2011). Superfast muscles set maximum call rate in echolocating bats. *Science* 333, 1885–1888.

Fee, M., Kozhevnikov, A., and Hahnloser, R. H. (2004). Neural mechanisms of vocal sequence generation in the songbird. *Ann. N. Y. Acad. Sci.* 1016, 153–170.

Feher, O., Wang, H., Saar, S., Mitra, P. P., and Tchernichovski, O. (2009). De novo establishment of wild-type song culture in the zebra finch. *Nature* 459, 564–568.

Fisher, S. E., and Scharff, C. (2009). FOXP2 as a molecular window into speech and language. *Trends Genet.* 25, 166–177.

Fisher, S. E., Vargha-Khadem, F., Watkins, K. E., Monaco, A. P., and Pembrey, M. E. (1998). Localisation of a gene implicated in a severe speech and language disorder. *Nat. Genet.* 18, 168–170.

Fitch, W. T. (2005). The evolution of language: a comparative review. *Biol. Philos.* 20, 193–203.

Fitch, W. T. (2006). The biology and evolution of music: a comparative perspective. *Cognition* 100, 173–215.

Fitch, W. T. (2010). *The Evolution of Language*. Cambridge: Cambridge University Press.

Fitch, W. T. (2011). "'Deep homology' in the biology and evolution of Language," in *The Biolinguistic Enterprise: New Perspectives on the Evolution and Nature of the Human Language Faculty*, eds A. M. Di Sciullo and C. Boeckx (Oxford: Oxford University Press), 135–166.

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

Fodor, J. A. (1996). The pet fish and the red herring: why concepts aren't prototypes, *Cognition* 58, 243–276.

Fong, S., and Di Sciullo, A.-M. (2005). "Morpho-syntax parsing," in *UG and External Systems*, eds A.M. Di Sciullo and R. Delmonte (Amsterdam: John Benjamins), 247–268.

Friederici, A., Bahlmann, J., Friederich, R., and Makuuchi, M. (2011). The neural basis of recursion and complex syntactic hierarchy. *Biolinguistics* 5, 87–104.

Friederici, A. D., and Brauer, J. (2009). "Syntactic complexity in the brain," in *Syntactic Complexity: Diachrony, Acquisition, Neurocognition, Evolution*, eds T. Givon and M. Shibatani (Amsterdam: John Benjamins), 491–506.

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

Gentner, T. Q., and Hulse, S. H. (1998). Perceptual mechanisms for individual recognition in European starlings (*Sturnus vulgaris*). *Anim. Behav.* 56, 579–594.

Gobes, S. M. H., and Bolhuis, J. J. (2007). Bird song memory: a neural dissociation between song recognition and production. *Curr. Biol.* 17, 789–793.

Gobes, S. M. H., Zandbergen, M. A., and Bolhuis, J. J. (2010). Memory in the making: localized brain activation related to song learning in young songbirds. *Proc. R. Soc. Lond. B Biol. Sci. B* 277, 3343–3351.

Gold, E. (1978). Complexity of automaton identification from given data. *Inform. Contr.* 37, 302–320.

Goldberg, A. (2006). *Constructions at Work*. Oxford: Oxford University Press.

Haesler, S., Wada, K., Nshdejan, A., Edward, E., Morrisey, E., Lints, T., Jarvis, E. D., and Scharff, C. (2004). FoxP2 expression in avian vocal learners and non-learners. *J. Neurosci.* 24, 3164–3175.

Halder, G., Callaerts, P., and Gehring, W. J. (1995). New perspectives on eye evolution. *Curr. Opin. Genet. Dev.* 5, 602–609.

Halle, M., and Chomsky, N. (1968). *The Sound Patterns of English*. New York: Harcourt-Brace.

Halle, M., and Idsardi, W. J. (1995). "General properties of stress and metrical structure," in *A Handbook of Phonological Theory*, ed. J. Goldsmith (Oxford: Blackwell), 403–443.

Heinz, J. (2010). Learning long-distance phonotactics. *Linguistic Inquiry* 41, 623–661.

Heinz, J., and Idsardi, W. (2011). Sentence and sound complexity. *Science* 333, 295–297.

Hopcroft, J., and Ullman, J. (1979). *Introduction to Automata Theory, Languages, and Computation*. Reading, MA: Addison-Wesley.

Horning, J. (1969). *A Study of Grammatical Inference*. Ph.D. dissertation, Stanford University, Stanford.

Hsu, A. S., Chater, N., and Vitanyi, P. M. B. (2011). The probabilistic analysis of language acquisition: theoretical, computational, and experimental analysis. *Cognition* 120, 380–390.

Hurford, J. R. (2011). *The Origins of Grammar*. Oxford: Oxford University Press.

Huybregts, M. A. C. (1984). "The weak adequacy of context-free phrase structure grammar," in *Van Periferie Naar Kern*, eds G. J. de Haan, M. Trommelen, and W. Zonneveld (Dordrecht, Foris), 81–99.

Ikebuchi, M., and Okanoya, K. (2000). Limited memory for conspecific songs in a non-territorial songbird. *Neuroreport* 11, 3915–3919.

Imada, T., Zhang, Y., Cheour, M., Taulu, S., Ahonen, A., and Kuhl, P. K. (2006). Infant speech perception activates Broca's area: a developmental magnetoencephalography study. *Neuroreport* 17, 957–962.

Jackendoff, R. (1977). *X-Bar Syntax: A Theory of Phrase Structure*. Cambridge, MA: MIT Press.

Jackendoff, R. (2010). "Your theory of language evolution depends on your theory of language," in *The Evolution of Human Languages: Biolinguistic Perspectives*, eds R. Larson, V. Deprez, and H. Yamakido (Cambridge: Cambridge University Press), 63–72.

Jarvis, E. D. (2007). Neural systems for vocal learning in birds and humans: a synopsis. *J. Ornithol.* 148, 35–54.

Jarvis, E. D., Güntürkün, O., Bruce, L., Csillag, A., Karten, H., Kuenzel, W., Medina, L., Paxinos, G., Perkel, D. J., Shimizu, T., Striedter, G., Wild, J. M., Ball, G. F., Dugas-Ford, J., Durand, S. E., Hough, G. E., Husband, S., Kubikova, L., Lee, D. W., Mello, C. V., Powers, A., Siang, C., Smulders, T. V., Wada, K., White, S. A., Yamamoto, K., Yu, J., Reiner, A., Butler, A. B., and Avian Brain Nomenclature Consortium. (2005). Avian brains and a new understanding of vertebrate brain evolution. *Nat. Rev. Neurosci.* 6, 151–159.

Jarvis, E. D., and Mello, C. V. (2000). Molecular mapping of brain areas involved in parrot vocal communication. *J. Comp. Neurol.* 419, 1–31.

Jarvis, E. D., Ribeiro, S., da Silva, M., Ventura, D., Vielliard, J., and Mello, C. (2000). Behaviourally driven gene expression reveals song nuclei in hummingbird brain. *Nature* 406, 628–632.

Kakishita, Y., Sasahara, K., Nishino, T., Takahasi, M., and Okanoya, K. (2009). Ethological data mining: an automata-based approach to extract behavioural units and rules. *Data Min. Knowl. Discov.* 18, 446–471.

Katahira, K., Suzuki, K., Okanoya, K., and Okada, M. (2011). Complex sequencing rules of birdsong can be explained by simple hidden Markov processes. *PLoS ONE* 6, e24516. doi:10.1371/journal.pone.0024516

Kayne, R. S. (1994). *The Antisymmetry of Syntax*. Cambridge: MIT Press.

Kleene, S. (1953). *Introduction to Metamathematics*. Amsterdam: North-Holland.

Kleene, S. C. (1956). "Representation of events in nerve nets and finite automata," in *Automata Studies*, eds C. Shannon and W. Ashby (Princeton: Princeton University Press), 3–42.

Lambek, J. (1958). The mathematics of sentence structure. *Am. Math. Mon.* 65, 154–170.

Langendoen, T. (1975). Finite-state parsing of phrase-structure languages and the status of readjustment rules in the lexicon. *Linguist. Inq.* 6, 533–554.

Lasnik, H. (2000). *Syntactic Structures Revisited*. Cambridge, MA: MIT Press.

Laurin, M., and Reisz, R. (1995). A re-evaluation of early amniote phylogeny. *Zool. J. Linn. Soc.* 113, 165–223.

Marantz, A. (1982). Re-reduplication. *Linguist. Inq.*13, 435–482.

Marler, P. (1998) "Animal communication and human language," in *The Origin and Diversification of Human Language*, eds N. G. Jablonski and L. E. Aiello (San Francisco, CA: California Academy of Sciences), 1–19.

Mauner, G. A., Tanenhaus, M. K., and Carlson, G. N. (1995). A note on parallelism effects on processing verb phrase anaphors. *Lang. Cogn. Process.* 10, 1–12.

McNaughton, R., and Yamada, H. (1960). Regular expressions and state graphs for automata. *IEEE Trans. Electr. Comput.* 9, 39–47.

Miller, G., and Chomsky, N. (1963). "Finitary models of language users," in *Handbook of Mathematical Psychology*, Vol. 2, eds D. Luce, R. Bush, and E. Eugene Galanter (New York: Wiley), 419–491.

Miller, G. A., and Isard, S. (1964). Free recall of self-embedded English sentences. *Inform. Contr.* 7, 292–303.

Minsky, M. (1967). *Computation: Finite and Infinite Machines*. Englewood Cliffs, NJ: Prentice-Hall.

Mooney, R. (2009). Neural mechanisms for learned birdsong. *Learn. Mem.* 16, 655–669.

Moorman, S., Mello, C. V., and Bolhuis, J. J. (2011). From songs to synapses: molecular mechanisms of birdsong memory. *Bioessays* 33, 377–385.

Moro, A. (2000). *Dynamic Antisymmetry*. Cambridge: MIT Press.

Moro, A. (2008). *The Boundaries of Babel: The Brain and the Enigma of Impossible Languages*. Cambridge, MA: MIT Press.

Moro, A. (2011). A closer look at the turtle's eyes. *Proc. Natl. Acad. Sci. U.S.A.* 108, 2177–2178.

Musso, M., Moro, A., Glauche. V., Rijntjes, M., Reichenbach, J., Büchel, C., and Weiller, C. (2003). Broca's area and the language instinct. *Nat. Neuro.* 6, 774–781.

Niyogi, P. (2006). *The Computational and Evolutionary Nature of Language*. Cambridge, MA: MIT Press.

Okanoya, K. (2007). Language evolution and an emergent property. *Curr. Opin. Neurobiol.* 17, 271–276.

Oldfield, J., Williams, K., and Wiseman, N. (1987). Content-addressable memories for storing and processing recursively subdivided images and tress. *Electron. Lett.* 23, 262–263.

Pallier, C., Devauchelle, A.-D., and Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proc. Natl. Acad. Sci. U.S.A.* 108, 2522–2527.

Perfors, A., Reiger, C., and Tenenbaum, J. (2010). The learnability of abstract syntactic principles. *Cognition* 3, 306–338.

Perruchet, P., and Rey, A. (2005). Does the mastery of center-embedded linguistic structures distinguish humans? *Psychon. Bull. Rev.* 12, 207–313.

Petersson, K. M., Folia, V., and Hagoort, P. (2012). What artificial grammar learning reveals about the neurobiology of syntax. *Brain Lang.* 120, 83–95.

Petitto, L. A. (2005). "How the brain 2732 begets language," in *The Cambridge Companion to Chomsky*, ed. J. McGilvray (Cambridge: Cambridge University Press), 84–101.

Petitto, L. A., Holowka, S., Sergio, L., Levy, B., and Ostry, D. (2004). Baby hands that move to the rhythm of language: hearing babies acquiring sign languages babble silently on the hands. *Cognition* 9, 43–73.

Petri, J., and Scharff, C. (2011). Evo-devo, deep homology and FoxP2: implications for the evolution of speech and language. *Phil. Trans. R. Soc. B* 1574, 2124–2140.

Pulvermüller, F. (1993). "On conecting syntax and the brain," in *Brain Theory – Spatio-Temporal Aspects of Brain Function*, ed. Aertsen (New York: Elsevier), 131–145.

Pulvermüller, F., and Knoblauch, A. (2009). Discrete combinatorial circuits emerging in neural networks: a mechanism for rules of grammar in the human brain? *Neural. Netw.* 22, 161–172.

Rabin, M., and Scott, D. (1959). Finite automata and their decision problems. *IBM J. Res. Dev.* 3, 114.

Radford, A. (1997). *Syntactic Theory and the Structure of English: A Minimalist Approach.* Cambridge: Cambridge University Press.

Riebel, K., and Slater, P. J. B. (2003). Temporal variation in male chaffinch song depends on the singer and the song type. *Behaviour* 140, 269–288.

Rogers, J., and Pullum, G. K. (2011). Aural pattern recognition experiments and the subregular hierarchy. *J. Log. Lang. Inform.* 20, 329–342.

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

Sag, I. A., Wasow, T., and Bender, E. (2003). *Syntactic Theory: A Formal Introduction*, 2nd Edn. Stanford: CSLI Publications.

Sasahara, K., Kakishita, Y., Nishino, T., Takahasi, M., and Okanoya, K. (2006). "A reversible automata approach to modeling birdsongs," in *Proceedings of 15th International Conference on Computing (CIC2006)* (New York: IEEE Computer Society Press), 80–85.

Scharff, C., and Petri, J. (2011). Evo-devo, deep homology and FoxP2: implications implications for the evolution of speech and language. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 366, 2124–2140.

Shettleworth, S. J. (2010). *Cognition, Evolution, and Behavior.* Oxford: Oxford University Press.

Shubin, N., Tabin, C., and Carroll, S. (1997). Fossils, genes and the evolution of animal limbs. *Nature* 388, 639–648.

Shukla, M., White, K. S., and Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proc. Natl. Acad. Sci. U.S.A.* 108, 6038–6043.

Sipser, M. (1997). *Introduction to the Theory of Computation.* Boston, MA: PWS Publishing.

Solomonoff, R. (1964). A formal theory of inductive inference. *Inform. Contr.* 7, 1–22.

Stolcke, A. (1994). *Bayesian Learning of Probabilistic Learning Models.* Ph.D. dissertation, Stanford University, Stanford.

Steedman, M. (2000). *The Syntactic Process.* Cambridge, MA: MIT Press.

Stevens, K. (2000). *Acoustic Phonetics.* Cambridge, MA: MIT Press.

Suge, R., and Okanoya, K. (2010). Perceptual chunking in the self-produced songs of Bengalese Finches (*Lonchura striata* var. *domestica*). *Anim. Cogn.* 13, 515–523.

Suh, A., Paus, M., Kiefmann, M., Churakov, G., Franke, F., Brosius, J., Kriegs, J., and Schmitz, J. (2011). Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. *Nat. Commun.* 2, 1–7.

Todt, D., and Hultsch, H. (1996). "Acquisition and performance of repertoires: ways of coping with diversity and versatility," in *Ecology and Evolution of Communication*, eds D. E. Kroodsma, and E. H. Miller (Ithaca, NY; Cornell University Press), 79–96.

Uddén, J., Ingvar, M., Hagoort, P., and Petersson, K. (2011). Implicit acquisition of grammars with crossed and nested non-adjacent dependencies: investigating the push-down stack model. *Cogn. Sci.* (in press).

Vernes, S. C., Oliver, P. L., Spiteri, E., Lockstone, H. E., Puliyadi, R., Taylor, J. M., Ho, J., Mombereau, C., Brewer, A., Lowy, E., Nicod, J., Groszer, M., Baban, D., Sahgal, N., Cazier, J.-B., Ragoussis, J., Davies, K. E., Geschwind, D. H., and Fisher, S. E. (2011). Foxp2 regulates gene networks implicated in neurite outgrowth in the developing brain. *PLoS Genet.* 7, e1002145. doi:10.1371/journal.pgen.1002145

Webb, D., and Zhang, J. (2005). FoxP2 in song-learning birds and vocal-learning mammals. *J. Hered.* 96, 212–216.

Weir, D. (1988). *Characterizing Mildly-Context Sensitive Grammar Formalisms.* Ph.D. dissertation, University of Pennsylvania, Philadelphia.

Wohlgemuth, M. J., Sober, S., and Brainard, M. S. (2010). Linked control of syllable sequence and phonology in birdsong. *J. Neurosci.* 29, 12936–12949.

Woods, W. (1970). Transition network grammars for natural language analysis. *Commun. ACM* 13, 591–606.

Yip, M. (2006). The search for phonology in other species. *Trends Cogn. Sci.* 10, 442–446.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 December 2011; accepted: 20 March 2012; published online: 13 April 2012.

Citation: Berwick RC, Beckers GJL, Okanoya K and Bolhuis JJ (2012) A bird's eye view of human language evolution. Front. Evol. Neurosci. ***4****:5. doi: 10.3389/fnevo.2012.00005*