3

Cross-Cultural Universals and Communication Structures

Stephen C. Levinson

Abstract

Given the diversity of languages, it is unlikely that the human capacity for language resides in rich universal syntactic machinery. More likely, it resides centrally in the capacity for vocal learning combined with a distinctive ethology for communicative interaction, which together (no doubt with other capacities) make diverse languages learnable. This chapter focuses on face-to-face communication, which is characterized by the mapping of sounds and multimodal signals onto speech acts and which can be deeply recursively embedded in interaction structure, suggesting an interactive origin for complex syntax. These actions are recognized through Gricean intention recognition, which is a kind of "mirroring" or simulation distinct from the classic mirror neuron system. The multimodality of conversational interaction makes evident the involvement of body, hand, and mouth, where the burden on these can be shifted, as in the use of speech and gesture, or hands and face in sign languages. Such shifts having taken place during the course of human evolution. All this suggests a slightly different approach to the mystery of music, whose origins should also be sought in joint action, albeit with a shift from turn-taking to simultaneous expression, and with an affective quality that may tap ancient sources residual in primate vocalization. The deep connection of language to music can best be seen in the only universal form of music, namely song.

Introduction

To approach the issues surrounding the relationship between language and music tangentially, I argue that the language sciences have largely misconstrued the nature of their object of study. When language is correctly repositioned as a quite elaborate cultural superstructure resting on two biological columns, as it were, the relationship to music looks rather different.

This chapter puts forth the following controversial position: Languages vary too much for the idea of "universal grammar" to offer any solid explanation of our exclusive language capacity. Instead we need to look directly for our

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

biological endowment *for* language, communication, and culture. Part of this may involve the neural circuitry that is activated in language use (see Hagoort and Poeppel, this volume), although the innate nature of this is still unresolved, since it apparently develops in part parallel to the learning of language (Brauer et al. 2011b). Two systems, however, clearly contribute to our native language-ready capacities: (a) an evolved set of interactive abilities, which makes it possible to learn the cultural traditions we call languages, and (b) a specialized vocal-learning system (an auditory-vocal loop). These two systems have distinct neurocognitive bases and different phylogenetic histories. Judging from traces of parallel material culture, system (a) is well over 1.5 million years old—a time period when system (b) was not yet in place. Here I concentrate on system (a), our interactive abilities, because its contribution to linguistic capacity has not been properly appreciated. I begin with a brief description of this story and then explore its implications for language, music, and their interrelation.

Language Diversity and Its Implications

Let us begin with the observation that human communication systems are unique in the animal world in varying across social groups on every level of form and meaning. There are some 7000 languages, each differing in sound systems, syntax, word formation, and meaning distinctions. New information about the range of language diversity and its historical origins has undercut the view that diversity is tightly constrained by "universal grammar" or a language-specialized faculty or mental module (Evans and Levinson 2009). Common misconceptions, enshrined in the generative approach to language universals, are that all languages use syntactic phrase structure as the essential foundation for expressing, for example, grammatical relations, or that all languages use CV syllables (a Consonant followed by a Vowel), or have the same basic set of word classes (e.g., noun, verb, adjective). Instead, some languages make little or no use of surface phrase structure or immediate constituents, not all languages use CV syllables, and many languages have word classes (like ideophones, classifiers) that are not found in European languages. The entire apparatus of generative grammar fails to have purchase in languages that lack phrase structure (e.g., the so-called "binding conditions" that control the distribution of reflexives and reciprocals). Nearly all language universals posited by generative grammarians have exceptions in one set of languages or another.

The other main approach to language universals, due to Greenberg (1966), escapes this dilemma by aiming for strong statistical tendencies rather than exceptionless structural constraints. The claim would then be that if most languages follow a tendency for specific structures, this reflects important biases in human cognition. Greenberg suggested, for example, that languages tend to have "harmonic" word orders, so that if a language has the verb at the end of the clause, it will have postpositions that follow the noun phrase (rather than

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

prepositions that precede the noun phrase). Dryer (2008) has recently tested a great range of such predictions, with apparently good support. These generalizations rely on sampling many languages, both related and unrelated; one can hardly avoid related languages because a few large language families account for most of the languages of the world. One problem that then arises is that related languages may be similar just because they have inherited a pattern from a common ancestor. One recent solution has been to control for relatedness by looking at, for example, word order wholly within large language families. It turns out that the Greenbergian generalizations about harmonic word order do not hold: language change within language families often does not respect the postulated strong biases, and language families show distinctly different tendencies of their own (Dunn et al. 2011).

The upshot is that although there are clear tendencies for languages to have certain structural configurations, much of this patterning may be due entirely to cultural evolution (i.e., to inheritance and elaboration during the processes of historical language change and diversification). All the languages of the world outside Africa ultimately derive (judging from genetic bottlenecks) from a very small number that left Africa at the time of the diaspora of modern humans not later than ca. 70,000 years ago, with the possible proviso that interbreeding with Neanderthals and Denisovans (now known to have occurred) could have amplified the original diversity (Dediu and Levinson 2013).[1]

There are three important implications. First, we have underestimated the power of cultural transmission: using modern bioinformatic techniques we can now show that languages can retain strong signals of cultural phylogeny for 10,000 years or more (Dunn et al. 2005; Pagel 2009). Consequently, language variation may tell us more about historical process than about innate constraints on the language capacity. Those seeking parallels between music and language be warned: in neither case do we have a clear overview of the full range of diverse cultural traditions, universal tendencies within each domain, and intrinsic connections across those tendencies. Over the last five years, linguists have made significant progress in compiling databases reflecting (as yet still in a patchy way) perhaps a third of the linguistic diversity in the world, but no corresponding database of ethnomusicological variation is even in progress.[2]

Second, the observed diversity is inconsistent with an innate "language capacity" or universal grammar, which specifies the structure of human language in anything like the detail imagined, for example, by the "government and binding" or "principles and parameters" frameworks in linguistics (Chomsky

[1] The recently discovered Denisovans were a sister clade to Neanderthals, present in eastern Eurasia ca. 50 kya; they contributed genes to present-day Papuans, just like Neanderthals did to western Eurasians.

[2] Useful leads will be found in Patel (2008:17ff), who points out that cultural variability in scale structure (numbers from 2–7 tones, differently spaced) and rhythm (2008:97ff) makes strong universals impossible. See also Nettl (2000), who quotes approvingly Herzog's title "Musical Dialects: A Non-Universal Language."

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

1981; Baker 2001). Even the scaled-back Minimalist program makes claims about phrase structure that are ill-fitted to the diversity of languages. There is no doubt a general "language readiness" special to the species, but this does not seem well captured by the major existing linguistic frameworks. We seem to be left with general architectural properties of languages (e.g., the mapping of phonology to syntax, and syntax to semantics), abstract Hockettian "design features," and perhaps with stronger universals at the sound and meaning ends (i.e., phonology and semantics, the latter pretty unexplored) than in morphosyntax.

Third, since the diversity rules out most proposed linguistic universals, we need to look elsewhere than "universal grammar" for the specific biological endowment that makes language possible for humans and not, apparently, for any other species. Apart from our general cognitive capacities, the most obvious feature is the anatomy and neurocognition of the vocal apparatus, and our vocal-learning abilities, rare or even unique among the primates.[3] These input/output specializations may drive the corresponding neurocognition, the loop between motor areas and the temporal lobe. They may even, during human development, help build the arcuate fasciculus (the fiber bundle that links the frontal lobes with the temporal lobes, i.e., very approximately, Broca's and Wernicke's areas; Brauer et al. 2011b).[4] The neural circuitry involved in language processing may thus have been "recycled," rather than evolved, for the function (Dehaene and Cohen 2007).

Only slightly less obvious is a set of abilities and propensities that are the essential precondition for language: advanced theory of mind (ToM) and cooperative motivations and abilities for coordinated interaction, which together form the background to social learning and makes possible both culture in general and the learning of specific languages. These aspects of cognition and, in particular, the grasp of Gricean communication ($\text{meaning}_{\text{nn}}$) seem to have their own neural circuitry, distinct from vocal circuitry and mirror neuron circuitry (Noordzij et al. 2010).[5] These interactional abilities are much more central to language than previously thought; together with vocal learning, they provide the essential platform both for cultural elaboration of language and for infants to bootstrap themselves into the local linguistic system. Correspondingly, they may play some

[3] See, however, Masataka and Fujita (1989) for monkey parallels.

[4] The crucial experiment—checking on the development of these structures in deaf children and home-signers—has not to my knowledge been done.

[5] Gricean signaling (producing a noninstrumental action whose sole purpose is to have its intention recognized) is a kind of second-order mirror system: perception (decoding) depends on (a) seeing the noninstrumental character of the signal, and (b) simulating what effect the signaler intended to cause in the recipient just by recognition of that intention. Consider my signaling to you at breakfast that you have egg on your chin just by energetically rubbing my chin: a first-order mirror interpretation is that I have, say, egg on my chin; a second-order one is that I'm telling you that it is on your chin.

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

parallel role in musical learning and performance. They have preceded modern language in both ontogeny and phylogeny, which we turn to next.

The Timescale of the Evolution of Speech and Language

In a metastudy drawing on the most recent discoveries, we have argued that the origins of these vocal abilities can be traced back over half a million years to *Homo heidelbergensis*, who exhibited a modern human vocal tract, modern breathing control, modern audiograms, and the FOXP2 variants inherited in common by his descendants: Neanderthals, Denisovans, and modern humans (Dediu and Levinson 2013). At 1.6 mya, *H. erectus* lacked these vocal specializations but exhibited control of fire and complex tool traditions (the Acheulian or Mode 2 type), arguing for a communication system able to support advanced cultural learning. Such a system presupposes the cooperative interaction style of humans, which in turn relies on advanced ToM capacities. Therefore, *H. erectus* (or *H. ergaster* as some prefer to call the African variant) had some quite advanced form of language that was less vocally specialized than that used by *H. heidelbergensis*. *H. erectus*, in turn, is the presumed descendant of *H. habilis*, who already used a varied stone tool kit at 2.5 mya, with the first stone tools in use as early as 3.4 mya. Thus social learning and cooperative communication have deep phylogenetic roots.

Speech and language, as we know them, evolved in the million years time between 1.5 mya–0.5 mya (Dediu and Levinson 2013). Modern language is thus of a much greater antiquity than usually assumed (Klein 1999; Chomsky 2007, 2010 presume the last 50–100 kya). Nevertheless, from a geological or genetic time perspective, a million years pales in comparison to the 50 million years existence of birdsong or bat echolocation: language has been able to develop so fast because much of its complexity was outsourced to cultural evolution. There is, however, a deep biological infrastructure for human language: the vocal-auditory system, on the one hand, and the cooperative communicative instincts, on the other, on which cultural elaboration is based.

Most importantly, a fully cooperative communication system and interaction style evolved gradually in our line over the three million years of tool use leading up to *H. heidelbergensis*. Judging from the development of material culture, ToM capacities and advanced cooperative abilities are very ancient, tied to increasing encephalization and group size. They are the crib for language in both phylogeny and ontogeny.

Two controversial issues should be raised here. First, the Darwinian view that speech evolved not for language but for musical use, with songbirds as the animal model, has recently been revived, for example, by Mithen (2005) and Fitch (2009b). This view is, to my mind, a nonstarter. As just explained, our speech system evolved after at least a million years of functional communication geared to handing on cultural learning and tool traditions. The

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

preconditions for culture involved prolonged infancy, intensive cooperative social interaction, and the large social groups that motivated increased encephalization. It is not plausible that all this developed without some kind of protolanguage. Thus language (perhaps in gestural form) preceded speech, not the other way around as Darwin had imagined. Thus, songbirds probably do not provide the right analogy; vocal learners among more social species (e.g., sea mammals) may provide a better animal model (for a contrary view, see Fitch and Jarvis, this volume).

The second controversial issue is indeed the possible gestural origin of language (cf. Arbib 2005a, b). If language was carried by a medium other than fully articulate and voluntarily controlled speech for a million years or more, gesture is a prime candidate. Call and Tomasello (2007) make a good case for gesture being the voluntary, flexible medium of communication for apes, with vocal calls being more reflex. On phylogenetic grounds, then, one might indeed argue for gestural precursors to language. However, first, there is no specialization of the hand for communication that parallels the evolution of the vocal system, which one would expect if it played such a crucial role. Second, the human communication system is properly thought of as based on hand + mouth, allowing greater loading of the hand (as in sign languages) or of the mouth (as in spoken languages), but always involving both. It is therefore likely that this joint system has great antiquity. What seems plausible is that during the million years preceding *H. heidelbergensis* (by which time speech was fully formed), the burden of communication was shifted relatively from hand to mouth. More generally, human communication is intrinsically multimodal, as reflected, for example, in the general purpose nature of Broca's area (Hagoort 2005).

The Interactive Niche

Every language is learned in face-to-face interaction in a special context that is unique to the species. Most animals avert gaze except in aggression; in contrast, within restrictions, mutual gaze is tolerated or even required in many kinds of human interaction (for a cross-cultural study, see Rossano et al. 2009). This is a token of the *presumption of cooperation* which operates (again with limitations) in intragroup human interaction—a persistent puzzle from an evolutionary point of view (Boyd and Richerson 2005). Under this cooperative envelope, interaction consists of a sequenced exchange of actions, following specific turn-taking rules geared to the structure of minimal contributions (e.g., clauses in spoken communication), and which permit one action at a time (see Sacks et al. 1974). The expectation is that each such action unit is tied to the prior one by a "logic" of action: a request is met with a compliance or denial; a greeting by a greeting; a question with an answer or evasion; a pointing by a gaze following; and so forth. The structure of action sequences can be complex,

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

arguably as or more complex than anything seen in natural language syntax, as we shall see. Yet an elementary system of this kind is visible in the earliest prelinguistic mother–infant interaction ("proto-conversation," Bruner 1975; for resonances in the musical domain, see Malloch and Trevarthen 2009).

This interactional envelope is the context in which the great bulk of language use occurs; monologue is the exception, and in some societies hardly occurs at all. Narrative, likewise, plays a small role, statistically speaking, in the use of language. The basic niche for language is the tit-for-tat of informal conversation: action–response, action–response, and so on. It is intriguing to wonder what the equivalent natural ecological niche for music might be; perhaps Western music, with its division of performers and audience, is entirely misleading, like comparing lecturing to the natural niche for language use.

Two aspects of this interactional envelope are much more complex and intricate than meets the eye. The first is turn-taking. The fact that turns at talking alternate seems at first quite trivial, but consider this: the gap between turns is on average 200 ms across a wide variety of languages (and the mode offset between turns is 0 ms, without any gap at all, in all languages tested; Stivers et al. 2009). Since it takes at least 600 ms to crank up the speech production system, speakers must be anticipating the last words of their interlocutors' turns; they must also predict the content in order to respond appropriately (direct EEG measurement suggests actual launch of production is quite a bit earlier than this on average). The whole system is built on predicting what the other will say part way through the saying of it, and since what is said has all the open-endedness that syntax delivers, this is no mean feat. This system exerts tremendous cognitive demands: comprehension and production must run in parallel, at least part of the time (see Figure 3.1). It is perhaps not fanciful to imagine that this universal pace or fixed metabolism of the turn-taking system was set up early in the phylogeny of protolanguage, so that we inherit a system ill-adapted to the complexity of the structures we now thrust into these short turns.

The second complexity is action sequencing: questions expect answers, requests compliance, offers acceptances, etc. This requires recognizing a turn as a question, early enough in its production to allow time to formulate the response. This recognition might be imagined to be based on syntax or lexical cues, but corpus work shows, for example, that most questions in English are yes-no questions, and most of these are in declarative form with falling intonation. So, most questions are recognized by means other than direct linguistic

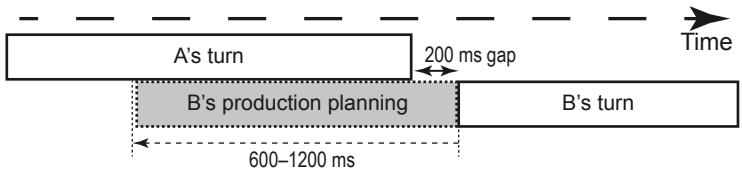

Figure 3.1 Overlap of comprehension and production processes in conversation.

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

cues, for example by noting that B is making a statement about a subject for which he knows I have more knowledge (e.g., "You've had breakfast"). The same holds for most kinds of speech acts: they don't come wrapped in some canonical flag. This problem of "indirect speech acts" has been neglected since the 1970s, but it is *the* fundamental comprehension problem: the speech act is what the hearer needs to extract to respond in the tight temporal frame required by the turn-taking system. Likewise, the whole function of language is often misconstrued: the job of language is not to deliver abstract propositions but to deliver speech acts.

Since the job of language is to deliver actions explains, of course, why speech comes interleaved with nonverbal actions in any ordinary interchange: I say "Hi"; you smile and ring up my purchases saying, "You know you can get two of these for the price of one"; I explain that one will do and hand you a bill; you say "Have a good day." Words and deeds are the same kind of interactional currency, which is why language understanding is just a special kind of action understanding. In cooperative interaction, responses are aimed at the underlying action goals or plans. Consider the telephone call in Example 3.1 (Schegloff 2007:30):

1.	Caller:	Hi.	
2.	Responder:	Hi.	
3.	Caller:	Whatcha doin'?	
4.	Responder:	Not much.	(3.1)
5.	Caller:	Y'wanna drink?	
6.	Responder:	Yeah.	
7.	Caller:	Okay.	

Line 3 might look like an idle query, but it is not treated as one: the response "Not much" clearly foresees the upcoming invitation in line 5 and makes clear that there is not much impediment to such an invitation, which is then naturally forthcoming. Conversation analysts call turns like line 3 a "pre-invitation," and show with recurrent examples that such turns have the character of conditional offers. Just as the caller can hardly say at line 5 "Oh just asking," so the responder will find it hard to refuse the invitation she has encouraged by giving the "go-ahead" at line 4 (although a counterproposal might be in order at line 6). The underlying structure of such a simple interchange, translated into hierarchical action goals, might look something like the sketch in Figure 3.2, where *Whatcha doin'* acquires its pre-invitation character from a projection of what it might be leading to.

Action attribution thus plays a key role in the use of language and is based quite largely on unobservables, like the adumbrated next action if I respond to this one in such a way. The process is clearly based on advanced ToM capacities, and beyond that on the presumption that my interlocutor has designed his turn precisely to be transparent in this regard. This, of course, is Grice's insight, his theory of meaning$_{nn}$: human communicative signals work

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

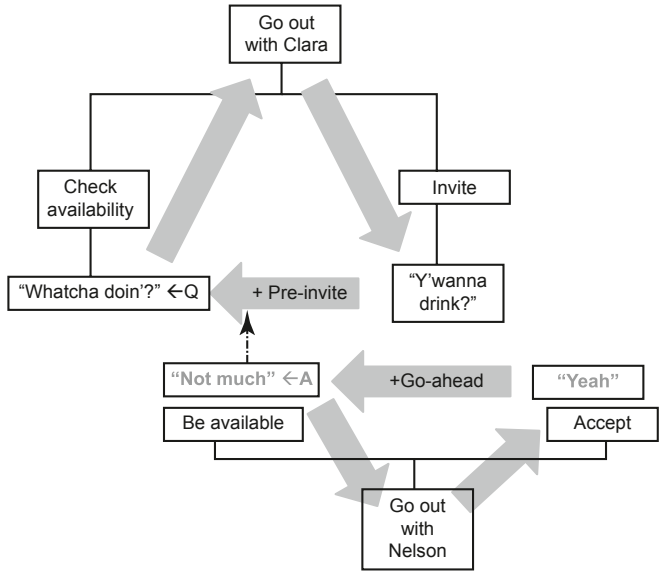

Figure 3.2 Action assignment based on plan recognition (see Example 3.1). Arrows indicate direction of inference from what is said to what is likely to come up next, which then "colors" the interpretation of the present turn.

by presenting an action designed to have its intention recognized, where that recognition exhausts the intention. In recent work, we have tried to isolate the neural circuitry involved in just this process and find it distinct from either the language circuitry or the mirror neuron circuits (Noordzij et al. 2010): we find overlapping areas of activation in the posterior superior temporal sulcus, in mirror-like fashion, in both signaler and receiver, interpreted as signaler's simulation of recipient's inferencing.[6]

The inferential character of action ascription makes it a complex process. However, an unexpected further order of complexity is that it has a quasi

[6] As a way to generalize over these observations and the classic mirror neuron system, it may be helpful to think (in a slightly different way than Arbib 2005b) of a *hierarchy of action–perception mirror loops*, as follows:

degree 0 (intra-organism): *Action–perception feedback*, as in proprioception or auditory feedback of one's own production, allows cybernetic feedback. Highly evolved systems include echolocation in bats and cetaceans.

degree 1 (cross-organism): *Classic mirror neuron system:* other's action recognition and self-action use overlapping neural resources. This can be further distinguished into degree 1.1 instinctive systems and degree 1.2 learned systems. Mouth mirror neurons might offer a route to vocal learning (Arbib 2005b:118).

degree 2 (cross-organism): *Gricean simulation systems:* applies to actions that self-advertise that they are signals (noninstrumental actions), so *discounting* mirror neuron systems of degree 1. Works by the recipient simulating what the signaler calculated the recipient would think/feel (that being the noninstrumental intention).

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

syntax (Levinson 1981). Consider the following simple exchange in Example 3.2 (Merritt 1976):

A:	Q_1	"May I have a bottle of Mich?"	
B:	Q_2	"Are you twenty one?"	
A:	A_2	"No."	(3.2)
B:	A_1	"No."	

This has a pushdown stack character: Q_1 is paired with A_1, but Q_2–A_2 intervenes. Many further levels of embedding are possible, and they can be characterized, of course, by the phrase-structure-grammar in Example 3.3:

$$\begin{aligned} &Q\&A \rightarrow Q\ (Q\&A)\ A \\ &Q\&A \rightarrow Q\ A \end{aligned} \qquad (3.3)$$

What is interesting is that this kind of center embedding has been thought to be one of the pinnacles of human language syntax. An exhaustive search of all available large language corpora has yielded, however, the following finding: the greatest number of recursive center embeddings in spoken languages is precisely 2, whereas in written languages the number is maximally 3 (Karlsson [2007] has found exactly 13 cases in the whole of Western literature).[7]

In contrast, it is trivial to find examples of center embeddings of 3 or greater depth in interaction structure. Example 3.4 (abbreviated from Levinson 1983:305) shows one enquiry embedded within another, and a "hold-OK" sequence (labeled 3) within that:

C: .. I ordered some paint… some vermillion… And I wanted to order some more, the name's Boyd
- R: .. Yes how many tubes would you like sir?
  - C: .. What's the price now with VAT?
    - R: I'll just work that out for you
      - [3] C: Thanks
      - [2] (10.0)
    - [1] R: Three pounds nineteen a tube sir
      - C: Three nineteen is it=
      - R: Yeah
      - C: That's for the large tube?
      - R: Well yeah it's the 37 ccs
  - C: I'll just eh ring you back I have to work out how many I'll need

[0] (3.4)

Examples 6 deep or more can arguably be found in conversation. When one finds a domain in which a cognitive facility is most enhanced, it is reasonable to assume that this is the home in which it originally developed. The implication is that core recursion—as expressed in center embedding—has its origin

[7] More precisely, Karlsson (2007) calls one center embedding "degree 1," a center embedding within a center embedding "degree 2," and shows that degree 2 is the maximal attested depth for spoken languages, degree 3 for written texts.

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

in interaction systems, not in natural language syntax. Exactly parallel arguments can, I believe, be made for so-called cross-serial dependencies, vanishingly rare in syntax but exhibited recurrently in conversational structure.[8] Why exactly it is so much easier to keep track of discontinuous dependencies in joint action than in solitary performance remains unclear; the mental registers required would seem to be the same, but the distributed production clearly helps cognition in some way.

More generally, the implication is that we have minds engineered for extraordinary coordination in joint action (see Fogassi, this volume). It may be interesting to reconsider music in this light: not as, in origin, a solitary enterprise or a performance to a passive audience, but as an interchange between actors, where guessing the next phrase is crucial to coming in on time, where one performer "answers" another, where the basic units are seen as "actions" rather than formal objects, and where extremes of coordination carry a deep satisfaction of their own (see Janata and Parsons, this volume). This suggests improvisational jazz as the model, not the sonata or the lullaby.

Language and Music

Sixty years ago the great anthropologist Levi-Strauss pointed out that music is the central mystery of anthropology.[9] Nothing has changed since. Contrary to the Chomskyan idea that language is a late evolutionary freak, a spandrel from some other evolutionary development, the fossil and archaeological record actually shows a steady, slow accumulation of culture which was only made possible by some increasingly sophisticated mode of communication, already essentially modern and primarily in the vocal channel by 0.5 million years ago (pretty much as Pinker and Bloom 1990 imagined on more slender evidence). But what is the story for music?

To what extent could music be parasitic on language, or more broadly on our communicative repertoire? First, some basic points. In small-scale societies with simple technology, music often equals song: that is to say, music only occurs with language. It is often imagined that music always involves instruments, but again small-scale societies often have no instruments, in some cases also avoiding any form of ancillary percussion (as in the elaborate, but purely vocal, range of song styles of Rossel Island, Papua New Guinea).[10] Phoneticians often distinguish language, prosody, and paralanguage, where the latter two are suprasegmental properties of speech (roughly tonal and wide

[8] Cross-serial dependencies have the form A1–B1–A2–B2 where the linkages cross over. Example 3.1 contains such a pattern, but I leave that as an exercise for the reader.

[9] Compare Darwin (1871:333) "As neither the enjoyment nor the capacity of producing musical notes are faculties of the least direct use to man in reference to his ordinary habits of life, they must be ranked amongst the most mysterious with which he is endowed."

[10] The Rossel Island observations come from my own ethnographic work.

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

timbre qualities, respectively), only partially bound into the linguistic system in rule-governed ways (see Ladd, this volume). Song is in a sense just language in a special, marked suprasegmental register or style or genre. Rossel Islanders, for example, do not have a category of "music" that would place each of their named types of song style (e.g., *tpile we*, "operetta"; *ntamê*, "sacred chants"; *yaa*, "laments") in opposition to speech of other types (e.g., *wii*, "fast-declaimed poetry"). It is thought-provoking to realize that "music" seems to be an ethnocentric category (Nettl 2000:466).

We are hampered, as mentioned, by having no ethnomusical databases that cover the world, but it is likely that song is in the unmarked case not a solo performance, but a joint activity involving a chorus (Nettl 2000).[11] Most of the song styles on Rossel Island, for example, are joint performances sung in unison, with the exception of laments (*yaa*) which are composed and sung by individuals, typically at funerals. This contrasts with normal conversation, which is composed from individual, short turns with rapid alternation. Song is thus a marked genre, in being predominantly jointly performed in unison (which is a rare, but observable occurrence in conversation, as in greetings or joint laughter). In some circumstances, but not all, song is like speech-giving, a performance by a set of performers with a designated audience. Linguistic systems make a lot of distinctions between speakers, addressees, auditors, and the like, originally explored by Goffman (1981, Chapter 3). For example, when I say, "The next candidate is to come in now," the syntax projects a second speech event, indicating that I am instructing you to go and ask the candidate on my behalf (see Levinson 1988). The same distinctions are relevant for song: both a song and a speech may be authored by one individual on behalf of another (the principal) and performed by a third (as in the praise songs of West Africa; see Charry 2000). In Rossel Island laments, author, principal, and mouthpiece are identical; sacred hymns (*ntamê*), however, are composed by the gods and sung by elder males to a precise formula, to a male-only audience.

Song, surely the original form of music,[12] makes clear the possibly parasitic nature of music on language: the tonal and rhythmic structure must to some extent be fitted to the structure of the language. The language of the lyrics determines both aspects of the fine-grained structure, the affectual quality matched to the words, and the overall structure, for example, the timing of subunits and nature of the ending (e.g., the number of verses).

The perspective adopted here, emphasizing the role of language in its primordial conversational niche, also suggests a possible take on the cultural (and possibly biological) evolution of music. The motivation for and structural complexity of music may have its origins in joint action rather than in abstract representations or solitary mentation. It may also rely on Gricean reflexive

[11] Patel (2008:371), however, reports one Papuan society where song is largely private and covert.

[12] The assumption makes the prediction that no cultural system of musical genres will be found without song genres (see Nettl 2000).

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

mirroring or simulation to achieve the empathy that seems to drive it,[13] together with the apparently magical coordination through prediction which is one source of the pleasure it gives. The rhythmic properties may owe at least something to the rapidity of turn-taking, the underlying mental metabolism, and the interactional rhythms that are set up by turn-taking. The multimodality of human communication allows the natural recruitment of additional channels, whether multiple voices or instrumental accompaniments, and of course dance (Janata and Parsons, this volume).

Still, few will be satisfied with the notion that music is, even in origin, just a special kind of speech (see Wallin et al. 2000; Morley 2011). They will point to the existence of (largely) independent cultural traditions of instrumental music, to the special periodic rhythms of music, and to its hotline to our emotions (Patel 2008). One speculation might be based on the Call and Tomasello (2007:222) argument that in the Hominidae, with the sole exception of humans, vocal calls are instinctive, reflex, and affectual ("Vocalizations are typically hardwired and used with very little flexibility, broadcast loudly to much of the social group at once—who are then infected with the emotion;" see also Scherer, this volume), in contrast to the gestural system which is more intentionally communicative and socially manipulative. If language began in the gestural channel and slowly, between 1.5–0.5 million years ago, moved more prominently into the vocal channel, it is possible that the vocal channel retains an ancient, involuntary, affective substrate. Note, for example, how laughter and crying exhibit periodic rhythms of a kind not found in language and are exempt from the regular turn-taking of speech. It could be this substrate to which music appeals, using all the artifice that culture has devised to titillate this system. Drugs—culturally developed chemicals—work by stimulating some preexisting reward centers. The ancient affective call system could be the addiction music feeds.[14]

Conclusion

The theory of language, properly reconstructed, yields much of the complexity of linguistic structure over to cultural evolution, seeking biological roots primarily in the auditory-vocal system and the species-special form of

[13] Gricean reflexive intentions may play a larger role in music than is obvious: Performers "work" an audience, intending to induce a feeling partly by affective evocation, but partly by getting the audience to realize that is what they are trying to do. This accounts for the difference between a recording and a live performance—the performer tries to persuade the particular audience to adopt the affective state intended. Thus all three types of action–perception loop mentioned in footnote 6 may apply equally well to music.

[14] A recent study of musical "chills" shows striking similarity with cocaine highs, providing "neurochemical evidence that intense emotional responses to music involve ancient reward circuitry" (Salimpoor et al. 2011).

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.

communicational abilities in cooperative interaction. What is peripheral in current linguistic theory (speech and pragmatics) should be central; what is central in much theory (syntax) may be more peripheral. Syntaxes are, I have suggested, language-specific cultural elaborations with partial origins in the interactional system, within bounds set by aspects of general cognition (Christiansen and Chater 2008). Viewed in this light, the relation of language to music shifts. The vocal origins of music may ultimately be tied to the instinctual affective vocal system found in apes, while the joint action and performance aspects may be connected to the interactional base for language. Just as syntaxes are artifacts honed over generations of cultural evolution, so are the great musical traditions.

Acknowledgment

Special thanks are due to Michael Arbib for helpful comments on both a long abstract and the draft paper, as well as to Penelope Brown and Peter Hagoort for helpful early comments. I also owe a diffuse debt to Ani Patel's stimulating 2010 Nijmegen Lectures and to discussions at the Strüngmann Forum meeting, where this paper was discussed.

From "Language, Music, and the Brain," edited by Michael A. Arbib.
2013. Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press. 978-0-262-01810-4.