Songs to Syntax:
Cognition, Combinatorial Computation, and the Origin of Language

Robert C. Berwick, Massachusetts Institute of Technology, USA

ABSTRACT

Language comprises a central component of a complex that is sometimes called "the human capacity." This complex seems to have crystallized fairly recently among a small group in East Africa of whom people are all descendants. Common descent has been important in the evolution of the brain, such that avian and mammalian brains may be largely homologous, particularly in the case of brain regions involved in auditory perception, vocalization and auditory memory. There has been convergent evolution of the capacity for auditory-vocal learning, and possibly for structuring of external vocalizations, such that apes lack the abilities that are shared between songbirds and humans. Language's recent evolutionary origin suggests that the computational machinery underlying syntax arose via the introduction of a single, simple, combinatorial operation. Further, the relation of a simple combinatorial syntax to the sensory-motor and thought systems reveals language to be asymmetric in design: while it precisely matches the representations required for inner mental thought, acting as the "glue" that binds together other internal cognitive and sensory modalities, at the same time it poses computational difficulties for externalization, that is, parsing and speech or signed production. Despite this mismatch, language syntax leads directly to the rich cognitive array that marks us as a symbolic species.

Keywords: *Cognition, Computation, Mathematics, Origin of Language, Syntax*

INTRODUCTION

It seems appropriate to address the full sweep of cognitive informatics and computing with an analysis of the origin and nature of that part of cognition that seems to be uniquely human, namely, language. There can be no doubt that language comprises a central component of what the co-founder of modern evolutionary theory, Alfred Russell Wallace, called "man's intellectual and moral nature," the human cognitive capacities for creative imagination, language and symbolism generally. In short, language makes us smart. In what follows, this article sketches how this remarkable ability might have arisen during the course of evolution and exactly how language boosts our cognitive capacity beyond that of all other animal species. To do this, it first outlines what we know about the evolution of modern humans. This will give us some important clues as to what marks out language as something uniquely human, leading naturally to a brief

DOI: 10.4018/jcini.2011100102

review of what it is that we humans have that other animals don't – what paleo-anthropologist Tattersall (1998) calls "flexibility instead of specificity in our behavior." After all, ants or bees can easily beat us at navigation, and it seems from recent studies that songbirds can do better than us at auditory production and perception. Yet we have surpassed them all in general intelligence.

Remarkably, as we shall see, it turns out that human language seems to arise from just a *single*, small evolutionary innovation, built on two already-available cognitive substrates, present separately in other animals, but brought together for the first time in modern humans. So human language is not just "more of the same," to use Tattersall's terms, but involves something entirely new, "how we integrate" cognitive competences that we share with other animals (Tattersall, 2010). In fact then, contrary to what is sometimes thought, human language is *not* complex – on the contrary, it is far simpler than anyone may have thought, certainly simpler than what one reads about in standard linguistic textbooks. But it *is* novel. On reflection, this is not at all surprising, given the relatively short time scale involved in evolutionary terms – not millions of years but just 100-50 thousand years, according to current accounts. Complicated evolutionary change typically occurs over the time span of many thousands or millions of generations. Given this, we might anticipate that any evolutionary change leading to language would be relatively small, since it seems to have occurred within the time of a few hundred generations, and a hundred generations already takes us back to the founding of the Roman Republic. There simply was not enough time to evolve something as radically new and complex as, say, the wings of birds. As always, evolution by natural selection had to make do largely with what is at hand. Once unleashed, language served as a kind of *lingua franca*, a kind of "cognitive glue" that lets all our other cognitive faculties talk to each other, in a way that is not available to other animals. And applied to other human, digital cognitive domains, it leads to the number system, mathematics, and music. All this appears to be the result of a single evolutionary innovation. So how did this all happen?

Before beginning it is worthwhile to clear away two common misconceptions. First, this view does not entail that 'thought' is co-extensive with 'language.' Obviously they are not. Why? We all know that there can be language without thought, as is demonstrated to us every day by our rote conversations that take place seemingly without any reflection whatsoever – or failing that, the discourse of politicians. Conversely, there can be thought without language, as evidenced by the visual computation inherent in, for example, Feynman diagrams. Nonetheless, it is clear that language plays a large role in our mental lives. Second, I would also like to emphasize that the emergence of language cannot simply have been due to purely an expansion in brain size. While it is true that there has been a *general* increase in brain size throughout the primate lineage, as we shall see, language cannot be the result of brain size alone, since Neandertals were bigger brained than us.

Putting these two matters to one side then, let me turn to a brief review of the paleontological record regarding us and our immediate ancestors as it is currently understood, following Tattersall (2010). A picture of the 'family tree' of our recent homin ancestors, distinct species, with time stretching back to 5-7 million years ago reveals two crucial properties. First, like virtually any other tree of related species, all of the family *Hominidae* – it is very bushy, just as Darwin taught us. In all there may have in fact been at least over 100 distinct species in our immediate family tree. Many of these died out, after making their brief appearance on the evolutionary stage, just as Darwin suggested. Second, at any one time in the past there have typically been several, often many *Hominidae* species that co-existed, often for many hundreds of thousands of years – for instance, *Homo sapiens* (modern humans) and *Homo neandertalesis*. As Tattersall notes, what is unusual is that at their present time we have no other living relatives – there is just a single Homo species left alive in the world after millions of years of co-existence: us (Tattersall, 2010).

Perhaps most strikingly, the several-million year period before the appearance of clearly behaviorally modern Homo sapiens approximately 75 thousand years ago (kYA) is marked in general by a 'disconnect' (Tattersall, 2010) between the appearance of each new hominid species and new technologies as evidenced by differences in, e.g., stone tool making. That is, most often a new species appears on the scene (with a different body morphology and larger brain capacity, etc.), but *without* any concomitant innovative change in external behavior. For example, there is nearly a 1 million year 'gap' between the appearance of the first, Type 1 'scraping tools' about 2.5 million years ago and and the type 2 ovoid Acheulean tools, appearing about 1.5 million years ago; crucially, these post-dated the appearance of the hominid that first made them, *Homo ergaster,* dating from about 1.9 million years ago. "As far as can be told, aside from the invention of the Acheulean in Africa (its spread beyond that continent occurred considerably later) the history of the genus Homo in the period between about two and one million years ago the Old World, but without radical physical or as far as we can tell cognitive innovation. It is not until 600 thousand years ago that we find, again first in Africa, a new kind of hominid with a significantly larger brain…This is *Homo heidelbergensis*" (Tattersall, 2008, p. 105).

Similarly, even though "anatomically recognizable" *Homo sapiens* appears about 200 – 150 thousand years ago in Africa, they come "bearing a technology that was basically indistinguishable from those of its contemporaries and immediate predecessors" (Tattersall, 2008), again a "disconnect" between anatomical and behavioral innovation. Remarkably, at one time in the Levant, at least 3 distinct species of *Homo* apparently lived side-by-side for a hundred thousand years – *Neandertals, heidelbergesis, and sapiens* – all at apparently the same level of tool-making.

All this changed starting about 75 thousand years ago, with the appearance of behaviorally modern humans in Europe, known informally as Cro-Magnons, who displaced Neandertals in Europe. Indeed, the contrast between the Neanderthals and Cro-Magnon provides a clear contrast of the differences between a cognitively competent species operating at the maximum level possible without something like language, the Neandertals, and a second species – us – who had already attained language and symbolic thought. A side-by-side look at the skeletons of *Neandertal* vs. *Homo sapiens* suffices to point out the remarkably different skull, thorax, and pelvis shapes; and in this regard, Neandertals more closely represents our common ancestor, highlighting the extensive, highly derived changes that took place yielding modern *Homo sapiens*. Note, however, that if one considers simply brain capacity that in fact Neandertals had, if anything, a *larger* cranial capacity. They evidently hunted in groups, perhaps even more effectively than Cro-Magnons (consider the metabolic needs during the colder climate at that time); used animal skins and built shelters; and much more. But in contrast to the lack of evidence of any symbolic behavior in Neandertals, with the coming of modern Homo sapiens, there is a virtual explosion of artifacts that all say that this species was just us, including the first sculptures of remarkable aesthetic skill and beauty; sophisticated musical instruments; the first 'written' records on plaques and bones; and the astonishing images on the Chauvet caves in France that can be seen in Werner Herzog's film.

So what happened? What unleashed this astonishing creativity? Clearly, some other animals possess formidable cognitive skills. We saw that our pre-Homo sapiens ancestors made increasingly sophisticated tools, albeit at a glacial pace. It was once thought that tool-making distinguished us from other species, but this has long been proved false. Birds, such as the corvids (ravens, crows, and the like) all can make sophisticated tools and engage in what seems to be quite sophisticated causal reasoning. For example, the Western scrub jay can tie strings around a bit of rock to lower into a hole to catch ants. In Tokyo, Japan, carrion crows have been observed to get at the inside of walnuts by using automobile traffic to crack them: patiently waiting until the pedestrian "walk" signs turn green, then placing the walnuts where automobiles will crack them, stepping out of

the way; then waiting until the "walk" signs turn green again to go and fetch their rewards. Finally, as Aristotle appreciated, songbirds are clearly superb at vocal production, perception, mimcry, and learning. In both birdsong and speech, auditory-vocal learning takes place during a sensitive period early in life, and there is a transitional phase of vocalization called 'babbling' in infants and 'subsong' in young songbirds. More recently, the parallels between speech and song have been extended to the neural and genetic levels. As is now familiar, the juvenile males acquire their songs by listening to con-specific adult male 'tutors', apparently molding an initial 'babbling' template into a progressively more accurate form.

But as complex as birdsong is, it lacks two essential ingredients for human language. First, birdsongs are songs lacking *words*. Birdsong, as varied and as sophisticated as it might be, is not varied to convey distinct meanings, but rather maps directly to some attentive/hormonal state: it is a monoblock signal to mark territory (Me, me, me!) or sexual availability (Ready, ready, ready!). While people quite easily morph "Obama likes Palin" into "Palin likes Obama" to mean something radically different, no bird juggles its song components in any comparable way to result in distinct meanings. No words, no language. Second, as will be described in more detail below, birdsong is not hierarchically compositional in the way that human language is. Though birdsong is sometimes described as being decomposable into repetitive motifs, which are in turn broken down into syllables, this structure does not come close to the way in which human language is formed.

What about the great apes, our closest living relatives? They are good at many cognitive tasks, including cooperative behavior, causal reasoning, and the like. Other primates probably have conceptual structures are found in other primates: probably actor-action-goal schemata, categorization, possibly the singular-plural distinction, and others. These were presumably recruited for language, though the conceptual resources of humans that enter into language use appear to be far richer.

But again there is something missing. Much effort has been spent to 'teach' chimpanzees or gorillas 'language' either by using sign language or, in the case of pygmy chimpanzees, bonobos, by means by plastic tokens denoting actions or objects. These efforts all failed, and failed miserably. No other living non-human ape or dolphin has attained anything close to human language. Rather, as Prof. Laura-Anna Petitto notes, "while apes can string one or two 'words' [or signs] together in ways that seem patterned, they cannot construct patterned sequences of three, four, and beyond…After producing [a] matrix of two words they then– choosing from only the top five or so most frequently used words that they can produce (all primary food or contact words, such as *eat* or *tickle*) – randomly constructing a grocery list. There is no rhyme or reason to the list, only a word salad lacking internal organization" (Pettito, 2005, pp. 85-86).

So what do *we* have that the other animals don't? Here's the surprising answer in a single point: a pencil. While other animals make tools, there is apparently no other animal that makes a *combinatorial* tool. There is no other animal that stitches together separate 'bits' like an 'eraser' and a 'stick of lead' that then can be manipulated as if it were a new, single object, that can be labeled as such – a *pencil.*

What is the analog in the case of language? To answer that, we have to consider the properties of language. In essence, this comes down to Humboldt's famous aphorism that language makes 'infinite use of finite means'. The most elementary property of our shared language capacity is that it enables us to construct and interpret a discrete infinity of hierarchically structured expressions. The expressions are discrete (or 'digital') because there are 5 word sentences and 6 word sentences, but no 5½ word sentences; infinite because there is no longest sentence; and hierarchical because what our language capacity assembles are *structures*, not mere strings of sounds – what are called phrases. Language is therefore based on some generative procedure that takes

elementary word-like elements from a mental store, and applies repeatedly to yield structured expressions, without bound. Operating unfettered, such a system can even arrive at astonishing language combinations such as this one: "Almost inconceivably, the gun into which she was now starting was clutched in the pale white hand of an enormous albino with long white hair."

This is this ability we immediately recognize as the hallmark of human language (even if it's not good language): the ability to produce a discrete infinity of possible meaningful 'signs' integrated with the human conceptual system, the algebraic closure of a recursive operator over our 'dictionary.' No other animal has this combinatorial promiscuity, an open-ended quality quite unlike the frozen 10-20 'word' vocalization repertoire that marks the maximum for any other animal species. Such combinatory promiscuity seemingly permeates all of human mental life, from our lexicon, to mathematics and music, as we shall see.

To account for the emergence of this new computational ability we have to account for its two key components. The first is the storehouse of words – commonly in the range of 30–50,000. The second comprises the computational properties of the language faculty. In turn, the computational properties of the language faculty that constructs internal mental representations may be subdivided further into two components, or *interfaces* with language-external (but organism-internal) systems: the system of thought, on the one hand, and also to the sensorimotor system, thus *externalizing* internal computations and thought. This is one way of reformulating the traditional conception that dates at least back to Aristotle, that language is sound with a meaning.

So what's the 'secret sauce' that lets us, but no other animal, grab any two individual 'words' and paste them together, assembling a new object that can itself be manipulated *as if* it were a single object? Whatever this procedure is, it can take two words, for example, *the* and *apples*, and glue them together into a single new object, here written as *the-apples,* This combinatory operation can in turn paste together a verb with this newly formed object, selecting the verb as 'most prominent' and yielding a verb-like chunk that forever after acts like a verb-like object and so on, yielding *ate the apples, John ate the apples, I know John ate the apples,* etc., the familiar open-ended creativity we associate with human language, an infinite number of (sound, meaning) pairs. If we assume, reasonably, that the human brain is finite, taking the computational theory of mind seriously, then all this must be produced by some *finite* number of rule or operators. But this logically entails that at least one of the operators or rules must apply to its own output, that is, the computational system must be *recursive*.

The simplest assumption is that this generative procedure emerged suddenly, in accordance with the archaeological evidence reviewed above. In that case we would expect the generative procedure to be very simple. Various kinds of generative procedures have been explored in the past 60 years. One approach familiar to linguists and computer scientists is context-free phrase structure grammar, developed in the 1950s. This fit very naturally into one of the several equivalent formulations of the mathematical theory of recursive procedures – Emil Post's rewriting systems – and it captured at least some basic properties of language, such as hierarchical structure and embedding. However, it was quickly recognized by the early 1960s that context-free phrase structure grammar is not only inadequate for natural language but is also quite a complex system with many arbitrary stipulations, and so unlikely to have emerged suddenly.

Over the years, research has found ways to reduce the complexities of such generative systems, and finally to eliminate them entirely in favor of the simplest possible mode of recursive application: an operation that takes two objects already constructed, say X and Y, and forms from them a new object that consists of the two objects unchanged, hence simply the set with X and Y as members, along with a *label* for the new object. Call this operation *cons*, after the familiar Lisp operation *constructor*. Provided with the conceptual atoms of words, the operation *cons* may be iterated without bound, yielding an infinity of hierarchically constructed expressions. If these can

be interpreted by conceptual systems, the operation provides an internal "language of thought." Note that there is no room in this picture for any precursors to true human language. To go from seven-word sentences to the infinity of human language requires the same recursive procedure as to go from zero to infinity. Further, there is no direct evidence for such "protolanguages."

Further, this tells us what the basic structure of human language is, akin to the spiral structure of DNA. But instead of DNA, the basic structure of language is an asymmetrical, hierarchical template. The right way to picture them is like this: as a pair of coat hangers stuck together, that are free to turn, in a mobile-like fashion, around the vertical axis. Thus, in the structure for *ate-the-apples*, the coat hanger unit corresponding to *the apples* is free to rotate around the higher coat hanger *ate*. So left to right order does not matter: indeed, in some languages, like German or Japanese, *the apples–ate* would be the correct order.

There are of course some constraints on *cons*. Not any two arbitrary objects can be combined: we cannot have *the the*, or *ate ate* for instance; this implies that one of the two objects *X, Y* glued together by *cons* has have what we might call and 'edge' feature, like the notch in a jigsaw puzzle piece, that matches up with the other object.

How do we know that these objects are hierarchical 'chunks' rather than, say, just flat strings? This is because one can show that the constraints and operations of human language respect hierarchical structure, not linear order. This is easy to demonstrate via simple examples. In the sentence, *Obama likes him*, 'him' cannot refer to Obama. However, if a chunk of hierarchical structure intervenes between 'him' and 'Obama', as in the example, *Obama thinks Palin likes him*, now 'him' *can* refer to Obama (but of course need not). It does not matter whether Obama is 'to the left' or 'to the right' of 'him'; what matters is the relationship between the two of *hierarchical* structure. But there is more that *cons* implies.

While in the previous cases of *cons*, the two items we combined, *X* and *Y,* were disjoint sets, suppose we have the case where *Y* is a subset of *X* (or vice-versa) where the set object *Y* is the structure corresponding to *the apples*, while the set object *X* corresponds to the structure associated with *John ate the apples.* In this case, *Y* is now a (proper) subset of *X*. In this case, *cons*(*X, Y*) yields the new set structure corresponding to, *the apples John ate the apples.* In effect, we have 'copied' the object of the verb *ate* to a position that is sometimes called the 'focus' of the sentence, to draw attention to it in the discourse. There is an additional principle at work that suppresses the pronunciation of the second copy of *the apples* when it is passed to the speech (or sign language) output machinery to get "flattened" onto a set of instructions to the speech apparatus in a left-to-right-fashion. So in fact what is pronounced, *the apples, John ate*, noting that internally the object of *ate* is in the proper place for interpretation. This is an important point: note that *the apples* must appear in *two* distinct places: one, the position for proper interpretation of *the apples* as the object of the verb (namely, directly after the verb); the second, the position for the proper interpretation of *the apples* as a 'focused' item for intonation (at the front of the sentence). The representation built by the generative apparatus is thus optimal in this regard: it yields exactly the right structure and no more.

More generally, the operation *cons* yields the familiar *displacement* property of language: the fact that we pronounce phrases in one position, but interpret them somewhere else. Thus in the sentence *guess what John is eating* we understand *what* to be the object of *eat* as even though it is pronounced somewhere else. This property has always seemed paradoxical, a kind of "imperfection" of language. It is by no means necessary in order to capture semantic facts. But it is found everywhere in human language. But it falls within *cons* automatically, as we have seen.

These observations generalize over a wide range of sentence types. The resulting representations are in the exact form needed for semantic interpretation: these *interior* mental representations yield a kind of 'logical form'. If you again remember the lambda calculus, or better, programming

languages built on *cons* like Scheme (or Lisp), then the representations are precisely those would we would expect to find if language takes *interior* representations, the interface to semantics, to be primary, so that these representations are 'easy' and transparent to process, involving no extra work for the semantic or interior representational apparatus. This asymmetry is pervasive. As a more complicated example, consider the question: *Which of his pictures did they persuade the museum that every painter likes best?* The answer to this question might be, 'his first one', crucially a *different* picture for each painter (Picasso, Manet, Rembrandt, ….). Now, this kind of answer is possible only if the human system of inference and interpretation constructs a representation that builds *two* instances of "his pictures", one that is logically present as the object of *likes* (and therefore hierarchically underneath), but is not pronounced, and one copy of "his pictures" that is the one you hear.

However, this dual representation, while making *semantics* easy does *not* yield representations that are equally transparent or easy to process for *external* processes like parsing or production. We do not say *guess what John is eating what*, but rather *guess what John is eating*. That is a universal property of displacement. The property follows from principles of computational efficiency. If we suppose serial motor activity to be computationally costly, a matter attested by the sheer quantity of motor cortex devoted to both motor control of the hands and for oro-facial articulatory gestures, then this follows. To externalize the internally generated expression *what did John eat what* it would be necessary to pronounce *what* twice, placing a heavier burden on computation, when we consider expressions of normal complexity and the actual nature of displacement *cons*. With all but one of the occurrences of *what* suppressed, the computational burden is greatly eased. The one occurrence that must be pronounced is the most prominent one, the last one created *cons*: otherwise there will be no indication that the operation has applied to yield the correct interpretation. It appears, then, that the language faculty recruits a general principle of computational efficiency for the process of externalization. The suppression of all but one of the occurrences of the displaced element is computationally efficient, but imposes a significant burden on interpretation, hence on communication. The person hearing the sentence has to discover the position of the gap where the displaced element is to be interpreted. That is a highly non-trivial problem in general, familiar from parsing programs. Sometimes the resulting sentence can be ambiguous, as in, *Who did you walk to the store,* and indeed, there are cases where externalization is impossible, even though the meaning is perfectly clear, as with the example, *Who is it you wonder about as to the reason why they quit school*?

There is, then, a conflict between computational efficiency and interpretive-communicative efficiency. Universally, languages seem to resolve the conflict in favor of computational efficiency and easier *internal* interpretation. That is, the system makes life easy for the *internal* system, rather than making life easy for the *external* system of parsing. These facts at once suggest that language evolved as an instrument of *internal thought*, with externalization by speech or sign a secondary process.

There are independent reasons for the conclusion that externalization is a secondary process. One is that externalization appears to be modality-independent, as has been learned from studies of sign language in recent years (Petitto, 2008). The structural properties of sign and spoken language are remarkably similar. Additionally, acquisition follows the same course in both, and neural localization seems to be similar as well. That tends to reinforce the conclusion that language is optimized for the system of thought, with the mode of externalization a secondary consideration.

The individual first endowed with *cons* would have had many advantages: capacities for complex thought, planning, interpretation, and so on. This capacity would presumably be partially transmitted to offspring, and because of the selective advantages it confers, it might come

to dominate a small breeding group. What it implies is that the emergence of language in this sense could indeed have been a unique event, accounting for its species-specific character. Such 'founder effects' in population bottleneck situations like those often assumed about our ancestral population 50-75 thousand years ago are not uncommon.

Returning to the general theme of cognitive computing, it is important to see what this new combinatorial ability unleashed – that is, how language lit a bonfire under the rest of cognition. How? Recall that there is plenty of evidence for specialized cognitive "modules" in other animals – like bee navigation, or bat echolocation. But characteristically, these narrowly defined modules are unable to "talk" to one another – bats cannot press their echolocation abilities into the service of solving some *other* cognitive task. This is quite different from the "Swiss army knife" character of human cognition – the ability to cobble together novel mental representations of complex events, well beyond the power of any single 'module'.

It is natural to suggest, then, that language acts as a kind of 'cross-module' cognitive glue that links all *other* representations together. In other words, *cons* lets us hook together words into 'chunks' that can then act as single units, but we should recall what stands behind the words, namely, *concepts*. By enabling the construction of extremely complicated, novel conceptual objects and events, language enables the internal construction of representations *of* representations, cross-wiring other mental modules.

Evidence for this cross-modal coupling comes from recent brain imaging evidence regarding the interaction between the brain regions often cited to be active during syntactic processing, e.g., Broca's area (Brodmann's area 44/45), a "phylogenetically younger" part of the cortex (Friederici et al., 2011), and areas involved recognizing events semantically or in relevant motor processing. Two brain tracts as seem to form two streams: one, a top, *dorsal* tract or bundle of fibers, that is involved in coupling sound to its articulation; and two, a separate, bottom, ventral tract that connects syntax to the retrieval of stored representations of objects and actions. Note how this picture corresponds exactly to the two interfaces we mentioned earlier, as well as to similar dorsal/ventral processing streams found in the visual system. Further, it exhibits explicitly how visual representations are cross-wired by language. For example, Aziz-Zadeh et al. (2006) showed that when adults read about action sentences such as "eating a peach", the brain areas that lit up were not only the classical language ones but also the same areas activated when just viewing the action visually Both Aziz-Zadeh et al. (2006) and Pulvermüller and Fadiga (2010) have demonstrated that particular networks are activated for distinct body parts and actions, e.g., arm as opposed to legs, for throwing as opposed to kicking an object. One interpretation of activation patterns such as these is that it is language that 'binds' vision and action together.

A second line of evidence for such 'cross-modal' coupling enabled by language comes from the experimental work of Elizabeth Spelke and her colleagues at Harvard (Hermer-Vasquez, Spelke, & Katsnelson, 1999). In one typical experiment, they probed the influence of language on the ability to integrate distinct perceptual cues. Children and adults were placed in a room with four screens in each corner, in which a desired object (such as a toy, for the children), was placed behind a screen in one corner, while the subjects watched. Then the subjects were blindfolded and disoriented by turning them around, and afterwards had the job of finding the hidden toy. The experimenters wanted to see whether the subjects made use of other perceptual cues about the room: for example, if one wall of the room was painted blue, then that would serve as a landmark to reduce the search for the hidden object, since a subject could simply remember whether the sought-for object was placed behind a blue-wall corner or not, reducing the search from four corners to two. If the perceptual cue is not used, for whatever reason, then performance changes, since any one of the four corners might be selected. The evidence that language is deeply involved with the perceptual integration of such cues is that if we 'overload'

the language system when carrying out the search, by having the subjects perform a simultaneous language task, like reciting a poem, then performance degrades as if the blue wall were not even noticed. In this sense, it seems that language acts to bind

With *cons* then, humans have apparently been liberated to develop ever richer descriptions of the world, involving descriptions of actions, objects, and the like. But there is more than this, not limited to language. Applied repeatedly to the domain of a single element, *cons* acts like the successor function of Peano arithmetic: *cons*(*cons*(*cons*(x))) = 3. This immediately yields the number system of integers, with all its familiar properties. Intriguingly, as soon as children begin to acquire the rudiments of syntax, and *cons*, they apparently an open-ended quantificational ability with large numbers, unlike any other animals. While crows and chimpanzees can seemingly 'count' up to 5-7, beyond that, they deal with large numbers as though they were quantities of 'stuff' weighed on a scale – a more-or-less affair. But children by age 5, or as soon as they have acquired language syntax, seem to easily grasp that if there's a number 100, then there can be 101, as anyone who's ever played the game with a child, "I can find a bigger number than you can" might attest. This kind of ability lies beyond the reach of any other species we know.

Yet a third domain where the 'grouping' operations of *cons* again appears has to do with the domain of both rhythm and music. Consider the beat structure of a line of metrical poetry, for instance, *Tell me not in mournful numbers*. The pattern of strong and weak beats, and in fact that the strongest beat comes first on *tell*, then *not*, and then *mourn*, can be explained very simply by the same *cons* model as before. In this case, *cons* works its way left to right through the string, grouping together pairs of elements, here syllables, as before, just like *ate* and *the apples*, collecting them into a new group that is then supposed to be labeled as such. We then select one of the two units we have collected as the new label of the grouped hierarchical structure, obtaining a second level of representation. After one pass through the initial syllables, we make a second pass through this next level, as before, as shown, successively collecting groups of two, until we can do no more. Then the strongest 'beat' is just the stack with the greatest depth, and so on.

Note that there is a crucial difference here between full-fledged language and beat structure. The 'lexical items' here correspond to syllables, and are denoted simply as asterisks, because there are no elements in beat structure with features like 'verb' or 'noun.' There are just the 'marks' of the beats, corresponding to syllables. Thus, when two marks are collected together to form a new, higher level unit, there are no features to *label* the new unit, as we labeled *ate the apples* as a 'verb phrase'. That is, beat structure is what one gets if one applies the same *cons* operation as in the rest of language, but to a system *without* words. One and the same innovation leads to both language syntax and metrical structure. While there is no space to demonstrate it here, one can show that just a few different ways of 'passing through' the initial string of syllables – from right to left as opposed to right to left, perhaps alternating between levels, give rise to all the possible metrical patterns shown in all human languages (Fabb & Halle, 2008). We might even think of this 'beat structure' as the first glimmerings of language syntax – the platform for *cons* that existed before words were wired into language. If so, then a primitive ability like *cons* might actually be apparent in another species that exhibits metrical patterning to their vocal output – in particular, songbirds. And in fact, there is some suggestive evidence from genomic data, the famous FOXP2 gene, that this is the case: when this regulatory gene is disrupted in either songbirds or humans, then the ability to sequence either songs or speech into their proper linear arrangements, following the 'beats' of motor sequencing, is likewise disrupted (Haesler et al., 2007). We shall have to leave aside a detailed commentary on this point for now. What it suggests, though, is that the ability to carry out *cons* might have been available either many hundreds of millions of years ago, but that true language did not appear because there were no words to wire it to, or else that *cons* arose in songbirds by means of convergent evolution.

What we do know is that there is yet one more domain of human activity where *cons* surfaces, related to metrical structure, and it is this one: music. First, the beat structure of music works exactly like what we showed for metrical poetry: the same grouping-and-projection (without words). Second, extending this to a 'lexicon' that consists of melodic notes, leads to an analysis of 'tonal musical syntax' that looks a lot like language (with obvious differences again because melodic elements are not words). As an example, Katz and Pesetksy argue that language and tonal music are formally identical, differing only in the fundamental building blocks that they use. In the case of language, as we have seen, these 'atoms' are words (Katz & Pesetsky, 2011). In the case of music, Katz and Pesetsky maintain that these atoms are pitch-classes and chord qualities. Just as word-atoms may be combined into phrases by *cons*, Katz and Pesetsky demonstrate that the analog in music is so-called 'prolongation reduction' – "the hierarchical patterns of tension and relaxation in tonal harmony," which are also put together by the operation *cons*. So perhaps this was Mozart's deep secret: for whatever reason, the generative faculty for language that makes speaking for us so effortless, was somehow cross-wired in Mozart's brain at a very early age so that literally, for him, making music was just as natural and easy as speaking is for us.

Let us just summarize briefly what seems to be the current best guess about the origin of language and human cognition. In some as yet unknown way, our ancestors developed human concepts, as opposed to what chimps, birds, and bees possess. At some time in the very recent past, perhaps about 75,000 years ago, an individual in a small group of hominids in East Africa underwent what was likely a small mutation that provided the operation *cons* – an operation that takes human concepts as computational atoms, and yields structured expressions that provide a rich language of thought. The innovation had obvious advantages, and took over the small group. At some later stage, this internal language of thought was connected to the sensory motor system, with its obvious consequences for communication, both for good and ill. In this way, human language provided the platform for its expansion into all areas of thought, yielding a good part of our "moral and intellectual nature," in Wallace's phrase (Wallace, 1871), extending far beyond language, to mathematics and music, indeed, much of what is distinctive about the human condition.

REFERENCES

Aziz-Zadeh, L., Wilson, S., Rizzolati, G., & Iacobani, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Current Biology*, *16*, 1818–1823. doi:10.1016/j.cub.2006.07.060

Fabb, N., & Halle, M. (2008). *Meter in poetry*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511755040

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

Haesler, S., Rochefort, C., Georgi, B., Licznerski, P., Osten, P., & Scharff, C. (2007). Incomplete and inaccurate vocal imitation after knockdown of FoxP2 in songbird basal ganglia nucleus Area X. *PLoS Biology*, *5*, e321. doi:10.1371/journal.pbio.0050321

Hermer-Vasquez, L., Spelke, E. S., & Katsnelson, A. S. (1999). Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology*, *39*, 3–36. doi:10.1006/cogp.1998.0713

Katz, J., & Pesetsky, D. (2011). *The identity thesis for language and music*. Cambridge, MA: MIT. Retrieved from http://ling.auf.net/lingBuzz/000959

Petitto, L. (2005). How the brain begets language . In McGilvray, J. (Ed.), *The Cambridge Companion to Chomsky* (pp. 84–101). Cambridge, UK: Cambridge University Press. doi:10.1017/CCOL0521780136.005

Pulvermüller, F., & Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nature Reviews. Neuroscience*, *11*, 351–360. doi:10.1038/nrn2811

Tattersall, I. (1998). *The origin of the human capacity (68th James Arthur Lecture on the Evolution of the Human Brain)*. New York, NY: American Museum of Natural History.

Tattersall, I. (2008). An evolutionary framework for the acquisition of symbolic cognition by *Homo sapiens*. *Comparative Cognition Behavior Reviews*, *3*, 99–114.

Tattersall, I. (2010). Human evolution and cognition. *Theory in Biosciences*, *129*, 193–201. doi:10.1007/s12064-010-0093-9

Wallace, A. (1871). *Contributions to the theory of natural selection*. New York, NY: Macmillan.

Robert C. Berwick is professor of Computational Linguistics in the Department of Electrical Engineering and Computer Science and the Department of Brain and Cognitive Sciences at the Massachusetts Institute of Technology. Professor Berwick received his A.B. degree from Harvard University in Applied Mathematics and his S.M. and Ph.D. degrees from the Massachusetts Institute of Technology in Computer Science in Artificial Intelligence. Since then he has been a member of the MIT faculty, and is currently co-Director of the MIT Center for Biological and Computational Learning. He is the recipient of a Guggenheim Award, and the author of 7 books and many articles in the area of natural language processing, complexity theory, language acquisition, and the biology and evolution of language. His latest book, to be published by Oxford University Press, is Rich Grammars from Poor Inputs.