*J. theor. Biol.* (2001) **211,** 277–295
doi:10.1006/jtbi.2001.2346, available online at http://www.idealibrary.com on IDEAL®

# Universal Scaling Laws for Hierarchical Complexity in Languages, Organisms, Behaviors and other Combinatorial Systems

MARK A. CHANGIZI*

*Department of Psychological and Brain Sciences, Duke University, Box* 90086, *Durham, NC* 27708, *U.S.A.*

(*Received on* 4 *December* 2000, *Accepted in revised form on* 3 *May* 2001)

There are many complex systems in nature where components, or "words", are combined together to make expressions, or "sentences". Such combinatorial systems include: (1) human language, where sentences are composed of words; (2) bird vocalization, where songs are built from syllables; (3) organisms, where organism-expressions (e.g. the tonsil) are made out of cells; (4) behavioral repertoire, where mammalian behavior consists of a temporal arrangement of muscle contractions; (5) universities, where student academic degrees are comprised of departmental concentrations; and (6) electronic devices, where the device's actions are implemented via strings of button-presses. My central aim here is to discover how combinatorial systems accommodate greater numbers of expressions; that is, what changes do combinatorial systems undergo when they "say more things?" Are there general laws characterizing the properties of combinatorial systems as the number of expressions increases? If so, what are they? My main result is that, in all the kinds of combinatorial system mentioned above, there appear to be general laws describing how combinatorial systems change as they become more expressive. In particular, in each of these cases, increase in expression complexity (i.e. number of expressions the combinatorial system allows) is achieved, at least in part, by increasing the number of component types. Each kind of system follows one of two kinds of scaling law. In the first kind of scaling law, expression complexity increase is carried out *exclusively* by increasing the number of component types; the number of components per expression (i.e. the expression length) remains invariant. This applies to human language over history, bird vocalization, organisms in phylogeny and ontogeny, and universities. In the second kind of scaling law, expression complexity is accomplished by increasing in a law-like manner *both* the number of component types *and* the expression length. This applies to two cases of the ontogeny of language—the development of words and sentences, and the development of phonemes and morphemes—and to mammalian behavior. By treating these diverse systems as combinatorial systems we, in addition to elucidating general principles underlying such systems, gain insight into each kind of system mentioned.



## Introduction

There are many *combinatorial systems* in nature in which *components* are combined to instantiate *expressions* of some kind. For example, bird vocalization appears to be a combinatorial system because syllables (the components) may be put together in many ways to make many different songs (the expressions). My interest here concerns combinatorial systems generally,

*E-mail: changizi@changizi.com; www.changizi.com

0022–5193/01/150277 + 19 $35.00/0

and my main question is: *are there universal laws governing how combinatorial systems achieve greater expression complexity* (*i.e. a greater number of expressions*)? For example, if one species of bird has twice the number of songs as another species, what can we say, if anything, about their respective number of syllable types and number of syllables in an expression? More generally, if a combinatorial system has $E$ expressions, $C$ component types, and a characteristic expression length of $L$, then, among combinatorial systems of the same kind, how do $C$ and $L$ change as $E$ is scaled up?

There are at least five possible, plausible, general hypotheses, falling into two categories. Since the data I have been able to acquire inform us on how $C$ changes as a function of $E$, in presenting each hypothesis I focus on the expected relationship between $C$ and $E$.

*Category* 1: $E \sim C^L$. In other words, for a given kind of combinatorial system, there is a constant proportionality factor relating $E$ to $C^L$. For example, over all song birds, the number of songs a bird knows may be proportional to the number of syllables it knows to the power of the characteristic number of syllables in a song. The proportionality factor is determined by the combinatorial system's "grammar", or the rules determining which arrangements of components are allowed. While it is, in principle, possible that the proportionality factor remains constant even though the grammar undergoes change, it seems much more plausible that the proportionality factor's invariance is due to the grammar's invariance; in other words, when the proportionality factor is a constant it strongly suggests that the grammar is scale invariant.

(a) $E \sim C^L$, *only $C$ increases, and it does so in a lawful way*: under this hypothesis, $E \sim C^L$ and $L$ is invariant. Thus, $C \sim E^{1/L}$, with $L$ constant. That is, $C$ and $E$ are related by a power law with exponent equal to the inverse of the expression length. Since $L \geqslant 1$, the exponent is in the interval (0, 1]. (This will be slightly amended in the next section.) Also, if $L = 1$, the system is a combinatorial system in name only. (For example, birds with more songs would have songs of approximately the same average length, but have a greater number of syllable types with which to build songs.)

(b) $E \sim C^L$, *only $L$ increases, and it does so in a lawful way*: under this hypothesis, $E \sim C^L$, and since $C$ does not increase ($C \sim E^0$), with a little manipulation it follows that $L \sim \log E$. (For example, birds with more songs would build them with the same number of syllable types, but string them together into longer songs.)

(c) $E \sim C^L$, *both $C$ and $L$ increase, and they do so in a lawful way*: for this hypothesis, $E \sim C^L$ and neither $C$ nor $L$ is invariant. For this to be the case, $L$ must scale more slowly than logarithmically (because if $L \sim \log E$, then $C \sim E^0$). One natural possibility is that $L \sim (\log E)/\log(\log E)$, in which case a little algebra shows us that $C \sim \log E$. (For example, birds with more songs would have both longer songs and more syllable types with which to build songs.)

(d) $E \sim C^L$, *one or both of $C$ and $L$ increase, but they do* not *do so in a lawful way*: this hypothesis states that $E \sim C^L$, but when $E$ increases, sometimes $C$ increases, sometimes $L$ increases, and sometimes both increase, with no overall regularity governing which of these occurs. (For example, some birds with more songs would have a greater number of syllable types in their repertoire but have no increase in song length, some birds with more songs would have longer songs but no increase in the number of syllable types, and some birds with more songs would have both more syllable types and longer songs. No general law governs the manner in which birds increase their expressivity, although $E \sim C^L$ still holds.)

*Category* 2: It is *not* the case that $E \sim C^L$. In other words, for a given kind of combinatorial system, there is *no* constant proportionality factor relating $E$ to $C^L$. Since the proportionality factor is determined by the combinatorial system's grammar, a non-constant factor entails that there is no invariant grammar in these systems: the grammar evolves with increasing expression complexity $E$. There are many hypotheses falling

within this category, but only the following one seems to me prima facie plausible.

(e) *Not $E \sim C^L$, and neither C nor L increases*: for this hypothesis it is not the case that $E \sim C^L$, and greater numbers of expressions are accommodated not via an increase in either $C$ or $L$, but, instead, by increasing the proportionality factor, i.e. by allowing a greater and greater percentage of all possible combinations to count as grammatical expressions. It can be difficult to distinguish between this hypothesis and hypothesis (a) when $L$ is large, because when $L$ is large hypothesis (a) leads to an effectively invariant $C$. (For example, birds with more songs might have the same number of syllable types and the same length songs, but just utilize more and more ways of arranging the syllables into songs.)

We cannot dismiss, *a priori*, any of these five possibilities. In order to test which of these possible hypotheses seems to occur in actual combinatorial systems, I undertook a study of the 12 kinds of combinatorial system (falling under six categories) displayed in Table 1 (where each row is a kind). For each kind of combinatorial system I acquired from many combinatorial systems of that kind the number of component types and the number of expressions.

My main results are as follows. (1) In each kind of combinatorial system, the number of component types increases as a function of expressive complexity, and does so in a lawful way. Thus, hypotheses (b), (d) and (e) are not indicative of these kinds of combinatorial system. The remaining two hypotheses—(a) and (c)—satisfy the proportionality $E \sim C^L$, and thus the kinds of combinatorial system studied appear to have scale-invariant grammars. (2) Some of the kinds of combinatorial system—bird vocalization, human language over history, organisms both in phylogeny and ontogeny, and universities—appear to follow hypothesis (a). That is, in these latter combinatorial systems, increased expressivity is accommodated *purely* by increasing the number of component types; the length of an expression does not appear to change. This

TABLE 1

*Summary of the kinds of combinatorial system, the components, expressions, which hypothesis from the introduction the data confirm* [(*a*)–(*e*)], *the correlation, the combinatorial degree, and the corresponding figure in this paper. When it is unknown whether* (*a*) *or* (*c*) *applies, correlation and combinatorial degree are under the assumption of hypothesis* (*a*). *Correlation is highly significant* ($p < 0.01$) *in each case. When hypothesis* (*c*) *applies, the approximate range is shown over which combinatorial degree values increase*

| Kind of combinatorial system | Component | Expression | Relationship between $C$ and $E$ (and hypothesis from introduction) | $R^2$ | Combinatorial degree | Figure in text |
|---|---|---|---|---|---|---|
| Bird vocalization | Syllable | Song | Power law (a) | 0.702 | 1.23 | 1 |
| Human language | | | | | | |
| —Historical | Word | Sentence | Power law (a) | 0.795 | 5.02 | 2 |
| —Ontogeny 1 | Word | Sentence | Logarithmic (c) | 0.984 | 1–2.5 | 3 |
| —Ontogeny 2 | Phoneme | Morpheme | Logarithmic (c) | 0.959 | 2–4 | 4 |
| Behavior | Muscle | Behavior | Logarithmic (c)? | 0.772 | 3–9 | 5 |
| Organism | | | | | | |
| —Phylogeny | Cell | Expression | Power law (a) | 0.438 | 12.42 | 6 |
| —Ontogeny | Cell | Expression | Power law (a) | 0.988 | 1.02 | 7 |
| University | Concentration | Degree | Power law (a) | 0.687 | 1.65 | 8 |
| Electronic device | | | | | | |
| —CD player | Button-press | Action | Power law or logarithmic (a or c) | 0.489 | 2.07 | 9 |
| —TV | Button-press | Action | Power law or logarithmic (a or c) | 0.842 | 1.58 | 9 |
| —VCR | Button-press | Action | Power law or logarithmic (a or c) | 0.508 | 3.95 | 9 |
| —Calculator | Button-press | Action | Power law or logarithmic (a or c) | 0.875 | 8.77 | 9 |

conclusion is made via demonstrating that the number of component types scales against the number of expressions as a power law with exponent in the interval (0,1]. Alternatively, some of the kinds of combinatorial systems—both cases of the ontogeny of language, and mammalian behavior—appear to be described by hypothesis (c). That is, increased expressive complexity is accommodated by increasing *both* the number of component types *and* the expression length. This conclusion is made by showing that the number of component types increases logarithmically as a function of the number of expressions. There were four kinds of combinatorial system—the four kinds of electronic device—for which I could not ascertain which of hypotheses (a) or (c) holds.

## Results

In this section, I report the results for each of the kinds of combinatorial system. Although my principal task is the examination of combinatorial systems in general, it is also my hope to communicate that analysing the scaling behavior of any particular kind of system can give us insights into systems of that kind. With this in mind, for some of the kinds of combinatorial system I devote a little time to possible explanations for the particular scaling behavior.

Before proceeding, some comment is needed concerning the interpretation of the constant exponent $b$ in cases where $C \sim E^b$. When $C \sim E^b$ with $b \in (0, 1]$ it entails that $E \sim C^d$ where $d = 1/b \in [1, \infty)$. In hypothesis (a) above I stated that $d$ will equal the expression length. This is not exactly true. Instead, $d$ is a measure of the *combinatorial degree* of the kind of combinatorial system, where a combinatorial degree of $d$ means that it is as if, on average, $d$ components are required to implement an expression and all combinations (up to a constant proportion) of the components are expressions. I say "as if" because although it is possible that a system with combinatorial degree three has, on average, three components for each expression, it is also possible that the system has, on average, 100 components for each expression, but so few of the possible combinations are utilized that it only has the combinatorial degree of a three-component-per-expression entity; i.e. it has an "effective" expression length of only three. Therefore, when we measure the exponent $b$ in a log–log plot of $C$ vs. $E$, we can only conclude that $1/b$ is the combinatorial degree, and that this may or may not be equal to the expression length $L$. I will suppose that if the combinatorial degree is invariant, then the expression length is also invariant; that is, I will conclude from the fact that the data fits a power law that the combinatorial degree is invariant, and thus that probably is so in the expression length. In principle, this need not be the case: the expression length could, say, increase as $\log E$, yet the number of "effective components" might stay invariant. This would seem, however, quite improbable.

### BIRD VOCALIZATION

Syllables (the components) for bird vocalizations are combined to make songs (the expressions). The number of syllable types and the number of songs were acquired for 28 birds with expression complexities ranging over two and a half orders of magnitude: ten thrush species (Ince and Slater, 1985), seven wren species (Kroodsma, 1977), five male Bewick Wrens (Kroodsma, 1977), one magpie (Brown *et al.*, 1988), tufted and bridled titmouses (Hailman, 1989), one canary (Mundinger, 1999; Devoogd *et al.*, 1993), and alder and willow flycatchers (Kroodsma, 1984).

The number of syllable types increases with increasing expression complexity, and they appear to be related by a power law (Fig. 1). Under a power law hypothesis there is a high and significant ($p < 0.01$) correlation between them [Fig. 1(a)], and the scaling relationship is $C \sim E^{0.813}$ (where now $C$ is the number of syllable types and $E$ the number of songs). Under a logarithmic hypothesis the plot [Fig. 1(b)] is not at all linear, curving upward in a way characteristic of power laws plotted without the $y$-axis logged. Thus, bird vocalization appears to follow hypothesis (a) from the introduction: greater expressivity is handled by exclusively increasing the number of syllable types; the expression length appears to remain invariant (i.e. expression length does not vary as a function of the number of expressions, although, of course, it may as well be variable).

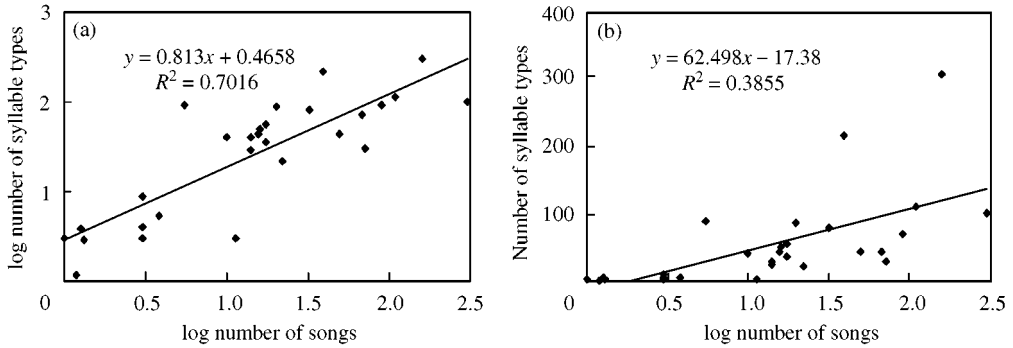


FIG. 1. (a) $\text{Log}_{10}$ of the number of syllable types vs. $\log_{10}$ of the number of songs in 28 birds. (b) Number of syllable types vs. $\log_{10}$ of the number of songs among the same birds.

The combinatorial degree for bird song measured here is $1/0.813 = 1.23$. How does this compare to the actual number of syllables appearing in bird song? We may estimate this from Read & Weary (1992), who compiled the number of syllables per bird song for 109 birds. Since the values range over a couple orders of magnitude, a log-transformed average (i.e. the inverse log of the mean of the logs) is a more appropriate measure of the central tendency, and is 3.72 syllables per word; the median is 3. I do not have any insight into what this means for the grammar of bird song. However, we shall see that the combinatorial degree for bird vocalization of 1.23 is much less than the combinatorial degree of 5.015 for English, which is not unexpected if the combinatorial degree is determined by neurobiological constraints. In fact, the 95% confidence interval for the slope of 0.813 is [0.599, 1.027], and thus the slope is not significantly different from 1, where the combinatorial degree would be 1. A combinatorial degree of 1 implies the system is not, in fact, combinatorial at all, and thus not language-like. Therefore, while at first glance bird vocalization appears to be language-like since songs are built out of syllables, an examination of the way bird vocalization scales up suggests that it may not be language-like after all.

## ENGLISH LANGUAGE, HISTORICALLY

Words (the components) in a language are combined to make sentences (the expressions). Each entry in the dictionary is a different word type. To study how the number of word types in a language increases as the number of sentences increases, I studied the rates of growth for each from the years 1200 to 1990 (Fig. 2); data are currently available and computer accessible only for English, and so only the English language was studied.

The number of new word types per decade was determined by searching for years within the decade in the etymologies of the Oxford English Dictionary (OED), Second Edition. The number of new word types per year $dC/dt$, where $C$ is now the number of word types, grows exponentially in time, following equation $dC/dt \sim 10^{0.001725t}$ ($\sim e^{0.003972t}$). Since the growth is exponential [Fig. 2(a)], the absolute number of word types is proportional to the number of new word types per year, i.e. $C \sim dC/dt$. Thus, $C \sim e^{0.003972t}$.

I used the number of books written in English in any given decade as a measure of the number of new English sentences in that decade. (Note that most written sentences are probably novel). The number of new books per decade was obtained by searching for publication dates within the decade for literature listed in WorldCat, an on-line catalog of more than 40 million records found in thousands of OCLC (Online Computer Library Center) member libraries around the world. The number of new books per year $dE/dt$, where $E$ is the absolute number of books, grows exponentially in time [Fig. 2(a)], and follows equation $dE/dt \sim 10^{0.008653t}$ ($\sim e^{0.01992t}$). As was the case for word types, $E \sim dE/dt$, and so $E \sim e^{0.01992t}$.

Since the equations for $C$ and $E$ are each exponential functions of $t$, they are related by

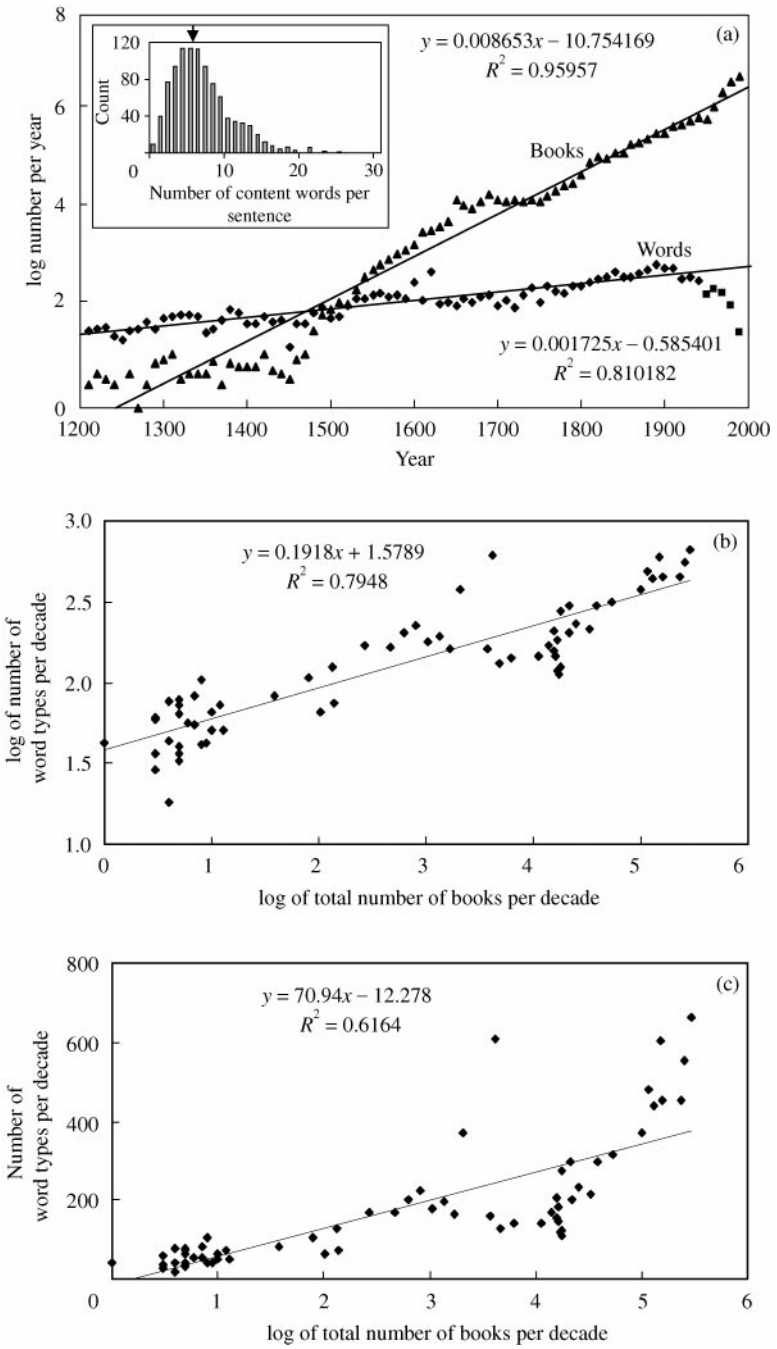


FIG. 2. (a) Growth rates in the decades from the years 1200 to 1990 for the number of new English word types and the number of new English books. Regression equations and correlation coefficients are shown for each (79 data points each). Unsure etymological dates tend to cluster at century and half-century marks and therefore century and half-century marks tend to be overcounted; accordingly, they were not included in the counts. The OED is conservative and undercounts recently coined word types; consequently, the exponential decay region (the last five square data points) was not included when computing linear regression. I do not have any way to similarly measure the number of word type extinctions per year, and so I have not incorporated this; my working assumption is that the extinction rate is small compared to the growth rate, but it should be recognized that the estimated combinatorial degree is therefore an underestimate. [(a), inset] Distribution of numbers of content words per sentence in English. A string of words was deemed a sentence if it represented a complete thought or proposition. So, for example, semicolons were treated as sentence delimiters, multiple sentences combined into one long sentence by, "and" were treated as multiple sentences, and extended asides within dashes or parentheses were not treated as part of the sentence. Arrow indicates the log-transformed mean. (b) $\text{Log}_{10}$ of the number of new words per decade directly against the $\log_{10}$ of the number of new books per decade results in a moderately linear plot (note that the $x$-axis here ranges over more than five orders of magnitude), whereas (c) plotting the number of new words per year directly against the $\log_{10}$ of the number of new books per year results in a plot that is nonlinear, bending upward and scattering in a fashion distinctive of power laws plotted with only the $x$-axis logged; this further confirms the power-law relationship between the number of word types and the number of sentences.

a power law. If hypothesis (a) from the introduction holds here, then it must also be the case that the exponent is in the interval (0, 1]. From the exponential equations for $C$ and $E$ we can conclude that $C \sim E^{(0.003972/0.01992)}$, or $C \sim E^{0.1994}$. Alternatively, we may plot $\log \mathrm{d}C/\mathrm{d}t$ directly against $\log \mathrm{d}E/\mathrm{d}t$ [Fig. 2(b)], and the best-fit slope by linear regression is 0.1918; since $\mathrm{d}C/\mathrm{d}t \sim C$ and $\mathrm{d}E/\mathrm{d}t \sim E$, this slope may be used as another estimate of the exponent for $C$ as a function of $E$. A plot of (unlogged) $\mathrm{d}C/\mathrm{d}t$ directly against $\log \mathrm{d}E/\mathrm{d}t$ appears very nonlinear, and upwardly curved [Fig. 2(c)]. These analyses demonstrate that, for English, component-type complexity increases as expression complexity increases, and they are, indeed, related as a power law with exponent in (0, 1]. Therefore, increasing expressivity in English appears to be achieved exclusively by increasing the number of word types.

Although the total number of words in an English sentence tends to be in the range of 10–30 (Scudder, 1923; Hunt, 1965), the combinatorial degree is a much lower $1/0.1994 = 5.015$. In an effort to understand this particular combinatorial degree value, consider that there are two kinds of words in English: *content* and *function*. The set of content words, which refer to entities, events, states, relations and properties in the world, is large (hundreds of thousands) and experiences significant growth (Clark & Wasow, 1998). In contradiction to content words, the set of function words, which includes prepositions, conjunctions, articles, auxiliary verbs and pronouns, is small (around 500) and relatively stable through time (Clark & Wasow, 1998). The scale-invariant combinatorial degree of English suggests that the average number of words per sentence is invariant. Under the simplifying assumption that there are, on average, $n$ "slots" in a sentence for content words and $m$ "slots" for function words, the total number of sentences $E \sim N^n M^m$, where $N$ is the total number of content words in English and $M$ the total number of function words. Since $n$, $m$ and $M$ are invariant (as discussed above), $E \sim N^n$, and since the total number of word types $C \approx N$ (because $C = M + N$ and $M \ll N$), $E \sim C^n$. That is, this reasoning suggests that the combinatorial degree of around five is due to there being, on average, around five content words per sentence, and they may be combined in any order (up to a constant proportion). To test this, I sampled 984 sentences from 155 authors from texts in philosophy, fiction, science, politics and history; I chose the second sentence on each odd numbered page. A word was deemed a function word if it was among a list of 437 such words I generated. The distribution is log-normal [Fig. 2(a), inset], and the mean of the logs is 0.7325 ($\pm$ 0.2987); the log-transformed mean is thus 5.401, and one standard deviation around this corresponds to the interval [2.715, 10.745]. Perhaps there are this many content words per sentence because of neurobiological constraints on understanding a sentence (Miller, 1956).

## ONTOGENY OF LANGUAGE IN CHILDREN

Words (the components) in a child's developing language are combined to make sentences (the expressions). I acquired data for the number of word types and the number of distinct sentences produced by a child named Damon for 41 weeks from 12 to 22 months of age (Clark, 1993). The numbers of word types and sentences do not appear to be related by a power law [Fig. 3(a)] as can be seen by how the plot flattens out. A logarithmic plot appears comparatively linear [Fig. 3(b)], providing support for hypothesis (c) from the introduction. [Actually, the data are perhaps best understood as two distinct power law regimes, with slope decreasing (and thus combinatorial degree increasing) in the second regime.] Children, then, appear to increase their number of sentences by increasing both the number of word types and the combinatorial degree. This conclusion is, of course, hardly surprising: children learn words, and their ability to string together words into sentences increases with age (Pascual-Leone, 1970; Case *et al.*, 1982; Siegel & Ryan, 1989; Adams & Gathercole, 2000; Robinson & Mervis, 1998; Corrigan, 1983). We may, however, use the data to determine how the combinatorial degree actually changes over this period. By determining the inverse of the instantaneous best-fit slope of the log–log plot, we can see that the combinatorial degree increases for Damon from around 1 to around 2.5 over this period, which is consistent with the increase in the mean length utterances for children in this age range (Robinson & Mervis, 1998). No

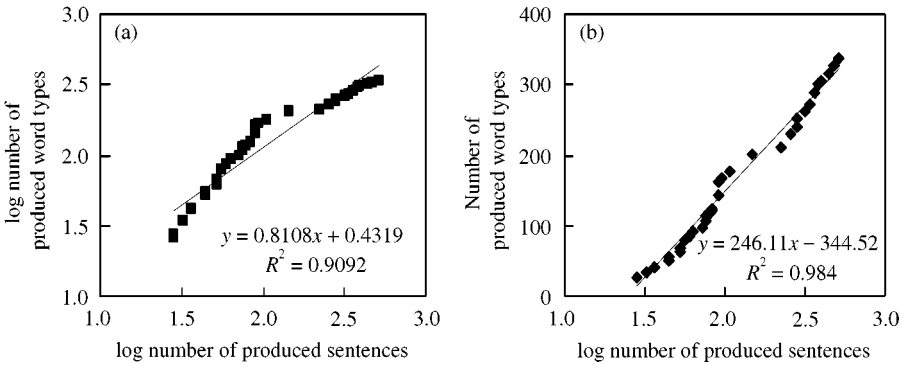


FIG. 3. (a) $\text{Log}_{10}$ of number of word types vs. $\log_{10}$ of number of sentences, as produced by one child named Damon from 12 to 22 months (Clark, 1993). Plot confined to multiword utterance ages, which began at about 14 months. (b) Number of word types vs. $\log_{10}$ of number of sentences. Since the average length utterance increases through time for children (probably due to increases in working memory), the combinatorial degree is not invariant during the ontogeny of language, and we expect the number of word types, $C$, to increase more logarithmically with $E$, which is what we see here.

utterance length information is given for Damon, so these combinatorial degree values cannot be directly compared to his utterance lengths.

Along the same lines as the ontogeny of the word–sentence combinatorial system, we may look at the ontogeny of the phoneme–morpheme combinatorial system. (A *morpheme* is the smallest meaningful linguistic unit.) Again, a child's ability to string together sequences of phonemes into morphemes increases through time, and we therefore expect the number of phoneme types to scale not as a power law against the number of morphemes, but, instead, to scale logarithmically with the number of morphemes. Figure 4(a) and (b) show the number of phoneme types plotted against the number of morphemes, as produced by a child named Jean from 11 to 30 months of age (Velten, 1943). As expected, the plot is (a little) more linear in Fig. 4(b) under the logarithmic assumption. Examination of the inverse of the instantaneous slopes from the log–log plot shows that the combinatorial degree does, indeed, increase as the number of morphemes increases [Fig. 4(c)], beginning at around two phonemes per morpheme (such as in "ma"), and rising to around four phonemes per morpheme at the end (such as in "dada"). This correlates very well with the actual increase in Jean's number of phonemes per morpheme, as measured by the maximum at each age [Figure 4(c) and (d)].

These results suggest that, when the child's number of component types is $C$ and combinatorial degree is $d$, the child produces $E \sim kC^d$ expressions, where $k$ is constant (i.e. children are "grammar-invariant"). That is, these results imply that as $C$ and $d$ increase, the child increases $E$ as fast as possible; intuitively, the child says, up to a constant proportion, everything he possibly can given the component types and combinatorial degree he has at his disposal. This is true for the phoneme–morpheme systems and for the word–sentence systems.

## BEHAVIORAL REPERTOIRE

Muscle contractions (the components) in vertebrates are combined to implement behaviors (the expressions); behaviors are "musical scores" with muscle-contractions as "notes" (Gallistel, 1980). In what fashion do vertebrates achieve greater behavioral complexity, i.e. a greater number of behaviors in their repertoire? To study this I would like to measure the number of muscle contraction types, $C$, as a function of the number of behaviors, $E$. The number of types of muscle contraction I measured as simply the number of muscle types, which I obtained for 12 land mammals and birds (see Fig. 5, legend).

How are we to measure an animal's numerous behaviors? As a proxy for this I used an animal's *encephalization quotient* (Jerison, 1973), which is a measure of how much "extra" brain there is above and beyond that expected by virtue of the animal's mass. Since brain volume appears to scale as $M^{3/4}$ (Allman, 1999; Changizi, 2001), the

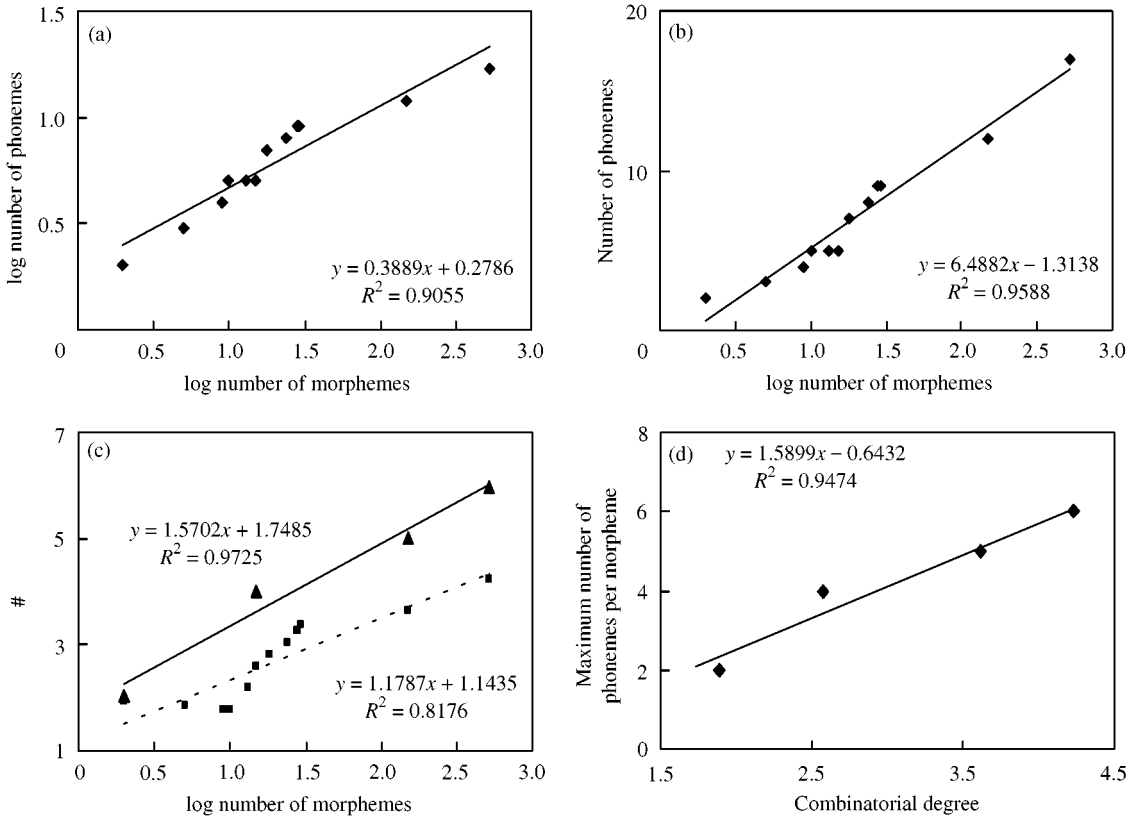


FIG. 4. (a) $\text{Log}_{10}$ of number of phoneme types vs. $\log_{10}$ of number of morphemes, as produced by one child named Jean from 11 to 30 months (Velten, 1943). (b) Number of phoneme types vs. $\log_{10}$ of number of morphemes. Since the average and maximum number of phonemes per morpheme increases through time for children, the combinatorial degree is not invariant during ontogeny, and we expect the number of phoneme types to increase logarithmically with $E$, which is what we see, to a small extent, here ($R^2 = 0.9055$ for log–log plot, and $R^2 = 0.9588$ for semilog plot). (c) Maximum number of phonemes per morpheme (a measure of expression length $L$) and combinatorial degree vs. $\log_{10}$ of number of morphemes. Combinatorial degree is measured from inverse of instantaneous slope (measured via linear regression for a moving window 12 data points wide) from the log–log plot. Max # phonemes per morpheme, ▲; combinatorial degree, ■. One can see that both the maximum number of phonemes per morpheme and the combinatorial degree increase together, and are well correlated with one another, as shown in (d).

encephalization quotient is $(V_{\text{brain}})/(M^{3/4})$. Encephalization quotient is probably highly correlated with behavioral complexity: ordering animals on the basis of their encephalization quotients leads to an ordering of animals that seems, intuitively, to reasonably, correctly order the animals on the basis of behavioral complexity. For example, if one looks at Fig. 5 and concentrates just on the $x$-axis values for each animal, one will see that human has greater encephalization quotient than macaque, which has greater quotient than dog, next in the list is cat, followed by elephant, rat and so on. Although encephalization quotient may well be positively correlated with behavioral repertoire size, I have no reason to expect them to be proportional to one another. Nevertheless, I *will* assume that encephalization quotient is related to behavioral complexity as a power law with unknown exponent. If the muscle-behavior combinatorial system follows a power law, then, with the assumption in hand, we expect a plot of a number of muscle types to scale against encephalization quotient as a power law as well. (That is, if $C \sim E^b$ and $E \sim Q^a$ (where $Q$ is encephalization quotient), then $C \sim Q^{ab}$.)

The encephalization quotient was computed for the 13 animals for which I acquired muscle type counts (see Fig. 5, legend). The number of muscle types was plotted against the encephalization quotient for these animals on a log–log plot [Fig. 5(a)], the $x$-axis which ranges over about

two orders of magnitude. There is a high and significant ($p < 0.01$) correlation, and the data appears to follow a law-like curve. The inverse of the best-fit slope is 6.34; i.e. this is the combinatorial degree *if* we could assume encephalization quotient is proportional to behavioral complexity. This suggests that either hypothesis (a) or (c) applies to animal behavior as a combinatorial system; the number of muscle types increases to achieve greater behavioral complexity. Is the combinatorial degree invariant? A plot of the data under the logarithmic assumption [Fig. 5(b)] suggests not, as the plot is slightly more correlated than under the power assumption, and thus hypothesis (c) receives more support. To further test whether hypothesis (c) applies, I computed from the inverse of the instantaneous slope of the log–log plot how the combinatorial degree changes as the encephalization quotient increases [Fig. 5(c)], and one can see that the combinatorial degree does tend to increase, from around 3 to around 9. I should reiterate that the absolute magnitudes of these combinatorial degree values are only meaningful if we can assume that encephalization quotient scales proportionally with behavioral complexity; the fact that they are *increasing* is more important at the moment. These analyses suggest preliminary, weak support for the proposition that behavioral complexity is increased (among mammals) by increasing both the number of muscle types *and* the average number of muscles involved in a behavior.

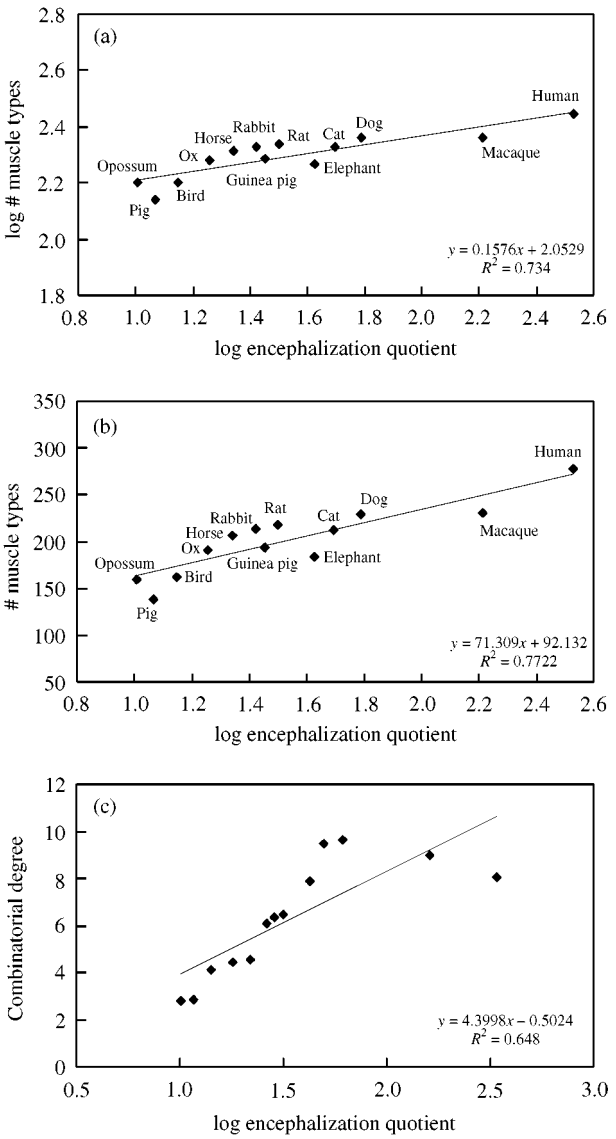


FIG. 5. (a) $\mathrm{Log}_{10}$ of the number of muscle types vs. $\log_{10}$ of the encephalization quotient, for 12 land mammals and for birds. One possible worry is that humans may have a greater muscle type count purely because they are much more well studied, and that the correlation may be being driven by this. Removing human from the plot leads to the equation $y = 0.164x + 2.044$ with a correlation only slightly reduced to $R^2 = 0.621$. Numbers of muscle types were taken as the maximum estimate counted from the following sources for each animal: human (Agur & Lee, 1991; Netter, 1997; Rohen & Yokochi, 1993; Anson, 1966), macaque (Bast *et al*., 1933), cat (Boyd *et al*., 1991; McClure *et al*., 1973; Reighard & Jennings, 1929; Hudson & Hamilton, 1993), rat (Greene, 1935; Hebel & Stromberg, 1976; Howell, 1926), rabbit (Wingerd, 1985; Craigie, 1966; McLaughlin & Chiasson, 1990; Popesko *et al*., 1990), guinea pig (Cooper & Schiller, 1975), dog (Adams, 1986; Boyd *et al*., 1991), horse (Budras & Sack, 1994; Way & Lee, 1965), ox (Ashdown & Done, 1984; Singh & Roy, 1997), pig (Sisson & Grossman, 1953), elephant (Mariappa, 1986), bird (Chamberlain, 1943; Kaupp, 1918; Nickel *et al*., 1977), opossum (Ellsworth, 1976). Encephalization quotients were computed using brain volumes from Frahm *et al*. (1982), Hofman (1982a, b, 1983, 1985), Stephan *et al*. (1981), Haug (1987), Hrdlicka (1907); and body masses from Nowak (1999), Hrdlicka (1907). (b) Number of muscle types vs. $\log_{10}$ of the encephalization quotient, for the same data. Correlation is slightly higher under this logarithmic assumption than in the power-law assumption from (a). If it is truly logarithmic, then combinatorial degrees measured by the inverse of the instantaneous slope (measured via linear regression for a moving window 13 data points wide) from the log–log plot in (a) should increase as the encephalization quotient increases; and they do, as shown in (c).

ORGANISMS

Cells (the components) in organisms are combined to make organism-expressions in organisms (the expressions). Estimates of the number of cell types for a large range of organisms have been catalogued by Bell & Mooers (1997). In making sense of what an "organism-expression" is, I assume that cells combine to make expressions or parts (probably functional parts) of some kind, and by an "organism-expression" I refer to such parts. Maybe organs are one kind of organism-expressions. For our main task it will not be crucial to pinpoint to what kind of biological structure organism-expressions correspond; we will make a general assumption below about organism-expressions—assumption (*)—that will enable us to test between hypotheses (a) and (c) without having to answer the question of what do organism-expressions correspond to .

How should the number of organism-expressions be measured? One possible measure of the expression complexity of an organism is the amount of the organism's DNA which codes for proteins, the *coding genome size*. There is a large and unsolved explanatory gap between genome complexity and phenotype complexity, but it is reasonable to expect coding genome size to *correlate* with the number of organism-expressions. Figure 6(a) shows that the number of cell types increases with increasing coding genome size ($p < 0.01$). If we believe that coding genome size is highly and significantly positively correlated with expression complexity, then it follows from Fig. 6(a) that (i) the number of cell types is probably positively correlated with expression

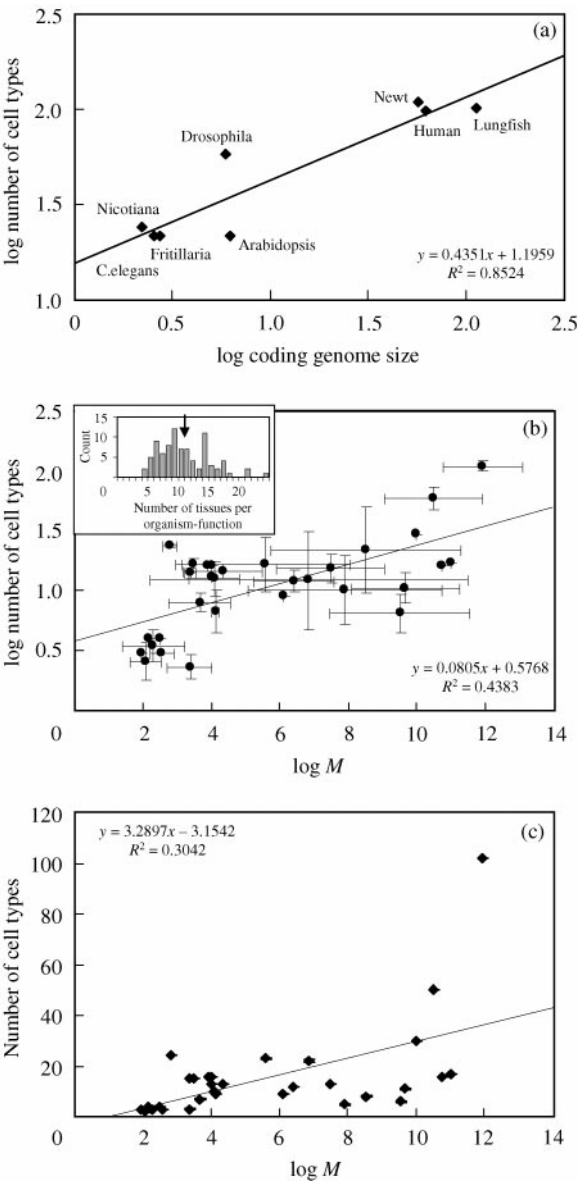


FIG. 6. (a) $\text{Log}_{10}$ of the number of cell types vs. $\log_{10}$ of the coding genome size (picograms of genes coding for proteins) for some plants and animals ($n = 8$). Coding genome size data are taken from Cavalier-Smith (1985). Cell type data are obtained from Bell & Mooers (1997). A similar plot, but for total DNA rather than coding DNA, appears in Kauffman (1969, Fig. 13; 1993, Fig. 12.7), motivated by different reasons. (b) $\text{Log}_{10}$ of the number of cell types vs. $\log_{10}$ of the total number of cells for plant, animal, fungi, Chlorophyta, Phaeophyta, Rhodophyta, Ciliata, Acrasiomycota and Myxomycota phyla ($n = 31$). Value for each phylum from Bell & Mooers (1997); error bars show standard deviation. [(b), inset] Distribution of number of tissues per organ in vertebrates. Arrow indicates the mean. Organs used are: heart, aorta, tonsil, lymph node, spleen, thymus, skin, endocrine and apocrine sweat glands, sweat and sebaceous glands, hair follicle and nail, tongue, submandibular gland, parotid gland, sublingual gland, soft palate, teeth, lip, esophagus, esophagogastric junction, stomach, gastroduodenal junction, duodenum, jejunum, ileum, colon, appendix, anorectal junction, liver, gallbladder, pancreas, olfactory mucosa, larynx, trachea and bronchus, bronchiole and respiratory passages, alveoli, kidney, ureter, urinary bladder, pituitary gland, pineal gland, parathyroid and thyroid glands, thyroid follicle cells, adrenal gland, testis, ductuli efferentes and epididymis, spermatic cord and ductus deferens, prostate gland, seminal vesicle, ovary, corpus luteum, oviduct-uterine tube, uterus, cervix, placenta, vagina, mammary glands, eye, ear, organ of corti. (c) Number of cell types vs. $\log_{10}$ of the total number of cells for the same data as in (b).

complexity, and (ii) the relationship appears to be lawful. Thus, this suggests that organisms as combinatorial systems are described by either hypothesis (a) or hypothesis (c) from the introduction; that is, it seems clear from Figure 6(a) that the number of cell types increases. Although coding genome size and the number of organism-expressions of an organism are probably positively correlated, I do not know their quantitative relationship, and so this plot cannot help us to determine whether or not the expression length is invariant. [One might be worried that Fig. 6(a) is uninteresting because a greater number of cell types would imply a greater coding genome size, so of course they are correlated. But coding genome size is probably due to organism complexity at a higher level than cell type complexity, possibly what I am calling expression complexity. It is *a priori* possible that a large variation in expression complexity is achieved by an invariant number of cell types [hypothesis (b)], in which case Fig. 6(a) would be flat. Alternatively, it is also *a priori* possible that a large variety of numbers of cell types all lead to the same expression complexity (e.g. because when $C$ is greater, either $L$ or the proportionality constant is smaller), in which case Fig. 6(a) would have variation on the $y$-axis, but it would not correlate with the $x$-axis.]

In order to test the two hypotheses—hypothesis (a) that $C \sim E^b$, and, alternatively, hypothesis (c) that $C \sim \log E$—we must make the simplifying assumption (∗) that *as organisms get more expressively complex, the number of times any given organism-expression is instantiated in the organism does not, by itself, tend to change*. Another way of stating this plausible, but difficult to defend, assumption (∗) is this: organisms become more expressively complex by increasing the number of different organism-expressions, not by increasing the number of occurrences of any given organism-expression. An immediate consequence of assumption (∗) is that the total number of cells in an organism will be proportional to the number of organism-expressions multiplied by the typical number of cells in an organism-expression; that is, $M \sim EL$, where $M$ is the total number of cells. With assumption (∗) in hand, we will be able to use the mass of an organism as a surrogate for the number of organism-expressions.

Before continuing, note that we are interested here in the relationship between the number of cell types and expression complexity *at the largest phylogenetic scales*. So, for example, if I say that $L \sim M^0$, I mean that across *all* (multicellular) organisms there is no overall trend, either up or down, in a plot of $L$ vs. $M$. This does not, however, preclude certain *local* trends, such as perhaps an increase in $L$ as a function of $M$ among mammals.

Let us now test hypothesis (a). If $C$ and $E$ are related by a power law, then expression length $L$ is invariant, and so $M \sim E$ (recall that from assumption (∗) we can conclude that $M \sim EL$). That is, we can use the total number of cells in the organism—or, equivalently, organism mass—as a proxy for expression complexity. Thus, if $C \sim E^b$ for some constant $b$, then it follows that $C \sim M^b$. We may test this prediction by plotting $\log C$ vs. $\log M$. Figure 6(b) shows that the plot, for which $M$ ranges over 31 phyla and about 10 orders of magnitude, appears to be reasonably linear, although by no means hugging a line. We will see that this is in stark contrast to the plot resulting from the hypothesis that $C$ is logarithmic in $E$, where the plot is very far from predicted.

To test the second hypothesis—hypothesis (c), that $C \sim \log E$—we again assume that it is true, make a prediction, and see if the prediction holds up. Under assumption (∗), it follows just as before that $M \sim EL$. But now $L \sim (\log E)/\log(\log E)$ (see introduction), and so $M \sim (E \log E)/\log(\log E)$. Ideally, we would like to solve for $E$ in terms of $M$—i.e. to have $E(M)$—and to plot $C$ vs. $\log(E(M))$. $M$, however, is a close enough approximation to $E(M)$ for our purposes, as $(E \log E)/\log(\log E)$ scales against $E$ nearly proportionally (exponent is 1.047) over our range of interest. Thus, we may plot $C$ vs. $\log M$. In sum, if $C \sim \log E$, then it should approximately be the case that $C \sim \log M$. We may test this prediction by plotting $C$ vs. $\log M$. Figure 6(c) shows that the plot is far from linear, curving upward which is the usual behavior when power laws are plotted with only the $x$-axis logged.

Hypothesis (a), then, is best supported by the data: at the largest phylogenetic scale, organisms appear to achieve greater expression complexity via an increase in the number of cell types, not via

an increase in the number of cells involved in any given organism-expression. It is interesting to note, then, that one kind of organism complexity, expression complexity (or the number of organism-expressions), may be very easy to measure, in contrast to the difficulties often inherent in measuring complexity (McShea, 1996): expression complexity may be, at the largest phylogenetic scale, directly proportional to mass. It should be reiterated, however, that these conclusions—and the upcoming discussion—require the simplifying assumption (∗).

The scaling exponent in Fig. 6(b) is 0.0805, implying a combinatorial degree of $1/0.0805 = 12.42$ (95% confidence interval is [10.07, 16.22]). Why might the combinatorial degree for organisms be in this range? Organisms do not seem to have many parts that could be said to have only a dozen or so cells, so why is the combinatorial degree so low? In an effort to understand this combinatorial degree value, I hypothesized that perhaps the more appropriate components for organism-expressions are *tissues*. The number of tissue types probably scales in direct proportion to the number of cell types, and so the scaling exponent above probably applies to the growth in the number of tissue types as well. Oversimplistically, the idea is that any given tissue type is built largely out of one type of cell (since they scale proportionally), and it is the arrangement of tissues that makes organism-expressions. In this way, the number of cells in an organism-expression may be much larger than the combinatorial degree. My hypothesis leads, then, to the prediction that there are, on average, around 12 *tissues* per organism-expression. To test this, I would like to be able to count up the actual number of tissues in a variety of organism-expressions. The difficulty, though, is that I have thus far left it extremely vague as to what an organism-expression corresponds to in an organism (assumption (∗) allowed us to, thus far, avoid this question). One very natural kind of organism-expression would seem to be *organs*. I do *not* wish to *define* organism-expressions as organs, but only wish to suggest that perhaps organs are paradigmatic cases of organism-expressions, and that counting the number of tissues in them may give us some insight into roughly how many tissues are in organism-expressions more generally. To do this we would ideally like estimates of the number of tissues making up organs for a variety of vertebrates, invertebrates, plants and other phyla, but in lieu of such a grand study, which is beyond the scope of this paper, as a preliminary test of this I used a standard vertebrate histology textbook (Ross and Romrell, 1995) to estimate the number of tissues present in 63 vertebrate organs [they are listed in the legend of Fig. 6(b), inset]. The average number of tissues is 10.52 ($\pm$ 4.17), indicated roughly by the arrow in the histogram (Fig. 6, inset), which is within the 95% confidence interval of the measured combinatorial degree of 12.42.

These results and ideas suggest a certain story concerning the evolutionary tendency for size to increase over the history of life (Bonner, 1988). Given assumption (∗), $M \sim EL$, and thus organisms that are twice as expressively complex must have at least twice the mass. If expression complexity has increased (either via selection or diffusion), then mass will have tended to increase. How mass increases with increasing expression complexity, however, depends on which way organisms as combinatorial systems work: hypothesis (a), (b) or (c), in particular. If $L \sim E^0$ as in hypothesis (a), then $M \sim E$. If, instead, $L \sim (\log E)/\log(\log E)$ as in hypothesis (c), then $M \sim (E \log E)/\log(\log E)$; that is, $M$ now scales more quickly as a function of $E$. Finally, if $L \sim \log E$ as in hypothesis (b), then $M \sim E \log E$; $M$ scales most quickly for this hypothesis. Therefore, if organisms have been selected against unnecessarily large size, then we would expect organisms to follow hypothesis (a); i.e. to keep $L$ invariant. Within this story, then, the number of cell types increases in order to minimize the mass required to obtain its organism-expressions.

An alternative viewpoint is that there is simply an upper limit to the number of cells (or tissues) that can be used in an organism-expression. If, as discussed above, the number of tissues per organism-expression really is around 12—and we should of course treat this with great skepticism at this point—*why* are there roughly a dozen tissues per organism-expression? Might there be some physico-mathematical barrier to more? One conjecture is that only so many tissues can be practically packed into one place in three dimensions. In particular, around 12 spheres can

be packed around a single sphere (where the spheres are all the same size), and perhaps the combinatorial degree is influenced by this. Within this alternative viewpoint, then, the number of cell types increases because more than around 12 tissues per organism-expression is impractical (without becoming highly convoluted and branched).

These previous two stories are not necessarily opposed to one another: perhaps mass-minimization drives the combinatorial degree to be invariant, but difficulties in having more than around 12 tissues per organism-expression explains the particular value of the combinatorial degree.

It is interesting to note that, in ontogeny, the number of cell types does *not* appear to increase in an interesting combinatorial fashion with expressive complexity, supposing still that expressive complexity may be measured by the total number of cells. Figure 7 plots the log of the number of cell types vs. the log of the total number of cells for the developing nematode *Caenorhabditis elegans*, and one may observe that for most of the development the number of cell types increases proportionally with the total number of cells.

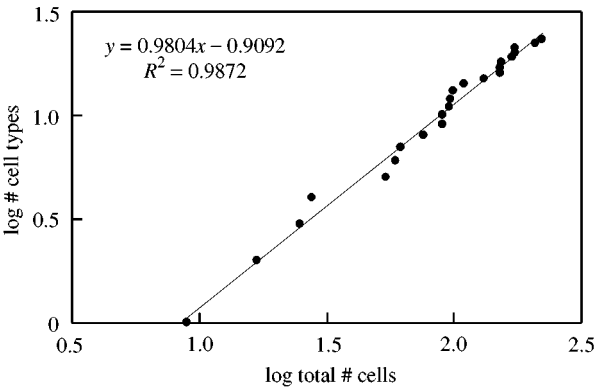


FIG. 7. $\text{Log}_{10}$ of number of cell types vs. $\log_{10}$ of total number of cells in *C. elegans* during gastrulation. The scaling exponent is 0.98 ($R^2 = 0.9872$, $n = 22$), or nearly 1. Here I list the cell type distinctions made; in square brackets next to each type I have put (a) the label, if there is one, for the type of cell from Sulston and White in Appendix 1 of Wood (1988), and (b) the number of cells of that type. *Six kinds of epithelial cell types*: (1) main hypodermis [hyp7, 83], (2) rectal hypodermis [rect, 4], (3) head hypodermis [27], (4) tail hypodermis [hyp8-12, 6], (5) interfacial [arc, 9], (6) seam [35]. *Three kinds of nervous tissue*: (7) neuron [302], (8) socket [23], (9) sheath [23]. *Ten kinds of mesoderm*: (10) head [hmc, 1], (11) anal depressor [mu anal, 1], (12) body [mu body, 79], (13) intestinal [mu int, 2], (14) pharynx [m, 46], (15) sphincter [mu sph, 1], (16) uterine [mu ut, 8], (17) vulval [mu vul, 8], (18) coelomocyte [cc, 6], (19) pharyngeal marginal [mc, 9]. *Two kinds of intestinal tissue*; (20) tube [int, 20], (21) valve [v, 8]. *Two kinds of gland*: (22) g1 [g1, 3], (23) g2 [g2, 2]. *Finally, four kinds of excretory cells*: (24) exc cell [1], (25) duct [1], (26) gland [2], (27) socket [1]. Data are for hermaphrodites only, and the founder and blast cells were excluded from the analysis.

## UNIVERSITIES

Departmental concentrations (the components) for undergraduate students are combined to get academic degrees (the expressions). One example degree might be "physics and mathematics", a degree which combines concentrations in physics and mathematics, and another degree might be "mathematics", which consists of just a single concentration in mathematics; each such degree counts as one degree, no matter how many students may earn it. I assumed that a school with twice the number of students will have twice as many distinct degrees—i.e. "we are all individuals (up to a constant proportion)"—and I used the number of students as a measure of the number of distinct degrees conferred at the university. I used the number of departments as a measure of the number of departmental concentration types.

Figure 8 shows the relationship between the number of departments $C$ and the number of students $E$ for 89 U.S. and Canadian colleges and universities, with the numbers of students ranging over two orders of magnitude. There is a highly lawful relationship between $C$ and $E$ here, and the number of departments can clearly be seen to greatly increase ($p < 0.01$) as the number of students increases. One of the hypotheses (a) and (c) seem, then, to apply here. Under a power law hypothesis [hypothesis (a)] the data appears to be fairly linear [Fig. 8(a)], but under a logarithmic law hypothesis [hypothesis (c)] the data bend upward and become increasingly scattered [Fig. 8(b)]. These results suggest that it is hypothesis (a) that applies to universities as combinatorial systems: demand for more academic degrees is filled by having students choose from more departmental concentration types, not by having students have degrees consisting of greater and greater numbers of departmental concentrations.

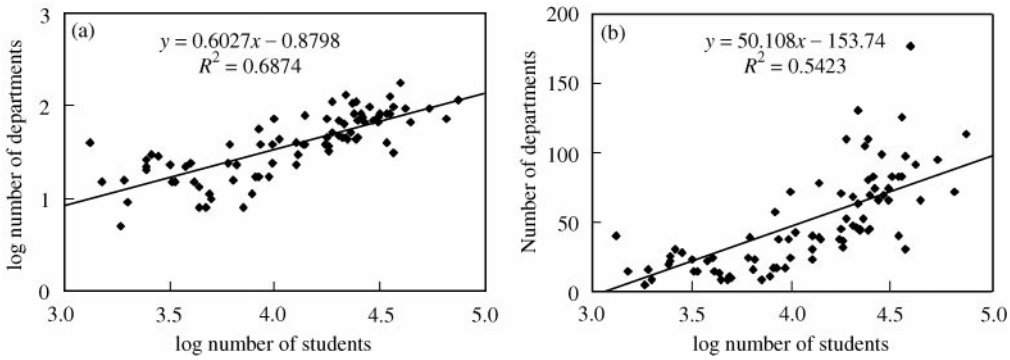


FIG. 8. (a) $\text{Log}_{10}$ of the number of departments vs. $\log_{10}$ of the number of students, for 89 U.S. and Canadian colleges and universities (*World of Learning*, 2000). (b) Number of departments vs. $\log_{10}$ of the number of students, for the same data.

In the power law plot [Fig. 8(a)] the scaling equation is $C \sim F^{0.6027}$. This scaling exponent implies a combinatorial degree of 1.66 (95% confidence interval [1.45, 1.94]). Why is the exponent what it is? One obvious possible explanation for this combinatorial degree is that it is set by the average number of departmental concentrations per student. Counting only majors and minors as concentrations, the average numbers of majors/minors per person at Duke, UVA and Williams College are, respectively, 1.75, 1.17 and 1.39, roughly within the range of the combinatorial degree. (Averages for Duke and UVA are for Spring semester, 2000. The value for Williams College is averaged over the average for each year from 1991 to 2000, and has standard deviation for those 10 years of 0.0344.)

## ELECTRONIC DEVICES

Button-presses (the components) for electronic devices are combined by the user of the device to carry out device-actions (the expressions). Note that the combinatorial system is not the electronic device itself; it is, rather, the electronic device's user-interface language that is the combinatorial system. Each button on the device is a type of button press. I used the number of pages of the device's user's manual as a measure of the number of device actions (i.e. of the number of things the device does); it seems reasonable to expect that if a device does twice as many things, its manual will be twice as long. Number of buttons (i.e. numbers of button-press types) and manual pages were gathered from calculators, compact disk (CD) players, televisions (TVs) and video cassette recorders (VCRs). For CD players, TVs and VCRs, the number of buttons was taken from remote controls. Calculators were limited to those without full a-to-z keyboards.

Plots (Fig. 9) show that component-type complexity increases with expression complexity in all four cases ($p < 0.01$), and does so in a law-like fashion: as electronic devices do more and more things via strings of button-presses, the number of button-press types (i.e. the number of buttons) increases. Distinguishing whether these plots follow power laws or logarithmic laws, though, is not possible given the very small range for the number of buttons; neither assumption leads to a significantly different fit with the data for any of the four kinds of electronic devices. The logarithmic plots (not shown) look nearly identical to the power-law plots.

Despite our inability to determine which of hypotheses (a) or (c) applies to electronic devices as combinatorial systems, it is instructive to suppose for the moment that they follow power laws [hypothesis (a)] and to look at what the scaling exponents are. The exponents are 0.114 for calculators, 0.4835 for CD players, 0.631 for TVs and 0.2529 for VCRs. The combinatorial degrees are, accordingly, 8.77 for calculators, 2.07 for CD players, 1.58 for TVs, and 3.954 for VCRs. I have not attempted to understand the particular combinatorial degree in each case, but it is clear that CD players, TVs and VCRs require many fewer button-presses to carry out an action than

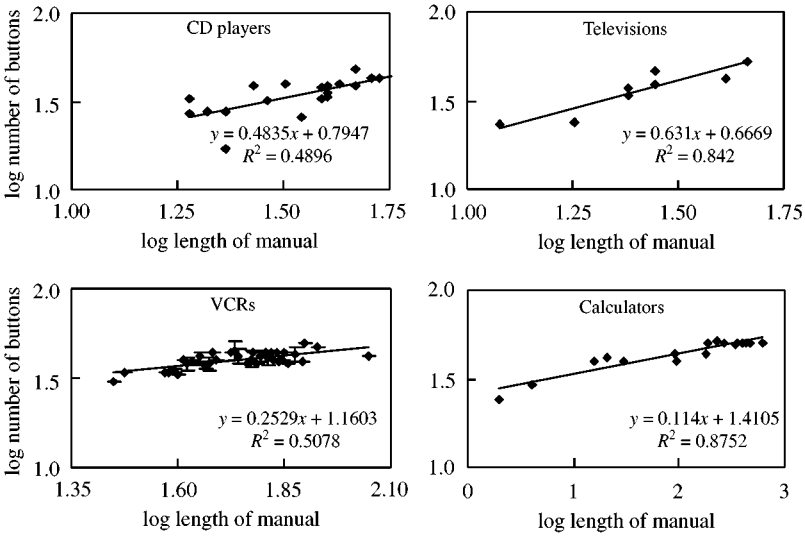


FIG. 9. $\text{Log}_{10}$ of the number of buttons vs. $\log_{10}$ of the length of manual, for CD players ($n = 20$), TVs ($n = 8$), VCRs ($n = 36$, averaged from 78 devices) and calculators ($n = 17$).

a calculator—for example, simply adding two plus two requires four button presses on a calculator—and this is evidenced by the much higher combinatorial degree for calculators.

## Discussion and Conclusion

We have seen that in each of the radically different kinds of combinatorial systems, (1) component-type complexity and expression complexity appear to be related in a law-like fashion with the number of component types increasing as expression complexity increases, (2) in five kinds of combinatorial systems (bird vocalization, English language over history, organisms in phylogeny and ontogeny, and universities; see Table 1) the increase in expression complexity appears to be accomplished solely by increasing component-type complexity, not by increasing the number of components per expression, and (3) in three kinds of combinatorial systems (both cases of ontogeny of language, and mammalian behavior; see Table 1), the increase in expression complexity appears to be achieved by increasing both component-type complexity and the number of components per expression. (For the four electronic device cases, it could not be determined which of the hypotheses (a) or (c) applies.)

The five kinds of systems fitting hypothesis (a) increase expression complexity by

(I) conforming to the proportionality equation $E \sim C^L$ (i.e. scale-invariant proportionality factor), and
(II) having a scale-invariant expression length—$L \sim E^0$—and combinatorial degree,

and thus the number of component types increases as a power-law function of the number of expressions ($C \sim E^d$). The three kinds of combinatorial systems conforming to hypothesis (c) increase expression complexity by

(I) conforming to the proportionality equation $E \sim C^L$ (i.e. scale-invariant proportionality factor), and
(III) having expression length and combinatorial degree that increases with the number of expressions more slowly than logarithmically,

and thus the number of component types increases logarithmically with the number of expressions ($C \sim \log E$). Are there general attributes of combinatorial systems that explain why they would satisfy (I), and one of either (II) or (III)? That is, what are the similarities and differences between these kinds of combinatorial system? It is to this we turn next.

Let us begin with (I): what property or properties might a kind of combinatorial system have that would lead it to satisfy $E \sim C^L$? We have already touched upon this in the introduction during our discussion of hypotheses of category 1: a constant proportionality factor relating $E$ and $C^L$ strongly suggests a *scale-invariant grammar*. Accordingly, I hypothesize that $E \sim C^L$ in these varieties of combinatorial system because each has a scale-invariant grammar (I will say "grammar-invariant").

Consider now attribute (II): what attribute or attributes might a grammar-invariant kind of combinatorial system possess that would explain why increasing expression complexity is achieved solely by increasing the number of component types [or, equivalently, explain why expression length (and combinatorial degree) is scale-invariant]? There are at least three possible attributes that would explain a scale-invariant combinatorial degree, and I will take each up in turn.

One reason a grammar-invariant combinatorial system may have an invariant combinatorial degree is simply that there is an upper limit on the combinatorial degree, but not an upper limit on the number of component types. We have implicitly suggested such an explanation for some of our combinatorial systems: for example, perhaps working memory limits in human and bird explain the combinatorial degrees of around 5 and 1 for, respectively, English and bird vocalization; and perhaps there could be tissue-packing limits explaining the combinatorial degree of around 12 for organism complexity.

A second reason a grammar-invariant combinatorial system may have an invariant combinatorial degree is if there is some kind of pressure to minimize the total number of components (not component types) in the combinatorial system. The total number of components in a system with $E$ expressions is $EL$. This is minimized when $L$ is minimized; in particular, when $L$ is invariant. This kind of explanation does not have the ability to predict why the combinatorial degree should be of any particular value, and although it may apply to organisms (as discussed earlier), it is not clear that it applies to the other systems studied here.

The third possible reason a grammar-invariant combinatorial system may have an invariant combinatorial degree is more subtle. No maximum limit on the combinatorial degree needs to be assumed. To understand this we must introduce the notion of compositionality. A kind of combinatorial system is *compositional* if, among combinatorial systems of that kind, components of different types do different things, and what an expression does (i.e. the expression's semantics) is a consequence of both the arrangement of the constituent components (i.e. the expression's syntax) and what each constituent component does (i.e. each constituent component's semantics). For compositional kinds of combinatorial system there is a clear reason why the number of component types should increase and expression length remain invariant: in order for an expressively more complex system to have expressions that do truly novel things, it is not enough to simply increase the expression length; truly novel component-types must be added. This will be clear with an example. Consider combinatorial systems that may only use the component types "table", "chair", "is on top of" and the standard symbols from logic (i.e. "for all", "there is", "and", "or", "not"). Let us suppose that combinatorial system A consists of all sentences that may be formed using two occurrences from either "table" or "chair", and $B_1$ consists of all sentences that may be formed using three occurrences from either "table" or "chair". System A possesses expressions like, "There is a table that is on top of a chair," whereas $B_1$ has, in addition, expressions like, "There is a table that is on top of a chair that is on top of a table." Combinatorial system $B_1$ certainly says more things than A, but, informally, nothing truly novel, or at right angles, to anything A can say. Consider now combinatorial system $B_2$, which consists of all sentences that may be formed using two occurrences from either "table", "chair", or the *new* component type "shelf". $B_2$ has all the expressions of A, but has, in addition, expressions like "There is a table that is on top of a shelf." Like $B_1$, $B_2$ has more expressions than A; but unlike $B_1$, $B_2$ says truly novel things—$B_2$ has expressions about shelves whereas $B_1$ does not. $B_2$ is intuitively *richer* than A, but $B_1$ is *not* richer than A. I will say that a kind of combinatorial system is *rich* if, among combinatorial systems of that kind, greater expressivity is obtained only for reasons of greater novelty in

the sense above. Now, if a grammar-invariant kind of combinatorial system is compositional *and* rich, then the expression length will not increase beyond some minimum length needed to have useful meaning, but the number of component types will increase instead. That is, the combinatorial degree will remain invariant, but not because of an upper bound limiting it, but because there is simply never anything interesting to say requiring more than a certain length of expression.

Attribute (III) from above, characteristic of combinatorial systems satisfying hypothesis (c), is most obviously understood in terms of the first reason discussed for attribute (II). When the upper limit on combinatorial degree increases as the number of expressions increases, but does so more slowly than logarithmically, the number of component types would increase more slowly than as a power law. This appears to be the case for the ontogeny of language and perhaps for mammalian behavior.

While all this may help us to understand possible similarities and differences among the diverse kinds combinatorial system, it does not aid in discerning their specific differences in scaling behavior: each kind of system has its own combinatorial degree (or range of combinatorial degrees), with values ranging from about 1 to 12. Explaining these particular values will depend on the particular kind of combinatorial system, and I entertained explanations for the particular values when I discussed each kind of combinatorial system.

I wish to thank Daniel McShea, W. G. Hall, Zhi-Yong Yang, Erich Jarvis, Christopher Sturdy, Robert McGehee, Dan Ryder, Brian Hayes, James L. Fidelholtz, Chip Gerfen, Reiko Mazuka, Craig Melchert, Geoffrey Sampson and Thomas Wasow for helpful advice and criticism.

## REFERENCES

ADAMS, A.-M. & GATHERCOLE, S. E. (2000). Limitations in working memory: implications for language development. *Int. J. Lang. Comm. Dis.* **35,** 95–116.

ADAMS, D. R. (1986). *Canine Anatomy*. Ames: The Iowa State University Press.

AGUR, A. M. R. & LEE, M. J. (1991). *Grant's Atlas of Anatomy*. Baltimore: Williams and Wilkins.

ALLMAN, J. M. (1999). *Evolving Brains*. New York: Scientific American Library.

ANSON, B. J. (1966). *Morris' Human Anatomy*. New York: McGraw-Hill.

ASHDOWN, R. R. & DONE, S. (1984). *Color Atlas of Veterinary Anatomy*: *The Ruminants*. Baltimore: University Park Press.

BAST, T. H., CHRISTENSEN, K., CUMMINS, H., GEIST, F. D., HARTMAN, C. G., HINES, M., HOWELL, A. B., HUBER, E., KUNTZ, A., LEONARD, S. L., LINEBACK, P., MARSHALL, J. A., MILLER JR, G. S., MILLER, R. A., SCHULTZ, A. H., STEWART, T. D., STRAUS JR, W. L., SULLIVAN, W. E. & WISLOCKI, G. B. (1933). *The Anatomy of the Rhesus Monkey*. Baltimore: Williams and Wilkins.

BELL, G. & MOOERS, A. O. (1997). Size and complexity among multicellular organisms. *Biol. J. Linnean Soc.* **60,** 345–363.

BONNER, J. T. (1988). *The Evolution of Complexity*. Princeton: Princeton University Press.

BOYD, J. S., PATERSON, C. & MAY, A. H. (1991). *Clinical Anatomy of the Dog and Cat*. St. Louis: Mosby.

BROWN, E. D., FARABAUGH, S. M. & VELTMAN, C. J. (1988). Song sharing in a group-living songbird, the Australian Magpie, *Gymnorhina tibicen*. Part I. Vocal sharing within and among social groups. *Behavior* **104,** 1–28.

BUDRAS, K-D. & SACK, W. O. (1994). *Anatomy of the Horse*: *An Illustrated Text*. London: Mosby-Wolfe.

CASE, R., KURLAND, D. M. & GOLDBERG, J. (1982). Operational efficiency and the growth of short-term memory span. *J. Exp. Child Psychol.* **33,** 386–404.

CAVALIER-SMITH, T. (1985). Eukaryote gene numbers, noncoding DNA and genome size. In: *The Evolution of Genome Size* (CAVALIER-SMITH, T., ed.), pp. 69–103. New York: John Wiley & Sons.

CHAMBERLAIN, F. W. (1943). *Atlas of Avian Anatomy*. East Lansing: Hallenbeck.

CHANGIZI, M. A. (2001). Principles underlying mammalian neocortical scaling. *Biol. Cybern.* **84,** 207–215.

CLARK, E. V. (1993). *The Lexicon in Acquisition*. Cambridge: Cambridge University Press.

CLARK, H. H. & WASOW, T. (1998). Repeating words in spontaneous speech. *Cog. Psychol.* **37,** 201–242.

COOPER, G. & SCHILLER, A. L. (1975). *Anatomy of the Guinea Pig*. Cambridge: Harvard University.

CORRIGAN, R. (1983). The development of representational skills. In: *Levels and Transitions in Children's Development*, (Fischer, K., ed.), San Francisco: Jossey-Bass. pp. 51–64.

CRAIGIE, E. H. (1966). *A Laboratory Guide to the Anatomy of the Rabbit*. Toronto: University of Toronto Press.

DEVOOGD, T. J., KREBS, J. R., HEALY, S. D. & PURVIS, A. (1993). Relations between song repertoire size and the volume of brain nuclei related to song: comparative evolutionary analyses amongst ocine birds. *Proc. R. Soc. Lond. B* **254,** 75–82.

ELLSWORTH, A. F. (1976). *The North American Opossum*: *An Anatomical Atlas*. Huntington: Robert E. Krieger Publishing.

FRAHM, H. D., STEPHAN, H. & STEPHAN, M. (1982). Comparison of brain structure volumes in Insectivora and Primates. I. Neocortex. *J. Hirnforschung* **23,** 375–389.

GALLISTEL, C. R. (1980). *The Organization of Action*: *A New Synthesis*. Hillsdale: Lawrence Erlbaum Associates.

GREENE, E. C. (1935). *Anatomy of the Rat*. Philadelphia: The American Philosophical Society.

HAILMAN, J. P. (1989). The organization of major vocalizations in the paradae. *Wilson Bull.* **101,** 305–343.

HAUG, H. (1987). Brain sizes, surfaces and neuronal sizes of the cortex cerebri: a stereological investigation of man and his variability and a comparison with some mammals (primates, whales, marsupials, insectivores, and one elephant). *Am. J. Anatomy* **180,** 126–142.

HEBEL, R. & STROMBERG, M. W. (1976). *Anatomy of the Laboratory Rat*. Baltimore: Williams and Wilkins.

HOFMAN, M. A. (1982a). Encephalization in mammals in relation to the size of the cerebral cortex. *Brain Behav. Evol.* **20,** 84–96.

HOFMAN, M. A. (1982b). A two-component theory of encephalization in mammals. *J. theor. Biol.* **99,** 571–584.

HOFMAN, M. A. (1983). Evolution of brain size in neonatal and adult placental mammals: a theoretical approach. *J. theor. Biol.* **105,** 317–332.

HOFMAN, M. A. (1985). Size and shape of the cerebral cortex in mammals. I. The cortical surface. *Brain Behav. Evol.* **27,** 28–40.

HOWELL, A. B. (1926). *Anatomy of the Wood Rat*. Baltimore: Williams and Wilkins.

HRDLICKA, A. (1907). *Brain Weight in Vertebrates*, pp. 89–112. Washington, D.C.: Smithsonian Miscellaneous Collections.

HUDSON, L. C. & HAMILTON, W. P. (1993). *Atlas of Feline Anatomy for Veterinarians*. Philadelphia: W. B. Saunders.

HUNT, K. W. (1965). A synopsis of clause-to-sentence length factors. *English J.* **54,** 300–309.

INCE, S. A. & SLATER, P. J. B. (1985). Versatility and continuity in the songs of thrushes *Turdus* spp. *Ibis* **127,** 355–364.

JERISON, H. (1973). *The Evolution of the Brain and Intelligence*. New York: Academic Press.

KAUFFMAN, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J. theor. Biol.* **22,** 437–467.

KAUFFMAN, S. A. (1993). *The Origins of Order*. New York: Oxford University Press.

KAUPP, B. F. (1918). *The Anatomy of the Domestic Fowl*. Philadelphia: W. B. Saunders.

KROODSMA, D. E. (1977). Correlates of song organization among North American wrens. *Am. Nat.* **111,** 995–1008.

KROODSMA, D. E. (1984). Songs of the alder flycatcher (*Empidonax alnorum*) and willow flycatcher (*Empidonax Traillii*) are innate. *Auk* **101,** 13–24.

MARIAPPA, D. (1986). *Anatomy and Histology of the Indian Elephant*. Oak Park, MI: Indira Publishing House.

MCLAUGHLIN, C. A. & CHIASSON, R. B. (1990). *Laboratory Anatomy of the Rabbit*. Dubuque: Wm. C. Brown Publishers.

MCCLURE, R. C., DALLMAN, M. J. & GARRETT, P. D. (1973). *Cat Anatomy: An Atlas, Text and Dissection Guide*. Philadelphia: Lea and Febiger.

MCSHEA, D. W. (1996). Metazoan complexity and evolution: is there a trend? *Evolution* **50,** 477–492.

MILLER, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* **63,** 81–97.

MUNDINGER, P. C. (1999). Genetics of canary song learning: innate mechanisms and other neurobiological considerations. In: *The Design of Animal Communication* (Hauser, M. D. & Konishi, M., eds), pp. 369–390. Cambridge: MIT Press.

NETTER, F. H. (1997). *Atlas of Human Anatomy*. New Jersey: East Hanover.

NICKEL, R., SCHUMMER, A., SEIFERLE, E., SILLER, W. G. & WIGHT, P. A. L. (1977). *Anatomy of the Domestic Birds*. New York: Springer-Verlag.

NOWAK, R. M. (1999). *Walker's Mammals of the World*. Baltimore: The Johns Hopkins University Press.

PASCUAL-LEONE, J. (1970). A mathematical model for the transition rule in Piaget's developmental stages. *Acta Psychol.* **32,** 301–345.

POPESKO, P., RAJTOVÁ, V. & HORÁK, J. (1990). *A Colour Atlas of the Anatomy of Small Laboratory Animals*. Bratislava: Wolfe Publishing.

READ, A. F. & WEARY, D. M. (1992). The evolution of bird song: comparative analyses. *Philos. Trans. R. Soc. Lond.* **B 338,** 165–187.

REIGHARD, J. & JENNINGS, H. S. (1929). *Anatomy of the Cat*. New York: Henry Holt and Company.

ROBINSON, B. F. & MERVIS, C. B. (1998). Disentangling early language development: modeling lexical and grammatical acquisition using an extension of case-study methodology. *Dev. Psychol.* **34,** 363–375.

ROHEN, J. W. & YOKOCHI, C. (1993). *Color Atlas of Anatomy*. New York: Igaku-Shoin.

ROSS, M. H., ROMRELL, L. J. & KAYE, G. I. (1995). *Histology: A Text and Atlas*. Baltimore: Williams and Wilkins.

SCUDDER, H. H. (1923). Sentence length. *English J.* **12,** 617–620.

SIEGEL, L. S. & RYAN, E. B. (1989). The development of working memory in normally achieving and subtypes of learning disabled children. *Child Dev.* **60,** 973–980.

SINGH, H. & ROY, K. S. (1997). *Atlas of the Buffalo Anatomy*. Pusa, New Delhi: Indian Council of Agricultural Research.

SISSON, S. & GROSSMAN, J. D. (1953). *The Anatomy of the Domestic Animals*. Philadelphia: W. B. Saunders.

STEPHAN, H., FRAHM, H. & BARON, G. (1981). New and revised data on volumes of brain structures in insectivores and primates. *Folia Primatol.* **35,** 1–29.

VELTEN, H. V. (1943). The growth of phonemic and lexical patterns in infant language. *Language* **19,** 281–292.

WAY, R. F. & LEE, D. G. (1965). *The Anatomy of the Horse*. Philadelphia: J. B. Lippincott.

WINGERD, B. D. (1985). *Rabbit Dissection Manual*. Baltimore: The Johns Hopkins University Press.

WOOD, W. B. (ed.) (1988). *The Nematode Caenorhabditis elegans*. Cold Spring Harbor: Cold Spring Harbor Press.

*World of Learning* (2000). London: Europa.