21

Culture and Evolution

Ian Cross, W. Tecumseh Fitch, Francisco Aboitiz, Atsushi Iriki, Erich D. Jarvis, Jerome Lewis, Katja Liebal, Bjorn Merker, Dietrich Stout, and Sandra E. Trehub

Abstract

This chapter captures extensive discussions between people with different forms of expertise and viewpoints. It explores the relationships between language and music in evolutionary and cultural context. Rather than trying to essentialize either, they are characterized pragmatically in terms of features that appear to distinguish them (such as language's compositional propositionality as opposed to music's foregrounding of isochronicity), and those that they evidently share. Factors are considered that constitute proximate motivations for humans to communicate through language and music, ranging from language's practical value in the organization of collective behavior to music's significant role in eliciting and managing prosocial attitudes. Possible distal motivations are reviewed for music and language, in terms of the potentially adaptive functions of human communication systems, and an assessment is made of the advantages which might accrue to flexible communicators in the light of ethological and archaeological evidence concerning the landscape of selection. Subsequently, the possible evolutionary relationships between music and language are explored, within a framework supplied by six possible models of their emergence. Issues of the roles of culture and of biology in the evolution of communication systems are then addressed within the framework of triadic niche construction, and the chapter concludes by surveying available comparative and phylogenetic issues that might inform the debate.

Distinguishing Music from Language

In placing music and language within the frames of culture and evolution, one is necessarily confronted by the question: "What is intended by the terms "music" and "language?" Are we dealing with culturally shaped distinctions or biologically distinct systems? Are music and language categorically discrete human faculties, or do they constitute different manifestations of the same underlying communicative capacities? Our initial strategy is to avoid definitions in favor of identifying features that distinguish between the two domains;

in postulating distinct features of music and language, we run the risk of essentializing ethnocentric concepts or stressing between-category differences and minimizing within-category differences, in effect, reifying distinctions that may not be supported by the evidence. It must be acknowledged, however, that in all known human cultures, the available suite of behaviors includes something that appears like music, just as it includes language, though the extent to which categorical distinctions are drawn between music and language, and the factors that motivate any distinction between the two domains, differ across cultures.

A key attribute that appears to distinguish between the domains is *propositionality*. Language, unlike music, provides a way of sharing information about states of affairs by means of truth-conditional propositions and thus of coordinating action. It enables mapping between worlds, thoughts, and selves, the formulation and exchange of information, and the coordination of joint, goal-directed action. Music appears to have none of these functional benefits, but it has others that we will consider subsequently. Nevertheless, music and language share the significant feature of *generativity*. Both afford complex combinatoriality and unlimited generativity via a few simple nonblending (particulate) elements into composite, individually distinctive patterns. Such a system has been called a Humboldt system, after Wilhelm von Humboldt, who first described language in these terms (for music, see Merker 2002). Combinatoriality is also found in vocal learning songbirds, such as the sedge warbler (*Acrocephalus schoenebaenus*) who varies the *sequencing* of his stock of some fifty different song elements to produce song patterns which essentially never repeat (Catchpole 1976). The sedge warbler's song, however, is not semanticized; the different patterns pouring out of the sedge warbler's throat are not invested with distinctive meanings. Moreover, other songbirds may only have one song, or very few variations. This is by way of contrast to the varied phoneme sequences in human speech, which may form words with learned meanings, the words in turn composing sentences, with the grammar of the language specifying how the meanings of words combine to imbue each sentence with distinctive meaning predicated on the specific assembly of phonemes/words of which they consist. This is what allows language to carry propositional meaning riding on the phonemic stream of speech, by contrast to the "note stream" of music, which has no corresponding compositionality of meaning. Music's combinatorial aspect falls closer to the sedge's warbler's use of combinatoriality to mount what we think of as an impressive aesthetic display. In any case, if the deep similarity between music and language is their hierarchical structure as yielded by Humboldt system generativity, the lack of formal semanticization in music (without lyrics) is the major contrast between music and language, fitting them for different uses in human communication. As language users, we need to share some common ground to conduct our dialogic and propositional transactions. This common ground is established largely by interaction within a shared community, being

built on commonalities of knowledge and belief mediated by the propositions shared in our linguistic exchanges or our observation of such exchanges (e.g., hearsay). As an evolutionary counterpoint, we may note that there is evidence from monkeys and chimpanzees (Crockford et al. 2004; Clay and Zuberbühler 2011) of control or combination of vocalizations which result in a change of meaning, although we have here just a few such vocalizations, with neither a Humboldt system nor a compositional semantics.

This account needs, however, to be supplemented by the realization that much linguistic dialog is not concerned with the exchange of formal propositions but rather with maintaining social networks (Dunbar 1996; Wray 1998), which is to say that a significant part of linguistic interaction is relational rather than transactional. Moreover, while music cannot communicate propositional information, the idea that music has meaning is widespread across cultures. In fact, music is frequently reported as bearing meanings similar to those transmitted by linguistic means. According to Leonard Meyer (1956:265), "music presents a generic event, a "connotative complex," which then becomes particularized in the experience of the individual listener." Such experience remains individual rather than being made mutually manifest to other listeners, as would be the case for language. If music is considered an interactive or *participatory* phenomenon, in contrast to the *presentational* form that it typically takes in Western conceptions (Turino 2008), close parallels emerge between the features of music and the relational features that sustain conversational interchange. Hence the criterial distinctions between music and language as interactive media may involve the extent to which each medium requires mutually comprehensible reference and foreground features concerned with sustaining the interaction.

While music and language appear to constitute discrete categories in contemporary Western societies, for many cultures they may be best conceived of as poles of a continuum, or there are divisions into more than two categories. For example, a complex set of distinctions is provided by Seeger (1987), who notes that primary distinctions made between "communicative genres" by the Suyá people of the Amazon are between three categories:

1. kaperní, which more-or-less corresponds to everyday speech, where there is a priority of text over melody, text and melody being determined by speaker, with an increasing formalization in public performance;
2. sarén, "telling" or instructional speech, where there is a relative priority of relatively fixed texts over relatively fixed melodies; and
3. ngére, song, where there is a priority of melody over text, and, importantly, time, text, and melody are fixed by a nonhuman source.

As Seeger (1987:50) notes, "Melody is not a particularly good way to distinguish between Suyá speech, instruction, and song." Some manifestations of *kaperní* may appear to shade into manifestations of *sarén*; similarly, it may be difficult to distinguish between instances of *sarén* as these may, in turn, begin

to shade into *ngére*. Here, modes of communication are being distinguished on the basis of their social function and their proper domain: *kapernί*, speech, is for mundane, everyday use, originating with—and being directed toward—humans; *sarén*, didactic talk, requires authority, whether deriving from present-day power structures or the invocation of a teacher from the past; whereas *ngére*, song, can constitute a special, liminally powerful medium, having nonhuman origins and being directed, in part, toward nonhuman agency. The Suyá are not alone in making such distinctions; other traditional cultures frequently embed what may appear as speech and song to Western observers in similarly complex communicative taxonomies (see Basso 1985; Feld 1982; Lewis, this volume).

We propose, therefore, that music and language constitute a continuum rather than discrete domains. This continuum can be interpreted in terms of at least two dimensions, the first running from definite to indefinite meanings and the second from greater to lesser affective potency. Music's power to form complex patterns (enabled by its generativity), its frequent repetition of elements (in comparison with language), together with its iconicity (i.e., its exploitation of biologically significant aspects of sound) endow it with an ambiguity and an immediacy that can be emotionally compelling. Language's capacity to formulate and exchange complex propositions allows it to represent an infinite variety of meanings and frees it, in principle, from the exigencies of affect. However, the discrete tones and pitch sets that supply grist for the musical mill in most cultures are rather unique to music, though a few birds (e.g., the pied butcher bird of Australia) do appear to feature them. Also, for humans the speaking voice is a highly significant biological sound whose emotional coloring draws on our repertoire of innate nonverbal emotional expressiveness. We routinely express emotion through the modality of speech rather than music; nothing compels music to convey emotion.

Given such blurrings of any strict dichotomy, it may be helpful to stress contexts of use, just as in the Suyá example above. The typical linguistic exchange is between two persons whereas, for most of its history, music has occurred in group contexts. Importantly, language and music differ in their power to coordinate human movement. There are differences in the regularity of timing between most registers of speech and most genres of music, with the latter featuring explicit use of isochrony, though it should be noted that this is a feature of both didactic talk and oratory, both oriented toward "musical" ends of capturing attention and enhancing a sense of mutual affiliation. The isochrony of music facilitates the timing of one's own movements and the prediction of others' movements, allowing for mutual co-adjustment of phase and period in simultaneous and sequential movements. Music's isochronicity and metrical structure may also underpin a greater mnemonic potential compared to language, or the musical feature of isochronicity itself may endow language with such mnemonic potential. In oral cultures, transmission

of cross-generational knowledge is likely to take forms that appear musical and poetic rather than discursive (see, e.g., Rubin 1995; Tillmann and Dowling 2007).

Low-level differences in acoustic attributes may also warrant a clear distinction between music and language, or more properly, between speech and song. Music (song) and spoken language differ in their inter-event transitions—the ways in which sounds succeed one another—in terms of rhythm (the previously noted tendency toward isochronicity) and formant transition, with sharper formant transitions in speech than in song. Schlaug (see, e.g., Özdemir et al. 2006) has suggested that the same pathways are used for the perception of music and speech in contrast to parallel pathways for the production of speech and singing, the latter arising, perhaps, from rate differences between speech and music. In contrast, Jarvis (2004) has suggested that the same pathways are used in different ways to produce song and speech; the latter is true of song learning birds, such as parrots, that can learn to sing as well as to imitate human speech.

The foregoing discussion has largely characterized music and language as an auditory-vocal phenomenon. Of course, both involve action in the form of gesture. Spoken language is typically embedded in a complex interactive matrix of gesture (see, e.g., Kendon 2004), and there are numerous signed languages. Music involves overt action, not only in its production but also as an interactive process or network of gestures among participants (Moran and Pinto 2007). Indeed, music is indissociable from dance as a cultural category in many societies (e.g., Stone 1998). As gestural media, music and language may be distinguishable in terms of timing and organization. Gestures in language tend to be sequential and timed in relation to prosody rather than an underlying rhythm, whereas those in music often involve temporal regularity and may involve simultaneity between participants. Nevertheless, there are counterexamples such as coincident gestures of participants in linguistic interaction, often at points of topical agreement in discourse (Gill et al. 2000), intermittent temporal regularity, or absence of meter in music.

Overall, no single criterial attribute, save perhaps that of propositionality, distinguishes between language and music clearly and comprehensively. As Wittgenstein (1953) noted some years ago, categories need not have criterial or defining features. Instead, instances of a category can have a "family resemblance" or one or more common attributes shared with some but not all instances of the category. As with any category (e.g., birds), there are prototypical (e.g., robins) and less prototypical (e.g., chickens) instances (Rosch 1975).

One can also ask whether the features of music and language are uniquely human. During our discussions, we listed (Table 21.1) behavioral and neural parallels that have been documented in nonhuman species. In some cases, nonhuman animals trained by humans have succeeded in recognizing many words (e.g., the dog Rico; Kaminski et al. 2004), phrases (e.g., the parrot

Table 21.1 Subcomponents of music and language.

1. Behavioral Components:
   a. Signal
      - Perception of speech (acoustic pattern recognition system). Lexical access may be unique to humans, since speech perception (involving lexical access) involves making lexical commitments. But what about Alex, the African grey parrot, and Rico (Fischer's dog)?
      - Production of speech and song
      - Limited vs. complex vocal learning: humans, birds
      - Opportunistic multimodality: ape gestural communication
      - Hypermeter: multilevel meter, hierarchical structure in whale song
      - Voluntary control of vocalizations
      - Instrumental, nonvocal music
   b. Structure and phonology
      - Syntax minus meaning, vocal combinatoriality (sequencing of learned syllables): any animal that has a complex song (e.g., humans and birds), but we don't know enough
      - Recursion
      - Scales
      - Relative pitch: ferrets
      - Working memory
   c. Pragmatics
      - Theory of mind, as evidenced in intentional communication
      - Extreme sociality or the motivation to share experience
      - Vocal maintenance of mother–infant bonds
      - Entrainment: frogs/insects vs. parrots (cross-modal, potentially communicative in relational terms)
      - Dyadic dialog (context of communication), face-to-face, addressed communication, deictic switch, multimodality (i.e., the extent to which contents of turns are conditional on partner's productions), agonistic versus cooperative engagement
   d. Semantics
      - Referentiality in the form of compositional semantics: unique, though precursors or minimal commonalities exist (e.g., monkey booming as signifying negation)
      - Predication (predicate/argument)
      - Cultural transmission at every level in vocal communication (extreme variability of human language)
      - Lack of signification (displacement of reference in space and time)
      - Notation

Alex; Pepperberg 1999), as well as octave-transposed melodies (e.g., in rhesus monkeys; Wright et al. 2000). There is no indication, however, of comparable feats in the natural environment in these or other nonhuman species. Moreover, the prevailing view is that language and music are unique products of human culture (and nature), although elements of each may be present in other species. At the same time, it is important to note that stimuli and tasks

Table 21.1 (continued)

2. Neural Components
    - Auditory forebrain pathway (Wernicke's area)
    - Forebrain vocal control path of vocal structure (including Broca's are, striatum, thalamus)
    - Direct connection from cortex to brainstem vocal-motor neurons: lateral motor area–laryngeal motor neurons
    - Between humans and nonhuman primates, there appears to be a direct connection between auditory and primary or secondary motor areas (arcuate fasciculus?)
    - Differential gene regulation and convergent mutation in genes that make direct projections in other forebrain areas that control vocalizations in both humans and songbirds (deep homology)
    - Auditory receptive field sharpness in humans
    - Lateralization: greater specialization in humans in the representation of communication sounds
    - Spindle cells in anterior cingulate cortex only in humans and great apes (also in dolphins?)
    - Heterochronicity of cortical synaptogenesis unique to humans (among primates)?
    - Does brain size matter?

involving nonhuman participants (and even human participants) typically lack ecological validity. In general, nonhuman species have difficulty recognizing transpositions of tone sequences, so it is of particular interest that European starlings can be trained to recognize transpositions of conspecific songs but fail to recognize transposed piano melodies after comparable training (Bregman et al. 2012). In any case, there is no evidence of a nonhuman species, whether in the wild or trained by humans, whose members combine Humboldt system generativity with a compositional semantics.

In this chapter, we view music and language as constituting different manifestations of the human capacity to communicate—manifestations which may take very different forms in different cultural contexts. Is that partly because, outside of its cultural context, music cannot be defined unambiguously? Persons within a culture usually have no difficulty differentiating most registers of speech from most forms of music. One complicating factor is that we have a reasonable understanding of the functions of language across cultures, but we have much less understanding with respect to music. In considering the place of music and language in culture and evolution, we must address the question of what impels humans to communicate—through language or music.

Proximate Motivators for Human Communication

That humans are highly motivated to communicate is unquestionable; the issue of what may underpin that motivation is, however, less certain.

Communication—at least, in the form of language—has immense value in helping groups of individuals shape their environments, individually or collectively, so as to attain goals. In the form of socially oriented or phatic talk, language can serve to build and maintain relationships in social interactions. There are, however, many other motivations to communicate that are likely to apply to a broader range of communicative systems than language alone. For vocal learning species, there seems to be an instrinsic pleasure in vocalizing (e.g., in forms such as babbling, subsong, or imitation). For humans (and perhaps some other primate species), vocal and gestural communication serves to co-regulate affective states between the caregiver and infant, and to enhance a sense of mutual affiliation. Communication can have prosocial effects, not just for dyads but also for larger groups: we may gain pleasure from collective and synchronized performance which, in turn, reduces social uncertainty and helps bond the group, enhancing the effectiveness of group action and identity, particularly when directed against potential external threats (e.g., other groups or prospective predators). Of course, once we can behave linguistically or musically, we can be motivated to co-opt these communicative resources for other ends; "inner speech" may be deployed to reduce uncertainty in attention-based coordination (Clark 2002) or to manage communication (Allwood 2007), whereas self-directed music may be produced as a means of affect regulation, as in the *dit* songs of the Eipo (Simon 1978).

Levinson (2006; see also this volume) argues for extraordinary human sociality grounded in an innate capacity for social interaction involving unique cognitive infrastructure (see also De Ruiter et al. 2010). Others emphasize the role of culture and experience in elaborating our inherited cognitive infrastructure (e.g., Vygotsky 1978). By 12 months of age, infants engage in declarative pointing to share their interest in events with others, to make requests, or to provide helpful information (Liszkowski 2011). They also vocalize to attract parents' attention. In fact, infants vocalize well before their vocalizations are intentionally communicative, perhaps because vocalizing is intrinsically pleasurable. However, the most significant motivation for human communication is the sharing of experience; that is, wanting another to see, feel, think, or know what I see, feel, think, or know. Early pointing in infancy is of the "look at that" variety rather than the instrumental or "get me that" variety. The pleasure of vocalization as the motivator would lead to a lot more vocalization in the absence of others, but this has nothing to do with communication. More generally, young children make greater use of gesture than spoken words in their early language development (Capirci et al. 2002). Tomasello (2008) emphasizes how different this is from the instrumental form of communication in ape gestures, in which one ape tries to modify the behavior of another. The pleasure of vocalizing has a more direct utility in mother–infant interactions although such interactions proceed equally smoothly in deaf mother–infant dyads, who use gestural rather than vocal signals. Early communicative mother–infant interactions have evident

functions in co-regulating the affective states of both participants. Such early experiences may underlie the ability of music to facilitate entry into states of shared intentionality or even trance-type states. These capacities may be built on a more general substrate.

Clearly, there is more that motivates humans to communicate than just vocal pleasure. We gain huge practical advantages from being able to exchange information linguistically and to coordinate our actions with others. Motivational factors may drive us not only to speak but also to sing and move together with others in dance or synchronous movement, and this may strengthen social bonds (McNeill 1995). Communication by means of language and music affords us, respectively, the capacity for information transfer as well as the formation and maintenance of group solidarity. We can use language to get what we want and to transfer information, whereas we can use music to give us pleasure and to achieve group solidarity as well as to relieve pain and suffering and to reduce stress (Knox et al. 2011). Indeed, a defining characteristic of the human species is a propensity for cooperation and prosociality (Levinson 2006; Tomasello 2008). We note, however, that much of speech does not appear to be oriented toward the transfer of information but to processes of establishing mutual affiliation with others (i.e., functions which may be hypertrophied in music). We seem motivated to order social life through language and music, but it is notable that music, rather than language, tends to be at the forefront of situations where social conflict is a potential threat to the social order (e.g., Marett 2005).

One key factor that orders the human motivation to communicate is that of culture, which plays a key role in shaping, structuring, and ordering the human motivation to communicate, although here we have an example of an expanding spiral: new means to communicate support developments in culture, and new cultural and social processes provide an ecological niche for the emergence of new communicative forms. While we may gain pleasure from communicating or synchronizing with others, different cultures sanction these behaviors in different ways, with enculturation processes shaping acceptable patterns of communication. Notable examples can be found in some traditional cultures, where silent co-presence can be privileged over relationally oriented speech (Basso 1970), as well as in a range of situations in all cultures where institutions constrain or facilitate the motivation to communicate.

While pleasure (the instrumental value of a means of information exchange) and the human benefits of interpersonal connections and group solidarity may provide proximal motivation in human communication, these forces must be situated in their broader evolutionary context, to which we now turn.

Adaptive Functions of Human Communicative Systems

The most direct evidence for the emergence of complex communicative faculties early in the hominin lineage is in the lengthy archaeological record of complex lithic technologies transmitted over multiple generations. That persistence of cultural transmission suggests that early hominin cognitions and interactions must have been characterized by intense social cooperativity and inhibition of aggression. Material technology was employed in food acquisition and preparation, including group hunting, which required the recognition of multiple levels of intention in order to second-guess prey and coordinate group hunting behavior. Also required was the capacity for planning, which involves the manipulation of nonexistent entities and the composition of structures free from the immediate constraints of the physical world. Together, all these factors create a fitness landscape within which communicative capacities—and a progressive enhancement of communicative capacities—would have been adaptive. Of course, there would have been other selection pressures for the emergence of flexible communicative capacities, perhaps arising in the context of within-group, or sexual, competition. In addition, the effects of aspects of music on arousal in nonhuman species reminds us that many of the factors that make up the modern human communicative repertoire are likely to be shared with a variety of other species. Different factors are likely to have arisen at different times under different selection pressures, and it is likely that evidence for these different evolutionary time depths is embodied in the structures and dynamics of our neural and genetic systems.

The emotional aspects of music are often conceived of as being specific to humans. However, the arousing dimension of responses to features that are evident in music may be shared by other species. For example, auditory rhythmic features arouse chickens (indexed by noradrenaline release) and affect memory consolidation (Judde and Rickard 2010; Rickard et al. 2005). The effect of subcomponents of music on other cognitive functions suggests that music can have fundamental as well as higher adaptive functions, and a comparative approach is needed to differentiate homology and analogy. Rather than taking the response to sound, in the form of music, as a starting point, perhaps learned vocal communication is being selected. In a range of species, learned vocal communication is used for mate attraction, on the basis of variability of F0 and syntax, which raises the question of why a "supranormal stimulus" effect of vocal sounds is not more common. Given the linkage of gesture with speech or of dance with music, it is a matter of debate whether the evolution of vocal learning was the driver for the emergence of language and music or was driven in part by the evolution of other embodied systems. For example, Arbib and Iriki (this volume) discuss the hypothesis that complex imitation of manual skills underwrote the evolution of manual gesture, and that the emergence of "protosign" provided a necessary scaffolding for the emergence of vocal learning in support of semantic expressivity. Alternatively,

the ability to regulate the expression of emotion, whether bodily (gestural, postural) or facial, may differentiate humans from other species. This hypothesis is rooted in our understanding of the human capacity to control the expression of emotion. At present, there is little evidence of comparable control of facial expressions and vocalization in nonhuman species, though some precursor ability has been shown in monkeys (Hihara et al. 2003). Other work (Slocombe and Zuberbühler 2007) suggests that chimpanzees have some control over the production of their vocalizations since they recruit specific group members to support them in aggressive encounters. We share with our closest relatives the capacity to produce an initial affect burst in response to situational stress (see Scherer, this volume), but we know little about their capacity to shape and redirect such affect bursts. It is certainly the case that apes can be opportunistic in exploiting different channels for communication (e.g., Leavens et al. 2004; Liebal et al. 2004), and it may be that the multimodality which characterizes speech (and music in action) has its origins in such capacities. Humans, like all primates, mammals, and indeed most vertebrates, have a multifaceted repertoire of largely innate nonverbal emotional expressiveness, which includes a rich repertoire of specifically vocal, emotional expressivity that is neither music, nor language, but which can be drawn upon by both of these for purposes of emotional coloring (e.g., in the dynamics and prosody of emotional speech). This preexisting largely innate repertoire is the key to the biology of emotional expressiveness in humans as in other species. However, if so, it must be stressed that the differences between such capacities and human music and language are immense.

Complex behaviors—such as acts of deception, binding the exercise of capacities for adopting the perspective of others with requirements to control mutually manifest behavior (e.g., vocalization)—may have provided grounds for the emergence of signals that have reference in relation to a state to be co-opted for proto-propositional use. Here, a parallel development of speech and music may be proposed, and the relationships between the raw expression of affect and the controlled articulation of art, whether linguistic or musical, could be explored. However, reasonably stable social groups would be needed to drive this process. One way of finding evidence for these speculations is to examine the range of emotional vocalizations from "raw affect bursts" to culturally defined quasi-lexical elements. This might shed light on the way in which raw vocalizations that we share with mammals have come under increasing control, both with respect to production and desired targets for communication. We note, however, that the control of the emotional expressions we share with other primates rests on medial circuitry (anterior cingulate as modulator of brainstem circuitry), whereas much of the circuitry associated with human language and music resides more laterally in the cortex (Jürgens 2009). Moreover, it is a classic observation going back to Hughlings Jackson in the 19th century that an aphasic may lose the propositional use of language yet still emit imprecations. Indeed, in humans, a crucial result of

evolution is that language can take over from direct, affect-induced action as a means of negotiating situations where different individuals' needs or desires are manifestly in conflict. A further factor that could have driven the emergence of something like language is an increase in the ability, and motivation, to make plans in conjunction with others. Such planning requires shared goals and manipulation of nonexistent entities, enabling the composition of structures free from the immediate constraints of the physical world. Here, the range of theories seeking to link the evolution of brain mechanisms supporting language to those supporting tool use become especially relevant, with the notion that visualization of a goal may play a crucial role in planning the means to achieve it (Stout and Chaminade 2012). Off-line planning may (but need not) render concrete phenomena less immediately relevant, affording a means to displace reference (cf. Iriki 2011). Such considerations may underlie the evolution of both language and music. Not only is language's propositionality built on reference to present and absent entities and events, but music affords an abstract domain for the construction of sound worlds that may be similarly grounded in experience yet divorced from immediate events.

The emergence of pedagogical capacities at some point in the hominin lineage may be a more specific driver for the propositional and intentional dimensions of language. Pedagogy involves the intentional alteration of one's behavior to influence the mental states (attention, knowledge, embodied skills) of other individuals. In Arbib's version of the gestural origins hypothesis (mirror system hypothesis; Arbib 2005), the transition to intentional communication requires a pantomimic/proto-sign phase. It could be argued that the intentionality of non-pantomimic communication in pedagogy shows that these substages may not be needed. The counterargument is that demonstration or modeling is an important part of pedagogy in natural environments. Gesture would be critical in such circumstances and would precede verbal instruction (Zukow-Goldring 1996, 2006). The need to communicate increasingly opaque causal relations in technological pedagogy also supplies a potential selective pressure for development of propositional meaning in language, but one must not conflate later stages of language evolution with their necessary precursors. Opaque causal relations are evidenced in skill transmission in modern humans, which involves not only direct communication, but also the creation of situations conducive for learning. This requires a high level of social cohesion, including (at least in modern humans) the development of appropriate skills and motivation for caregivers. The Vygotskian zone of proximal development (e.g., Vygotsky 1978) involves adult mentoring or scaffolding, which allows the learner to go beyond what he is capable of doing on his own. It refers to the difference between what the child can do independently and what he can do with adult assistance. The former indicates his state of knowledge or skill whereas the latter indicates his potential. In essence, it concerns culturally mediated learning rather than traditional pedagogy. This need for explicit support of the child's mental development may be an additional selective

pressure in the expanding spiral for language (storytelling, kinship, etc.) and music (social bonding).

Language is marked out not just by its propositionality but by its complex propositionality, which entails compositionality, hierarchical structure, and complex syntax. These constitute very general capacities that are taken to high levels in language and, in some instances, music. These features are probably important for many evolutionarily relevant behaviors, but they are visible and testable in the archaeological record of stone tools. The archeological record of tools can document the expression of a particular depth/complexity of hierarchical action organization at a particular time, which provides a *minimum* indication of past hominin capacities. Stout's work (e.g., Stout et al. 2008) provides PET and fMRI evidence of increasing activation of anterior inferior frontal gyrus (hierarchical cognition) in increasingly complex stone tool-making as well as activation of medial prefrontal cortex during observation of tool-making by experts (intention attribution). A three-year longitudinal study of tool-making skill acquisition, which involves behavioral, social, archaeological (lithic analysis), neurofunctional (fMRI), and neuroanatomical (VBM, DTI) observations, is currently in progress, one output of which will be an empirically derived action syntax of Paleolithic tool-making. This work provides a clear and testable set of hypotheses concerning the emergence of capacities for compositionality and hierarchical structure and the facilitative effects of pedagogy. If an association can be established between the presence of vocal learning and the importance of "teaching" in other animals, its implications would be substantially broadened. The mirror system hypothesis would view such skill transfer as driving gestural communication more directly, with this in turn providing scaffolding for increasingly subtle vocal communication. In any case, much of human culture, and most of animal life, proceeds without pedagogy in any explicit, formal sense. That includes the acquisition of skills in many useful arts for which observational learning with "intent participation" often suffices (Rogoff et al. 2003).

Pedagogy, in whatever form, appears to require the capacity for recognition of multiple levels of intention ("orders of intentionality") and may be tied to the emergence of that capacity. It is suggested that chimpanzees have two orders of intentionality ("I believe that you intend…"), whereas humans can manage up to five or six (Dennett 1983). In any case, there is a chicken-and-egg problem in placing language and intentionality in evolutionary perspective, as language itself promotes development of ToM abilities, as indicated by the considerable lag in deaf children's achievement of ToM milestones (Wellman et al. 2011).

One of the prime requisites for big-game hunting—a subsistence strategy of current hunter-gatherers and of several of our recent ancestor species—is the ability to second-guess prey and to coordinate group hunting behavior. A switch in the hominin lineage to social hunting, rather than scavenging, may have helped provide selection pressures for the emergence of the capacity for recognition of multiple levels of intention, though Bickerton (2009) argues

that scavenging, rather than hunting, provided the ecological niche that supported the emergence of language—perhaps too mono-causal a view of human evolution. It is notable that social hunting species, such as African hunting dogs and wolves, may have higher levels of intention recognition than nonsocial hunters, most likely driven by the demands of group hunting (Nudds 1978). However, this is without a hint of leading to either music or language, so one must still seek that "something extra" in human evolution.

Social hunting necessitates close cooperation with others, and there is extensive human evidence for cooperation, collaboration, reciprocity, and shared goals. Tomasello (2008) argues that such cooperation is a precondition for the development of complex culture (i.e., involving learning in several domains) and for complex communication systems such as language and music. He also emphasizes the importance of ratcheting, so that each skill becomes the building block for others (Tomasello 1999; Tennie et al. 2009), thus explaining why human culture is so much richer than that of chimpanzees (Whiten et al. 1999). In humans, cooperation or helping others is evident even when there is no obvious benefit to the helper. Planning becomes critical in attaining difficult goals involving two or more individuals (e.g., hunting, sharing the spoils, achieving a division of labor that increases efficiency). Moreover, effective planning is greatly assisted by effective communication. There is reported nonhuman evidence of cooperative hunting (i.e., hunting in groups or packs), but these instances of apparent cooperation may simply maximize self-interest (for evidence on the lack of reciprocity in chimpanzee food sharing, see Gilby 2006). It is therefore unclear whether group hunting involves genuine cooperation. If cooperative motives were involved, the collaborators would be unlikely to fight vigorously over the carcass, as they typically do. A major social change in our species might be revealed through the study of the social brain, or by means of social neuroscience. Indeed, the persistence over many generations of culturally transmitted behaviors, such as Acheulean technology in the *Homo* lineage, suggests that there must have been intense social cooperation and inhibition of aggression, which would predict significant frontal brain enlargement.

While these hypotheses stress the benefits conferred by linguistic and musical interaction to individuals within the group, questions remain about who accrues the advantage (individual, kin group). The aforementioned hypotheses do not necessitate group selection, but are instead concerned with standard processes of natural and sexual selection, or with standard natural selection operating within the context of cultural niche construction (see, e.g., Laland et al. 1996). We do compete within groups, and such competition is often evident in processes of sexual selection, where we find the aesthetic extravaganzas of nature such as the peacock's tail and elaborate bird song, which are intended to impress conspecifics. In line with Darwin's original suggestion that music arises as a consequence of processes of sexual selection (Darwin 1871), it is possible that aspects of music, such as pulse-based isochrony, might not

have derived from general processes of cooperation but from sexual selection pressures. Our ancestral setting of male territoriality and female exogamy could have led to synchronous chorusing by analogy with what occurs in some species of crickets and cicadas. Groups of territorial males could have become more effective at attracting migrating females by extending the reach of their hooting beyond territorial boundaries during the "carnival display." The key to such an extension would be precise temporal superposition of voices, requiring predictive timing, enabled by synchrony to a common pulse (Merker et al. 2009), although such a suggestion must remain speculative in the absence of clear evidence.

It must be noted that none of the above hypotheses are mutually exclusive. Instead, different strands and factors may have been operative at different times. While behavioral, cognitive, neuroscientific, anthropological, archaeological, and ethological evidence can be used to narrow the possible problem space and make predictions concerning efficacy and general chronological ordering of various factors, these predictions may be testable by means of emerging genetic techniques. For example, the effects of sexual selection in the hominin lineage in the emergence of communicative behaviors may be tracked by exploring the prevalence of sexual dimorphism (not just behavioral, but also in terms of brain developmental control by sex steroids) by analyzing gene expression as new techniques are developed to interrogate the fossil DNA of coexisting hominin species.

Much of this discussion concerns the emergence of human communicative capacities without attempting to delineate why humans should have a plethora of communicative capacities at their disposal. While proximate, and in some instances, ultimate, adaptive functions have been sketched out for aspects of language and music, we must question why we possess at least two communicative systems that overlap so significantly in their operational characteristics.

We considered six ways of conceiving of the evolutionary relationships between language and music (Figure 21.1). While the figure appears to present language and music as discrete or unitary domains, each may best be conceived of as opportunistic conflations of a mosaic of preexisting or extant capacities which themselves have diverse origins. Nevertheless, the models have heuristic value in delineating possible evolutionary relationships between music and language, given their current status and in the light of likely precursor capacities.

Of those precursor capacities, it can be suggested that the most compelling candidate for the origin of language and music is the capacity for vocal learning. All vocal animals produce innate calls expressive of emotional states. In the case of elaborate calls (still innate) these are sometimes called song, as in nonvocal learning songbirds (suboscines) or gibbons. In addition, a subset of these callers acquires and produces learned song (oscine birds, some cetaceans, and humans).Finally, a single species (humans) add a third something,

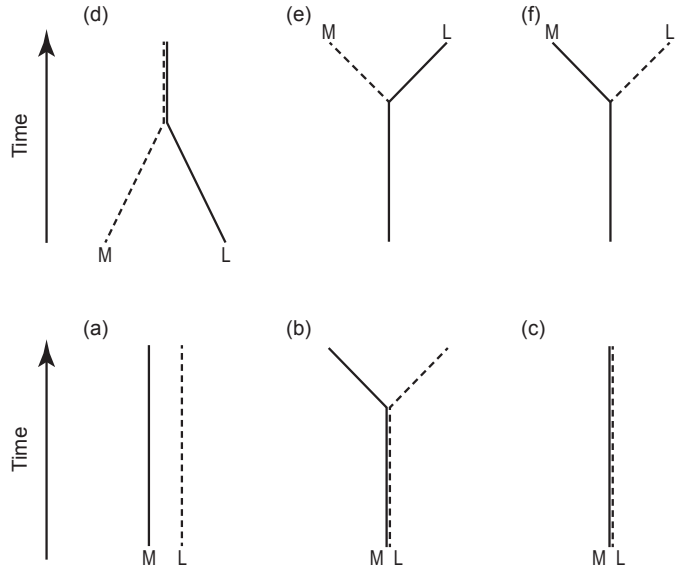

Figure 21.1 Six possible models for the evolutionary emergence of language (L) and music (M), with the timeline running from the bottom of the figure: (a) music and language have separate origins and remain distinct human faculties; (b) music and language have the same origin and diverge to become distinct faculties; (c) music and language have the same origin and remain indissociable; (d) music and language have separate and distinct origins and converge over time to share features; (e) language's origins precede those of music, which emerges as an offshoot (Herbert Spencer's view); (f) music's origins precede those of language, which emerges as an offshoot (Darwin's view).

dependent on the crux of the second (i.e., vocal production learning); namely spoken language and vocal music. All vocal learning species produce what has been interpreted by some as "music" in the form of complex sonic patternings ("song"). If one views vocal learning as providing a general form of "music" that has value in mediating social interactions but that does not embody propositionality, this may favor the last model (Figure 21.1f: language arising as a fairly late offshoot of music), with the emergence of language enabling semantic decompositionality and information transfer.

Even if we think in these terms, perhaps the distinction arises from a cultural bias, which would favor the third model (Figure 21.2c: common origins), with different cultures exploiting language and music for different ends. However, whether a culture distinguishes between language and music may not have the same perceptual consequences as cross-cultural differences in the use of color terms. It may be more relevant to aim to distinguish the ways in which music and language are bound to the evolution through natural selection of specific brain mechanisms or to processes of cultural evolution through the creation of ecological niches of cumulatively increasing social and artifactual complexity.

To understand other behaviors and capacities, broader contexts may be needed to assess the relationships between music and language. Perhaps phenomena such as language and music are different intersecting subsets of broader capacities, such as shared intentionality, or of general mimetic capacities. Moreover, to extrapolate from the kinds of enactment found in contemporary cultures to early human evolutionary history may well be unfounded. Such enactments work for contemporary humans because we are inclined to mine meaning from our physical and social environments. Perhaps that capacity, which involves a bird's eye view of the situation (in the form of a highly articulated theory of mind; Corradi-Del'Acqua et al. 2008), lies at the root of human communication. The emergence of a sense of self, a capacity to objectify ourselves and maintain a sense of self-continuity, and to relativize our experience of each other may underpin human communicative capacities.

Neither language nor music are purely vocal (or auditory); both constitute conceptual achievements that may be implemented by exploiting whatever tools are available at one's disposal (vocality, gesture, pantomime, external signing). Some of the traits that characterize both language and music, such as syntax and sequencing, are evident in other vocal learning species. The vocal part of those traits has been inherited in the production part of the neural circuitry subserving learned vocalization in humans. The issue is to understand why humans combine compositional semantics with their vocal learning whereas other species do not. We have seen that some gestural theories favor motor learning, based on pantomime, in the development of meaningful protolanguage as a scaffolding for vocal learning, rather than postulating that our ancestors first developed meaningless "song."

Revisiting the issue of humans' exquisite control over vocalization in contrast to chimpanzees, one can ask what allowed humans to gain that control. For instance, if a chimp consistently fakes its vocalizations, it is likely to be ignored. Assuming a similar tendency in our common ancestor, how did we start to control our vocalizations? One possibility is that through "performing" to out-groups—making sounds that are out of place to deter predators (cf. Hagen and Hammerstein 2005)—early hominins derived the ability for displaced reference that is central to the linguistic faculty. For example, among the contemporary Mbendjele forest-dwelling hunter-gatherers, women sing and co-talk in the forest to deceive other animals. That cooperative behavior drives bonding within the group, and the deception is oriented outside the group. Imitation skills, including nonconscious mimicry (Lakin et al. 2003), may be especially significant in the emergence of human cooperative and communicative capacities (Lewis 2009). If individual pleasure and group bonding derive from coordinated vocalization and movement, that would create pressure for more communication, with vocalizations and gestures moving from initially holistic (Wray 1998) or social (Dunbar 1996) significance to increasing analytic status. For example, in contemporary egalitarian societies based on sharing and absence of social hierarchicality,

explicit instruction is a claim to more knowledge and higher status and is thus rare. Most speech in such contemporary societies is "need-expression in the form of request," whereas much knowledge transmission is accomplished by means of pantomimetic display and mimicry (see Example 1 in the online supplemental information to this volume, http://esforum.de/sfr10/lewis.html: Mongemba's account of an elephant hunt), which highlight expressiveness rather than efficiency of information transmission. It is notable that participants may experience a form of "transportation" as consequence of pantomimic representation, as the interaction requires displacement of the experienced world, potentially providing a trigger for the emergence of propositionality. Were such gestural, mimetic and "displacing" interactions to have part of early hominin repertoires, then a general theory linking gestural and vocal language origins with pedagogical process appears viable.

Although language and music may be functionally differentiable, that difference may be marked in such a way as to indicate its origin. For example, play interactions in canids are marked by a "play bow" to signify that the social and physical consequences of the interaction—within limits—are to be discounted. Music's "lack of consequence"—the fact that engagement with others in music sanctions types of behavior which may be socially unacceptable in other contexts—seems parallel to play as a mode of social interaction. Perhaps music constitutes an offshoot of a common communicative faculty (see Figure 21.1f), emerging through pressures imposed by increasing altriciality to co-opt juvenile, exploratory modes of thought and behavior into the adult repertoire (Cross 2003a).

Irrespective of these considerations, the major obstacle to greater clarity in our understanding of the origins of music and language is our lack of knowledge of music in cultures other than those of the contemporary West and of its relationships to other aspects of culture, including language. Most cultures have been explored as linguistic cultures, not as linguistic and musical cultures. Our knowledge of the music of those cultures is simply not commensurable with our knowledge of the languages, in part because of a lack of consensus about the key elements of music that would allow for cross-cultural comparison (despite heroic but much-criticized efforts such as those of Lomax 1968; for a sympathetic critique see Feld 1984). Until we have a sample of the rich information required to elaborate a principled theory of the relationships between what appears, from a Western, "etic" (i.e., outsider) perspective, to constitute music and language (requiring close collaboration between culture members and a range of human sciences), it will be difficult to gain any certainty about the origins of these human capacities. Our understanding of the relationship between language and music may be even more limited than we think it is. Undoubtedly, the first music was based on the voice—a biologically significant timbre—and much music across cultures continues to be based on the voice. Music is also intrinsically linked to regular and entrained collective movement—dance—in many societies. It is surprising,

then, that most research on music cognition has used instrumental timbres, typically synthesized, rather than vocal timbre, and has only in recent years begun to explore music in the context of individual and collective movement. Recent work indicates, however, that adults remember melodies better when they are presented vocally (on the syllable "la") rather than instrumentally (Weiss et al. 2012), and that joint movement, in the form of dance, can enhance memory for person attributes (Woolhouse, Tidhar, and Cross, in preparation).

Triadic Niche Construction in Relation to Music and Language Origins

A significant role in any exploration and explanation of language-music relationships is likely to be played by Iriki's theory of triadic niche construction (Iriki and Taoka 2012; Arbib and Iriki, this volume). A niche is a fragment of available environmental resources, and the process of ecological niche construction is a modification implemented by an animal to create his own niche. The interaction between the activity of the organism and its environment changes the environment, thereby changing selective pressures acting on the organism. In classic niche construction theory, there is a two-way interaction between behavior and environment. Quallo et al. (2009) have found that tool-use training in macaques led to an expansion in gray matter volume, affording extra neural machinery for the brain. This expansion in brain volume constitutes a "neural niche"—a newly available resource in the form of extra brain tissue—for future exploitation, affording the organism an increased range of responses (i.e., a "cognitive niche") introducing selective pressures which could, under some circumstances, amplify evolutionary effects.

This triad of neural niche, cognitive niche, and ecological niche are all operational for humans, allowing for an acceleration of their interaction in the course of our evolution, behavioral changes opening the door for later genetic changes. In effect, by changing the context of selection, different selection pressures come into play which may afford the possibility for new types of genetic change. If the information generated in the interaction is embedded in the structure of the environment, then it may be inherited by the next generation. In the context of human evolution, it could then be postulated that post-reproductive survival—the "grandmother" hypothesis—together with a means of transmitting knowledge critical to survival (e.g., such as language, or more particularly, mimetic and musical modes of presentation, display and participation) can allow the genetic pathway to be bypassed in the transmission of skill (Iriki 2010; Iriki and Taoka 2012). This would afford time for genetic assimilation, if it is necessary in the hominin lineage. This "Baldwinian evolution"—a mechanism that initially induced modification within the range of preprogrammed adaptation, and is then available for later mutations to optimize it—would be particularly beneficial for species with long life spans

and low birth rates (e.g., in primates with humans at the extreme, who need to survive evolutionarily significant contingencies through an individual capacity to adapt). This stands in sharp contrast to species with short life spans and mass reproduction, which adapt to environmental changes through variations in their numerous offspring, expecting at least a few to survive. Both of these mechanisms, however, would aid the adaptive radiation of the species in the terrestrial ecosystem.

Comparative and Phylogenetic Issues

While triadic niche construction provides an extremely promising candidate mechanism for establishing and consolidating language and music in the human communicative repertoire, an exploration of origins requires consideration of evidence from beyond the hominin clade so as to avoid being blinkered by unacknowledged anthropocentrism (Figure 21.2a). Processes, structure, and behaviors in other species that are homologous to or convergent with those implicated in music and language are informative about their bases and manifestations in humans; after all, identification of sub-components of these complex capacities may be more directly observable in some nonhuman species. The concept of genetic or deep homology (see Fitch and Jarvis, this volume)—a genetic basis for behavioral capacities that may be common across different lineages, evidenced in the recruitment (particularly in ontogeny) of similar sets of complex genes to subserve similar functions—has significant potential to elucidate connections between types of behavioral capacity in different species: thoses which do not originate from a common ancestor as well as those that may be simply convergent, motivated by environmental selection pressures that operate on distantly related organisms to exploit specific types of environmental niche (Figure 21.2b).

While evolution is not progressive, there is a clear trend, at least in some lineages, toward increasing complexity, particularly in the hominin line. However, that complexity should not be considered independently of the systems that implement or enable it. With respect to song and language, when we compare, for example, a songbird with a human, we must first decide whether there is common design and then ask: How did these things emerge? Homology (i.e., the explanation that is likely the first port of call in answering the question) can be specified as either behavioral, anatomical or structural, developmental, or genetic (deep), this latter being evident in the common role played by certain genes (such as PAX6 in vision or FOXP2 in vocalization: see, e.g., White et al. 2006; Fernald 2000) in very distantly related species. We note, however, that a genetic network could have been recruited independently in two different species and may have functioned differently in different ancestor species. In the case of songbirds and humans, behavioral relationships in vocal capacity are clearly analogous rather than homologous, but may be motivated

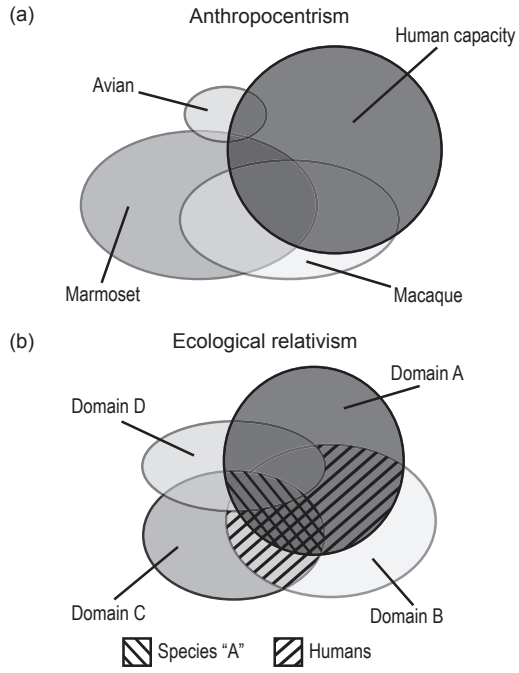

Figure 21.2 Venn diagrams depicting relationships among various cognitive capacities of different species. Sets are classified by (a) species (anthropocentrism) or (b) cognitive domains (ecological relativisim). In the anthropocentric view (a), cognitive domains are expressed as subsets within respective species set, partly overlapping with other species. In this way, humans tend to privilege only those included in the human set, making it difficult to recognize that other species may have cognitive abilities superior to humans (as shaded outside the "human set"). This perspective can lead to the misleading perception that nonhuman species are intrinsically inferior to humans. In contrast, when sets are classified by cognitive domains (b), species are depicted through a combination of subsets to illustrate inter-relationships between species' capacities. These cognitive domains and their combinations in species must be considered to have evolved through interactions with ecological conditions of habitats, thus, ecological relativism.

by deep homologies at the genetic level that afford the emergence of similarly functioning neural circuitry recruited for species-specific ends.

Hence it is possible to view aspects of the origins of music and language as embedded in a deep homology that is manifested at the genetic level; convergence may be occurring at the organ level (larynx in humans, syrinx in birds) but homology at the genetic level. The vocality that underpins speech and music may have deep homology across all vocal learners, with the motor learning circuitry being co-opted independently for vocal learning in different species. Nevertheless, vocal learning is only one of the constellations of features that can be identified as underpinning language and music. Humans'

complex sociality, excessive brain (cortex) size, and capacity for cultural conservation and transformation of knowledge all seem likely to have played a significant role in shaping our communicative capacities. It would be highly desirable to track the extent to which those aspects shared with our closest nonhuman relative represent true homologies. However, we are limited by a lack of knowledge of primate evolution immediately prior to our last common ancestor, whose capacities must be extrapolated (perhaps uninformatively) from those of their descendants. Nevertheless, even in absence of such data it might be possible to use datable divergences between existing nonhuman primate species to explore human cognitive functions such as language and music. For example, new world monkeys may provide a fertile experimental model as they have a wide range of vocal capacities as well as cooperative social structures. In the "old world," humans established their unique niche by dividing resources with other primates—apes and old world monkeys. In contrast, in the "new world," where humans did not exist, adaptive radiation should have developed differently. That is, the traits which characterize human-specific cognition, of which precursors should have derived from common ancestors and become extinct in nonhuman old world primates, might have preserved in the new world monkey lineages by deep homology and could be expressed in extant taxa through epigenetic interactions as convergent evolution. As such, new world monkeys could represent an ideal animal model to study various aspects of human-specific higher cognitive functions.

Conclusion

To return to the point made at the outset: when considering relationships between language and music from cultural and evolutionary perspectives, there is a pressing need to avoid presentist and anthropocentric biases in making inferences about cultural categories and evolutionary trajectories. Music and language may be different domains of human thought and behavior; they may be different manifestations of the same underlying capacities; or they may be the same suite of communicative capacities co-opted for different ends in different situations. They may have evolved separately or conjointly, or they may have merged or split over the course of human evolution. They or their subcomponents may be present in the repertoire of other species, or they may be unique to humans. Only by synthesizing evidence from the whole range of human sciences, in the context of investigations that are alert to cross-cultural differences in the conceptualization and implementation of communicative skills and the features shared with other species, can we achieve a degree of defensible clarity in our understanding.