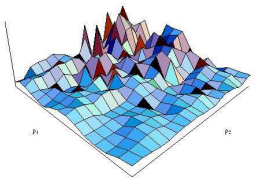
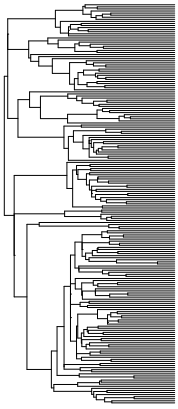
Lecture Notes: Language and Evolution

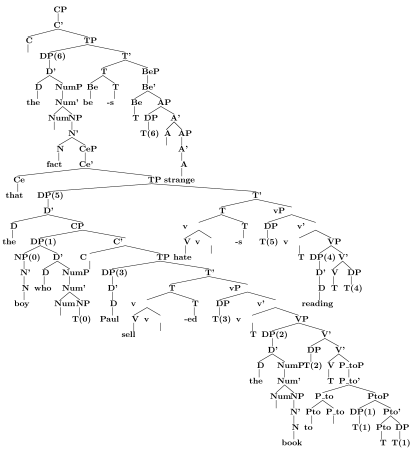
Edward Stabler

January 11, 2007

The study of evolution and language provides a unique opportunity for carefully examining basic questions about evolution, language, and the kinds of explanations available for sources of order in physical, biological, cognitive and cultural domains.







Spring 2006 Syllabus

The study of evolution and language provides a unique opportunity for carefully examining basic and important questions about evolution, language, and the kinds of explanations available for sources of order in physical, biological, cognitive and cultural domains.

Human languages provide a kind of mirror on human thought, and so we want to understand the forces that have shaped the structures we see there. Evolution provides a source of structure at two levels. First, the human organism has evolved, with linguistic abilities of certain kinds, by genetic transmission and natural selection. And second, each particular language is a cultural artifact, transmitted by learning and selected by various cultural and natural forces. In each case, we can ask: what aspects of language structure can be explained by evolutionary forces? And what other forces are shaping human languages?

These are fundamental questions that every thinking person is likely to be curious about, and so it is no surprise that there is a wealth of popular and scientific work addressing them. Readings will be drawn from classic and contemporary research, supplemented with lecture notes each week.

0.1 Some things you will know at the end of the class

- the fundamental axioms of Darwin's theory of evolution
- Mendel's argument for genetic "atoms" – "genes"
- Hardy's theorem about conditions under which a genetic population is stable
- some of the clues that led Watson and Crick to the discovery of DNA
- some idea of how DNA controls protein synthesis from amino acids
- how many basic elements are in the "language" of DNA (here's the answer now: 4)
- how many of these occur in the human genome (here's the answer now: ≈3,164,700,000)
- how many Mendelian atoms occur in the human genome (here's the answer now: ≈35,000)
- why AZT is ineffective against HIV (the terrible answer: HIV evolves *rapidly*)
- some critiques of the "neo-Darwinian" research program
- Frege's argument for "semantic atoms" and "compositionality" in human language
- how many basic gestures in each human language (here's the answer now: 11-160)

- how many Fregean atoms in each human language (here's the answer now: >10,000)
- what is the "Chomsky hierarchy," and where in it are DNA and human languages
- what do English, Quechua,[1] and American Sign Language (ASL) have in common
- how human language is "structure-dependent," unlike any other animal language
- how a language learner can be regarded as a mathematical function
- some critiques of the "neo-Lockeian" empiricist research program
- how nothing in the universe is really like Locke's "blank slate"
- ⋆ whether human language abilities emerged by selection: why the experts disagree
- how all these things relate to each other (!)

These are things everyone should know. Can we fit all this in? We'll try.

[1] Quechua is the language of the Incas, now spoken by approximately 7 million people in South America. The lecture notes have a glossary and index for less common names and terms like this one.

Contents

Lecture 0

Evolution and language

The main goal of the class will be to develop a clear conception of the fundamentals of evolution and of language, so that we can consider two questions: to what extent has the nature of human linguistic ability been shaped by natural selection, and to what extent can the development of particular languages, as cultural artifacts, be regarded as an evolutionary process of a similar kind. The class will be designed to foster clear, critical thinking about these matters, and to incite an interest in pursuing them further. With these goals, it is natural to focus on *clear* cases of evolution and on *well-understood* properties of language. Surprisingly, this makes the content of the class rather unusual, because it has been common to use evolutionary analogies as the basis for speculations about things that are not well-understood, and not in the mainstream of research in either evolutionary or linguistic theory: things like the extent to which Neanderthals might have had speech and communicative abilities, the relationship between brain size and "intelligence" (whatever that is), the nature of the neural structures that implement linguistic abilities, and how you should revise your religious beliefs in the light of all this. These matters are of great interest, of course, and so in some popular literature these are often the main thing (Deacon, 1997; Loritz, 1999; Lieberman, 2000; Calvin and Bickerton, 2000, …), but our concern here is the *real thing*, what is really known, and how. This is the secure foundation upon which the study of the more obscure topics must rest.

A long-standing and still essential component of linguistic study focuses on the description of particular languages, the origins of words, the variations in pronunciation, the shifts in meaning over time and across populations. But Chomsky and others in the 1950's and 1960's drew attention to some new kinds of questions: are there properties that all languages have in common, properties that must be due to something other than the particular characteristics of a common ancestral language? There are, and among those properties, some are apparently fundamental to our ability to acquire and use our language as we do. Of course, interests of the former kind continue, but this class will introduce questions of the latter kind, and consider some preliminary discussions of the clearest and most substantial results that have been obtained there. Very roughly, in all human languages, we find roughly similar kinds of pieces forming tightly interlocked structures, which over an infinite range can be "compositionally" interpreted by speakers of the language.

A similar shift can be seen in some reactions to the success of the "neo-Darwinian synthesis" in evolutionary biology. A long-standing and still essential component of evolutionary study focuses on the analysis of branching phylogeny (see Figure 1 on page 3): what emerged when, and why; what has been the role of selection and of drift in the history of life? But Gould, Kauffman and others in the 1970's and 1980's drew attention to some new kinds of questions: do common properties emerge on different branches of the phylogenetic tree, properties whose emergence must be due to some requirements of form that are independent of selection? There are, and understanding them is key to accounting for the emergence of complex "wholes." The account must go beyond the standard neo-Darwinian "myriad mutations, selection sifting" to recognize limitations of selection and the importance of other sources of order. This step brings earlier views like D'Arcy Wentworth Thompson's laws of form back into balance with Darwin's. This balance is essential for a proper appreciation of the question of how human language and other complex morphology and complex abilities could emerge in the forms we find.

Considering the clearest cases of evolutionary development first, the basic mechanisms of evolution will be studied. We will look briefly at some well-studied examples, including the clear and disastrous case of HIV evolution. This "retrovirus," with no DNA but only only two identical strands of RNA, is quite different from the large organisms that we are more familiar with, and because of its extremely rapid evolution, AZT and related treatments that succeeded in the short term have all failed in the long term. Such cases of extremely rapid evolution also provide examples of the emergence of dominant variants, "quasi-species." We then look briefly at some of the basic principles of evolutionary theory, using some examples to illustrate the roles of sources of order other than selection.

With this background, we turn to prominent ideas about the evolution of learning, noticing how selection can achieve a certain balance between rigidity and adaptation in organisms. This sets the stage for a rather careful look at human linguistic ability, which is rigid in some respects and plastic in others. We briefly survey first some of the distinctive features of human languages (features which, for the most part, spoken and signed and written languages all share). Finally, we will be in a position to really understand why the experts have conflicting views about the roles of natural selection, "exaptation," and laws of form and function in shaping human language.

As a last exercise, we turn to the study of how particular languages, particular cultural artifacts change over time. The tools for studying evolution can be applied to questions in historical linguistics. This is a relatively new field, but one that is booming with the advent of relevant computational methods for simulation, revealing the fundamental interplay between organismic plasticity and cultural transmission.

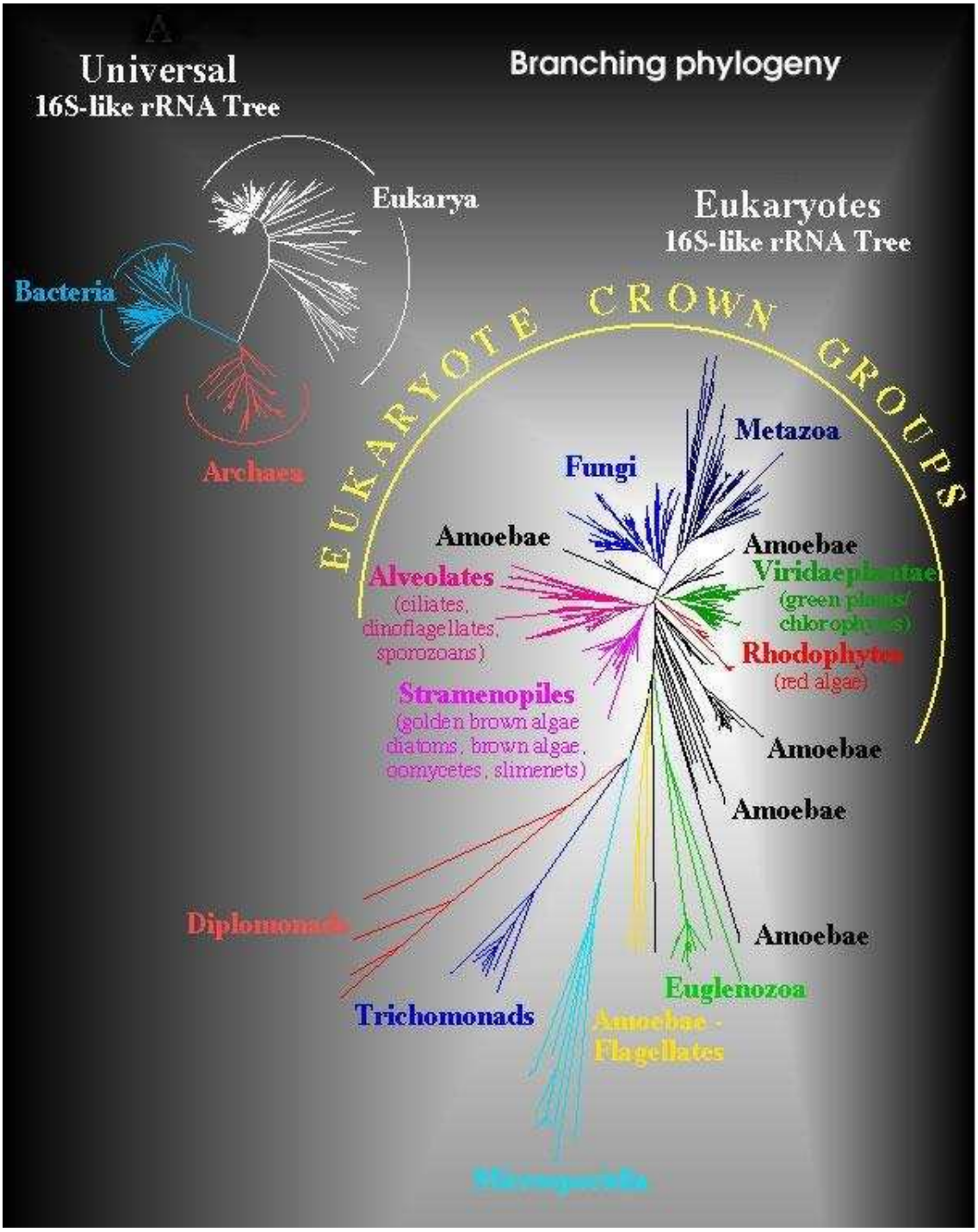
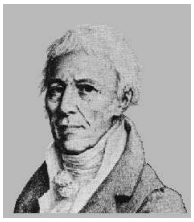


Figure 1: branching phylogeny calculated from genetic distances (Sogin and Patterson 1992)

0.1 Evolution: First ideas

Since every species has to exist in perfect harmony with its surrounding and since this surrounding is constantly changing, the species itself, too, has to change constantly, if it is to stay in a harmonic balance with its surrounding. If it would not adjust, the species would be threatened by extinction. – Lamarck 1809

I am almost convinced (quite contrary to the opinion I started with) that species are not (it is like confessing a murder) immutable. – 1844 letter from Darwin to Hooker


Lamarck

As Charles Darwin observes in the Preface to *Origin of Species*, the idea that the species are changing was nothing new. It had been already proposed by the French zoologist Buffon, and by Charles Darwin's grandfather, Erasmus Darwin. Buffon's student Jean-Baptiste Lamarck developed a more complete perspective in his 1809 Philosophie zoologique and his 1815 *Histoire naturelle des animaux sans vertèbres*. He proposes four fundamental laws. The second law specifies how characteristics are acquired (in response to some need of the organism), and the fourth law specifies a mechanism for transmission of change (inheritance). Lamarck formulates his four laws this way in *Histoire naturelle des animaux sans vertèbres*:[1]

1. Life by its proper forces tends continually to increase the volume of every body possessing it, and to enlarge its parts, up to a limit which it brings about.
2. The production of a new organ in an animal body results from the supervention of a new want continuing to make itself felt, and a new movement which this want gives birth to and encourages.
3. The development of organs and their force of action are constantly in ratio to the employment of these organs.
4. All which has been acquired, laid down, or changed in the organization of individuals in the course of their life is conserved by generation and transmitted to the new individuals which proceed from those which have undergone those changes.

The second law was hard to maintain even in 1815. And though it was common knowledge that offspring resemble the parents, the fourth law goes too far in suggesting that acquired traits are transmitted. But the combination of these ideas, change plus transmission by inheritance, provided a revolutionary perspective on life that is adopted and elaborated by Darwin.

[1] *Première loi:* La vie, par ses propres forces, tend continuellement à accroître le volume de tout corps qui la possède, e à éntendre les dimensions de ses parties, jusqu'à un terme qu'elle amène elle-même. *Deuxième loi:* La production d'un nouvel organe dans un corps animal, résulte d'un nouveau besoin survenu qui continue de se faire sentir, et d'un nouveau mouvement que ce besoin fait naître et entretient. *Troisième loi:* Le développement des organes et leur force d'action sont constamment en raison de l'emploi de ces organes. *4.*e *loi:* Tout ce qui a été acquis, tracé ou changé dans l'organisation des individus pendant le cours de leur vie, est conservé par la génération, et transmis aux nouveaux individus qui proviennent de ceux qui on éprouvé ces changemens. (pp182,185,189,199)


Darwin

Darwin sets out his basic ideas out in a clear and summary form in the last chapter of his *Origin of Species*. We can identify the following basic postulates on which his analysis is based:

1. There is variation in the traits of the individuals of each species
2. Some traits are inherited
3. Only some individuals survive long enough to reproduce; these are *naturally selected* to propagate their traits

We can notice some important differences from Lamarck. First, while Lamarck thought that the apparent adaptation of organisms to their ecological niches resulted from a history of traits acquired by the practice of one's ancestors, Darwin makes no such assumption. There is undirected variation, and there is selection. These factors alone are held to be responsible for the "creativity" that one seems to see in the adaptations of organisms. Also notice that Lamarck's second law is not replaced by any other idea about how variation is introduced into a species; it seems to be just provided by nature. Darwin had the view simply that all structures vary, and selection acts on the diversity nature provides.

What evidence is offered in support of these postulates, and the view that natural selection is a basic force behind the diversity of life?

Analogy with breeding and horticulture. This analogy is the main idea in *The Origin of Species*. The situation for all organisms is like breeding, except *first*, the selective force is not a human breeder, but the complex of natural forces that determines which organisms will survive to reproduce, and *second*, the period of time over which natural selection has acted far exceeds human history. Since domestication has given us animals fitting our various needs to such an extent, it is no surprise we see even more exquisite adaptations in the fit between organisms in the wild and their habitat, their "ecological niche." Darwin says (§14) "What limit can be put to this power, acting during long ages and rigidly scrutinizing the whole constitution, structure and habits of each creature – favoring the good and rejecting the bad? I can see no limit to this power, in slowly and beautifully adapting each form to the most complex relations of life."

Fossil records. In some popular accounts of evolution, it is suggested that fossils were Darwin's main evidence, but this is very far from the truth. The main evidence comes from the the 3 obvious axioms listed above, and the analogy with breeding and horticulture which shows how successive incremental changes can produce dramatic changes. The fossil evidence actually presents serious difficulties for Darwin's view that these mechanisms explain the enormous variation and adaptation that so impressed him, so he considers the problem at length. He suggests that missing intermediate forms and sudden appearances in the fossil record are plausibly attributed to the imperfection of the fossil record (§9), but nevertheless when fossils are present, adjacent strata tend to differ minimally, while differences between the organisms become larger as one considers strata that are far apart.

Distinct species on islands. Darwin notes that islands frequently have species peculiar to them. "Thus in the Galapagos Islands nearly every land-bird, but only two out of the

eleven marine birds, are peculiar; and it is obvious that marine birds could arrive at these islands more easily than land-birds." (§12) There are many correlations between geography and distribution of organisms: similar organisms tend to be geographically close to each other, even when the geographically close areas have very different climates. "In considering the distribution of organic beings over the face of the globe, the first great fact which strikes us is, that neither the similarity nor the dissimilarity of the inhabitants of various regions can be accounted for by their climatal and other physical conditions." (§11)

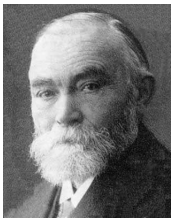
Vestigial organs. While the "perfection" of adaptation may result from selection over long periods of time, Darwin also observes that apparent "imperfections" may also have an explanation, as ancestral adaptations that are no longer used. "On the view of each organic being and each separate organ having been specially created, how utterly inexplicable it is that parts, like the teeth in the embryonic calf or like the shrivelled wings under the soldered wing-covers of some beetles, should thus so frequently bear the plain stamp of inutility! Nature may be said to have taken pains to reveal, by rudimentary organs and by homologous structures, her scheme of modification, which it seems that we wilfully will not understand." (§14) Gould calls this the "panda principle" after the peculiar panda's thumb, and sometimes the "orchid principle" because of the many contrivances Darwin noticed in the petals of orchids.

Homologous organs. The similarities among even very different species calls for some explanation: the explanation is that at least many of the similar organs are **homologous**, that is, they are inherited from common ancestors. Darwin says in §13, "What can be more curious than that the hand of a man, formed for grasping, that of a mole for digging, the leg of the horse, the paddle of the porpoise, and the wing of the bat should all be constructed on the same pattern, and should include the same bones, in the same relative positions."

Embryonic similarities. "The points of structure, in which the embryos of widely different animals of the same class resemble each other, often have no direct relation to their conditions of existence. We cannot, for instance, suppose that in the embryos of the vertebrata the peculiar loop-like course of the arteries near the branchial slits are related to similar conditions, in the young mammal which is nourished in the womb of its mother, in the egg of the bird which is hatched in a nest, and in the spawn of a frog under water…As the embryonic state of each species and group of species partially shows us the structure of their less modified ancient progenitors, we can clearly see why ancient and extinct forms of life should resemble the embryos of their descendants, our existing species" (§13)

Many puzzles and problems arise for the theory of natural selection. Some are recognized by Darwin and frankly discussed in his work. Others became clear only later. We will have more to say about all these things.

0.2 Language: First ideas



Frege

While language has always been of interest to people, the scientific study of language became much more feasible after we figured out how to precisely define logic, arithmetic, and other relatively simple symbolic systems. The definitions of these systems are "generative," allowing finite, precise definitions of infinite sets of expressions, as we will see later. One of the pioneering logicians was Gottlob Frege, best known now for his proposals about the meaning, the **semantics** of language. He was one of the first to be clear about the semantic value of quantifiers in language – words like *every, some, one, no*. Frege was also very interested in human language, and he made a number of important observations about it. One fundamental and simple idea is that human languages must allow a generative definition, and that this generative structure is part of the explanation of how it can be interpreted:

> It is astonishing what language can do. With a few syllables it can express an incalculable number of thoughts, so that even a thought grasped by a terrestrial being for the very first time can be put into a form of words which will be understood by someone to whom the thought is entirely new. This would be impossible, were we not able to distinguish parts in the thought corresponding to the parts of a sentence, so that the structure of the sentence serves as an image of the structure of the thought. (Frege, 1923)

The basic insight here is that the meanings of the limitless number of sentences of a productive language can be finitely specified, if the meanings of longer sentences are composed in regular ways from the meanings of their parts. We call this:

Semantic Compositionality: New sentences are understood by recognizing the meanings of their basic parts and how they are combined.

This is where the emphasis on basic units comes from: we are assuming that the reason you understand a sentence is **not** usually that you have heard it and figured it out before. Rather, you understand the sentence because you know the meanings of some basic parts, and you understand the significance of combining those parts in various ways. Frege observes sentences can have other sentences as parts. For example, sentence (3) has (2) and (1) as parts

(1) $(\frac{21}{20})^{100}$ is less than $\sqrt[10]{10^{21}}$

(2) $(\frac{21}{20})^{100}$ is greater than $\sqrt[10]{10^{21}}$

(3) $(\frac{21}{20})^{100}$ is less than $\sqrt[10]{10^{21}}$ and $(\frac{21}{20})^{100}$ is greater than $\sqrt[10]{10^{21}}$

Frege says that sentences also come with the "judgement or assertion" that each is true. (Frege chose these examples to illustrate that even if you do not know whether (2) or (1) is true (because you haven't thought about what numbers the arithmetic expressions denote), you can still know that (3) is false (because nothing is both greater and less than another thing.)

Given rigorous compositional accounts of simple mathematical languages, it did not take much longer to discover how a physical object could be designed to behave according to the formal rules of such a language – this is the idea of a computer. So by 1936, the mathematician Alan Turing showed how a finite machine could (barring memory limitations and untimely breakdowns) compute essentially anything (any "computable function"). In the short span of 70 or 80 years, these ideas not only spawned the computer revolution, but also revolutionized our whole conception of mathematics and many sciences. Linguistics is one of the sciences that has been profoundly influenced by these ideas: we recognize language structure by computing it, deriving it from our knowledge of the grammar of the language.

As we will see, a human language has some basic units, together with some ways for putting these units together. This system of parts and modes of combinations is called the **grammar** of the language. With a finite grammar, finite beings like humans can handle a language that is essentially unlimited, producing any number of new sentences that will be comprehensible to others who have a relevantly similar grammar. We accordingly regard the grammar as a **cognitive structure.** It is the system you use to "decode" the language.

In fact, human languages seem to require compositional analysis at a number of levels: speech sounds are composed from basic articulatory features; syllables from sounds; morphemes from syllables; words from morphemes; phrases from words. The semantic compositionality is perhaps the most intriguing, though. It is no surprise that it captured the imaginations of philosophers early in this century (especially Gottlob Frege, Bertrand Russell, Ludwig Wittgenstein). In effect, a sentence is regarded as an abstract kind of picture of reality, with the parts of the sentence meaning, or referring to, parts of the world. We communicate by passing these pictures among ourselves. This perspective was briefly rejected by radically behaviorist approaches to language in the 1950's, but it is back again in a more sophisticated form.

Another idea about the compositional structure of language is noted by Frege. He observes that certain parts of sentences require another to be "completed." For example, negation makes no sense by itself. The following sentences are fine:

(4) <u>It is not the case that</u> the cat likes dogs

(5) The cat likes dogs

But if we keep just the underlined part, the result is "incomplete," and would not usually be said by itself.

(6) * It is not the case that

We use the asterisk to indicate that there is something deviant about having the words *It is not the case that* by themselves. This string of words is incomplete until it is combined with *the cat likes dogs* or some other sentence. The same goes for *and, or, if…then*: these do not occur by themselves.

When we look into the structure of *the cat likes dogs*, we find a similar thing. The **subject** of the sentence is *the cat* and the **predicate** is *likes dogs*. The predicate seems to be "incomplete" in the same way as *it is not the case that*. The predicate *likes dogs* requires a subject to be

complete. We say that this predicate **selects** the subject. Going one step further, we can see that the **verb** *likes* selects a direct object too, since *the cat likes* is also incomplete.

In these first simple proposals, there are two very important claims:

(7) The structures of sentences are **recursive**, in the sense that inside a sentence, other sentences can "recur." This means that there is no longest sentence, and the language is infinite. Given any declarative sentence, you can make a longer one by adding *and* and another sentence.

(8) Certain parts of a sentence require other parts to be present.
In the common technical jargon: Certain parts **select** other parts.

These simple ideas will be important later.

0.3 Cognition: First ideas

Like the study of language, the study of things like learning, reasoning, and perception has taken a new shape with the advent of generative and computational models. Finally, these models can be clear and predictive, and even mathematical, where prior work had to be informal and more heavily judgement-laden. Still, these objects of study are very complex, and so it has been difficult to pin them down with the kind of generality and specificity that would be most useful in comparing the abilities of different organisms.

When comparing human cognition with non-human animal cognition, we face the great difficulty that we cannot explore what's going on by asking the animals about it. And of course we need to guard against attributing our own cognitive abilities to organisms that display similar abilities. A funny example of this is mentioned by (Pinker, 2002, p61), based on work by Laura Petitto, a psychologist who trained and studied a well-known chimpanzee named "Nim Chimpsky." She actually lived with Nim for a year in a large house in New York state, on a Columbia University research project. Nim seemed to imitate many things that he saw Petitto doing, but not in the way a human child would. For example, seeing Petitto washing the dishes, Nim would imitate her motions and enjoy the warm water. Looking more closely though, it turned out that he had no idea of what the activity was for. He would mimic her motions, rubbing dishes with a sponge, but he never got the idea that the object was to make the dish cleaner.

When we are interested in evolutionary connections between behaviors, there is another kind of confusion that we should guard against: similar behaviors in different organisms can sometimes derive from a common ancestor – in this case they are called "homologies" – but they can also arise independently – in which case they are called "homoplasies." The inference from mere similarity to an evolutionary connection is not generally a safe one.

Consider, for example, what goes on when you walk across the street: you visually perceive various familiar objects in spatial relationships of various kinds, and you coordinate the motions of your muscles in order to move your body. Since animals can navigate their environments, do they have similar and even homologous abilities to perceive and reason about

objects in the three dimensional world? When you irritate the skin on a frog's leg (a place that the frog cannot see), the frog will coordinate its muscles in a motion that results in scratching that part of its body (Fukson, Berkinblit, and Feldman, 1980). Given the changes in body positions and the restricted movements of the joints, it is easy to see that the frog is solving a very complex problem about three dimensional space in this task. When a wasp finds its back way to its nest after foraging, or when a rat returns to a place where food was found earlier, these are also complex tasks in three dimensional space. Do these animals have some conception of three-dimensional space similar to ours, or are they just using remembered landmarks to get around? This question can be studied. For example, you can take a rat that has learned how to navigate a maze to find food, and then systematically distort the shape and features of the maze to see how it affects the rat's navigation. It turns out that rats do attend to spatial geometry rather than navigating simply with landmarks (Cheng, 1986; Gallistel, 1990).

0.4 Some basic questions: Local and global properties

Darwin's claim is that natural selection is one of the influences on the development of organisms. The last sentence of the introduction to *Origins* says "I am convinced that natural selection has been the main, but not the exclusive means of modification."

There are many sources of order in biological forms. As we will see, natural selection, the survival of the fittest, provides a force external to the organism that can shape the changes in organisms over generations. But biological materials and biological complexes have properties that no force of evolution can change. In complex systems, some properties that emerge may be due to basic properties of their parts, rather than to some external shaping influence on the complex as a whole. With respect to these properties, the complexes are said to be *self-organizing.* Untangling the relative contributions of internal and external influences is one of the main challenges in evolutionary biology, and we will see that there are similar challenges in untangling the forces shaping human language. It is useful to reflect on some simple cases first, in order to raise more clearly the questions about what is going on in the complex cases.

0.4.1 Local and global properties in wave propagation: emergent physical regularities

When you throw a stone into a still pond, the ripples spread in almost perfect circles. This familiar fact has a kind of implication for physics that we are interested in here. The regularity of the ripples must somehow *emerge* from the local interactions among water molecules. The "global regularities," the circular ripples, emerge from strictly "local events." When you intentionally draw a circle, you execute a plan based on your conception of the intended result, but obviously no such thing is going on in the disturbed surface of a liquid. There is no plan, no external force creating the circular ripples or the spherical droplets in a splash. Rather, these emerge from the regular changes in forces among the water molecules and the surface tension. The forces around each water molecule are "local", but they have the "global," "emergent" effect of beautiful, expanding circles.

Similar puzzles about how local intrinsic properties of materials lead to global regularities come to mind when you contemplate the architecture of crystals and snowflakes, the regular shapes of sand dunes, the spectra of various light sources.

Harold Edgerton (1957) "Milk Drop Coronet"

0.4.2 Local and global properties in protein assembly: emergent organic regularities

We see similar emergent global regularities in much more complex phenomena. The structure of proteins provides a relevant example, since, as we will discuss in more detail later, heredity is controlled by DNA and RNA determination of protein synthesis.

All living things, from bacteria to humans, contain proteins built from the 20 naturally occuring amino acids, listed here with their (3 letter and 1 letter) standard abbreviations.

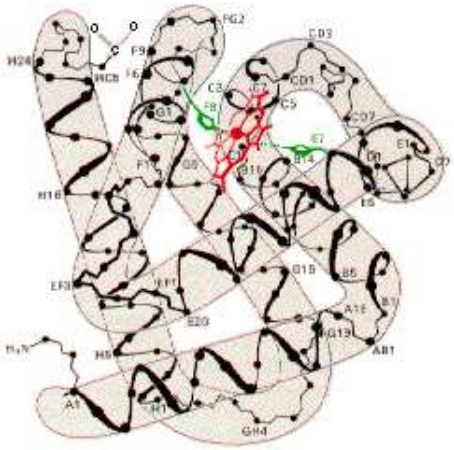
amino acid	abbreviations	amino acid	abbreviations
Alanine	Ala A	Cysteine	Cys C
Aspartic AciD	Asp D	Glutamic Acid	Glu E
Phenylalanine	Phe F	Glycine	Gly G
Histidine	His H	Isoleucine	Ile I
Lysine	Lys K	Leucine	Leu L
Methionine	Met M	AsparagiNe	Asn N
Proline	Pro P	Glutamine	Gln Q
pARginine	Arg R	Serine	Ser S
Threonine	Thr T	Valine	Val V
Tryptophan	Trp W	TYrosine	Tyr Y

These amino acids combine in sometimes very long sequences. Sequences of length less than 40 or so are often called **peptides**. Longer sequences are **proteins** or **polypeptides**. These proteins control most important cellular processes. For example, **hemoglobin**, the protein that carries oxygen in the blood, and which also appears in the cells of plants and even bacteria, has a distinctive sequence of amino acids:

NH_2—Val His Leu Thr Pro Glu Glu Lys Ser Ala Val Thr Ala

Leu Trp Gly Lys Val Asn Val Asp Glu Val Gly Gly Glu--

The long polypeptide chains twist, coil and fold up in complex ways, yielding shapes that are typically important determinants of their function. The shape of hemoglobin is thought to be something like this:

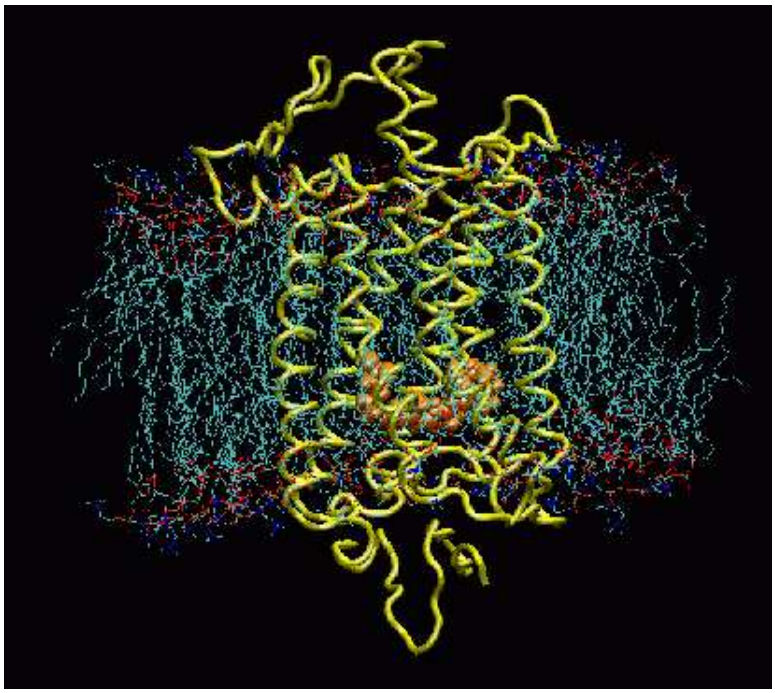


Another well-known protein is **rhodopsin**, found in light-sensitive cells in the eyes of vertebrates and invertebrates, and variants of this protein are even found in algae and in some light-sensitive bacteria. When rhodopsin is exposed to light, the surface of the protein becomes catalytically active, that is, capable of triggering reactions that change ion distributions

in the cell, in a way that affects its interactions with other cells: inducing synaptic potentials in the optic nerves of higher organisms. Bovine rhodopsin has the amino acid sequence:

MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIP
EGMQCSCGIDYYTPHEETNNESFVIYMFVVHFIIPLIVIFFCYGQLVFTVKEAAAQQQES
ATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSAV
YNPVIYIMMNKQFRNCMVTTLCCGKNPLGDDEASTTVSKTETSQVAPA

Electron microscopy and crystallographic studies suggest that mammalian rhodopsin has a 3D structure that is something like this (here the twisting molecule is shown in net-like membrane):



NIH Resource for Macromolecular Modeling and Bioinformatics

The shapes of these proteins are important for their functions, but how is shape determined? DNA and RNA provide external control on the amino acid sequence, but nothing "external" to the molecule plans its ornate 3D configuration. Instead, this must be determined by particular bonds and attractions between individual parts of the molecule. A great deal of effort has gone into mathematical models of how this works, and this research is ongoing today.

0.4.3 Local and global properties in phyllotaxis: emergent organismic regularities

Stepping up to the level of whole organisms, we again find regularities that must emerge from more basic and local self-organizing forces. For example, Turing (best known for his work on

computing, mentioned above) was intrigued by the number of petals on daisies (Turing, 1952).


The leftmost picture shows a shasta daisy, which has 21 petals. Ordinary field daisies have 34 petals, You can find daisies with 21, 34, 55 or even 89 petals, but not 4 or 8 or 25 or 36 petals. As in the previous examples, there is a question about how the number of petals is determined. Does each plant cell "know" how many other petals there are? If not, how can any cell know whether to initiate the development of a petal itself? But the daisies provide an extra clue to how this must work, with an extra puzzle: why these particular numbers of petals?

The numbers of daisy petals listed above are all "Fibonacci" numbers. The Fibonacci numbers are defined this way

$$\begin{aligned} fib(0) &= 1 \\ fib(1) &= 1 \\ fib(n) &= fib(n-1) + fib(n-2) \quad \text{for all } n > 1 \end{aligned}$$

Beginning with the third one, each Fibonacci number is the sum of the previous two. This definition is said to be **recursive**, since the definition of a Fibonacci number uses the notion of Fibonacci numbers. The definition is recursive but **not circular**, because the definition of each number depends only on earlier values. So we can calculate the values of fib for n=0,1,2,3,4,...to get the following numbers:

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, \ldots$$

Why are the numbers of petals on a daisy Fibonacci numbers?

Lilies, irises, and the trillium have three petals; columbines, buttercups, larkspur, and wild rose have five petals; delphiniums, bloodroot, and cosmos have eight petals; corn marigolds have 13 petals; asters have 21 petals; and daisies have 21, 34, 55, or 89 petals – all Fibonacci numbers. Flowers with other numbers of petals can be found, but the Fibonacci numbers are wierdly common.

see, e.g. http://ccins.camosun.bc.ca/ jbritton/fibslide/ for more examples

The number of spirals out from the center of a sunflower, a pinecone, a pineapple, or a cauliflower is usually a Fibonacci number. And in some plants with stems leaving a branch at various intervals, the rotation around the stem from one branch to the next is given by a

ratio of successive Fibonacci numbers. **Phyllotaxis** is the name for such patterns in the leaves or petals or branching patterns of plants. These regularities are clearly not imposed by some external controlling force. One recent proposal about them appears in (Douady and Couder, 1996). (a check on the web will reveal many others too!) Clearly something about the numbers of petals is genetically determined, and this may be influenced by natural selection, but there is some other force acting as well. The matter is still not well understood.

0.4.4 Local and global properties in axes formation: emergent organismic regularities

In the *metazoa*, the kingdom of multicelluar animals, we have similar puzzles. One that was noticed by Turing and has been recently studied by (Meinhardt and Gierer, 2000) and others, concerns the development of hydra, a freshwater organism with about 100,000 cells, about 1-2 millimeters long. This animal has a definite structure with a "head" and a "foot," and nerve cells for coordinating motions. The strange thing about these animals is that they can reproduce not only sexually, but also by "budding." (The figure here shows a bud forming.) In fact, a full hydra can be "regenerated" from even a small piece of any hydra, so we have to conclude that each cell, in some sense, has the complete body plan encoded in it. If this is true, how does each cell know which part of the body plan that it should realize? For example, a hydra has an axis of approximate symmetry extending along its length. How is this axis established in the development of a hydra from a body part?

Similar puzzles arise in more complex organisms, which typically have several axes of approximate symmetry that originate early in embryonic development. Vertebrates like us have at least 4 axes of symmetry in early embryonic development. Studies suggest that in all these organisms small initial asymmetries due to gravity, the exact point of sperm entry, and other things, get "amplified" somehow to form axes of symmetry for the developing embryo. This requires some kind of communication between cells, via some kind of diffusible substances. The basic idea is that cells developing in certain ways emit chemicals that inhibit certain kinds of changes in other cells. An existing hydra head inhibits the development of another one, and in this case the chemical basis of the signal is fairly well understood.

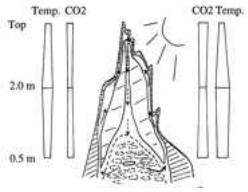
Again, the thing we see in these cases is an overall plan getting realized by local properties of the parts. In this case, the overall plan is rather complex, and some kind of "communication" among the parts is necessary. Clearly this is genetically determined, in part, may be influenced by natural selection, but the collaboration of physical determinants of axis formation is a complex matter, in which each cell relies on a complex of interaction of local stimuli.[2]

[2] There is a flurry of discussion about cell differentiation in vertebrates going on just recently (Dudley, Ros, and Tabin, 2002; Sun, Mariani, and Martin, 2002; Tickle and Wolpert, 2002).

0.4.5 Local and global properties in termite nests: emergent supra-organismic regularities

It is no surprise that we find emergent phenomena even above the level of the organism. One example that has been quite extensively studied is the structure of ant and termite nests. Termite nests in Africa are sometimes very large. The nests of *Macrotermes Bellicosus*, found in tropical Africa, can be 6 meters high and 30 meters across, with a main structure of sand and dried saliva. They are carefully structured, typically with a large central chamber and with a number of small passageways along the outer walls. The central areas have columns in which young termites are raised, and the queen stays in a small very sturdy, central room in the nest. Other chambers hold "fungus gardens" that grow in finely chewed wood. The overall structure provides temperature regulation and circulation to keep the level of CO_2 at reasonable levels in the heat of the savannah.

How does this structure get built? Surely no single ant directs the construction, and no single ant has the overall design in mind. In some sense, the structure must emerge from the behavior of each ant doing what seems natural to it. Studies show that the nest begins with a small underground chamber occupied by the "royal couple," then offspring of the couple construct surrounding chambers of chewed wood for cultivation of fungus, and finally the large superstructure starts to develop. Ants are already complex enough that we do not understand how "instincts" (genetically influenced and possibly naturally selected) and environment interact to yield such amazing structures.



(Korb and Linsenmair, 2000)

The construction of a particular human language like English may be like this last example in some important respects: in some sense, your English is represented in your mind, but no one planned its structure. And English as a cultural artefact takes its shape in part because of specific properties of the individual speakers (the language is self-organizing), but this happens over many generations.

0.5 Summary and poetry

This first section introduces a number of basic ideas that we will explore more carefully later. We introduced the basic axioms of **Darwin's theory of natural selection**; we briefly reviewed Darwin's **evidence for the theory**; we introduced Frege's idea that **human languages are compositional and recursive, with elements that select each other**; and finally we used 5 examples (waves, protein, phyllotaxis, axis formation, and termite mounds) to observe that in complex systems, regular properties can emerge which must be due to some **interaction between "local" properties of the parts and the "global" environment**. We expect to find this happening in the evolution of organisms, in the evolution of human language ability, and in the evolution of

languages. In all these cases, there will be other determinants of the complexes (the organisms, the cognitive abilities, the cultural artifacts) besides the external one of natural selection. As Darwin emphasizes, natural selection can only be part of the story.

The last paragraph of Darwin's *Origin of Species* poetically summarizes his vision of how we came to be:

> It is interesting to contemplate an entangled bank, clothed with many plants of many kinds, with birds singing on the bushes, with various insects flitting about, and with worms crawling through the damp earth, and to reflect that these elaborately constructed forms, so different from each other, and dependent on each other in so complex a manner, have all been produced by laws acting around us. These laws, taken in the largest sense, being Growth with Reproduction; inheritance which is almost implied by reproduction; Variability from the indirect and direct action of the external conditions of life, and from use and disuse; a Ratio of Increase so high as to lead to a Struggle for Life, and as a consequence to Natural Selection, entailing Divergence of Character and the Extinction of less-improved forms. Thus, from the war of nature, from famine and death, the most exalted object which we are capable of conceiving, namely, the production of the higher animals, directly follows. There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.

Has Darwin missed anything important? The puzzles about local and global effects in the previous section draw our attention to sources of order that do not seem to be on Darwin's list, sources that will act to limit the range of variations found in nature. We will say much more about these later, and we will look for similar factors shaping human languages.

Stuart Kauffman poetically expresses his optimism about theories that embrace both natural selection and self-organization in this passage, written 141 years after Darwin's:

> Whence the order out my window? Self-organization and selection, I think. We the expected and we the ad-hoc. We the children of ultimate law. We the children of the filigrees of historical accident.
>
> What is the weave? No one yet knows. But the tapestry of life is richer than we have yet imagined. It is a tapestry with threads of accidental gold, mined quixotically by the random whimsy of quantum events acting on bits of nucleotides and crafted by selection sifting. But the tapestry has an overall design, an architecture, a woven cadence and rhythm that reflect underlying law – principles of self-organization.
>
> How are we to begin to understand this new union? For "begin to understand" is all we can now hope for. We enter new territory . It would be presumptuous to suppose that we would understand a new continent when first alighting on its nearest shores. We are seeking a new conceptual framework that does not yet exist. Nowhere in science have we an adequate way to state and study the interleaving of self-organization,

selection, chance, and design. we have no adequate framework for the place of law in a historical science and the place of history in a lawful science.

But we are beginning to pick out themes , strands in the tapestry. The first theme is self-organization. Whether we confront lipids spontaneously forming a bilipid membrane vesicle, a virus self-assembling to a low-energy state, the Fibonacci series of a pinecone's phyllotaxis, the emergent order of parallel processing networks of genes in the ordered regime, the origin of life as a phase transition in chemical reaction systems, the supracritical behavior of the biosphere, or the patterns of coevolution at higher levels-ecosystems, economic systems, even cultural systems- we have found the signature of law. All these phenomena give signs of non-mysterious but emergent order. We begin to believe in this new strand, to sense its power. (Kauffman, 1995, pp185-186)

Lecture 1

Genetic variation, transmission, selection

…we must recognize and embrace natural history as a science of relative frequencies.
– (Gould, 2002, p147)

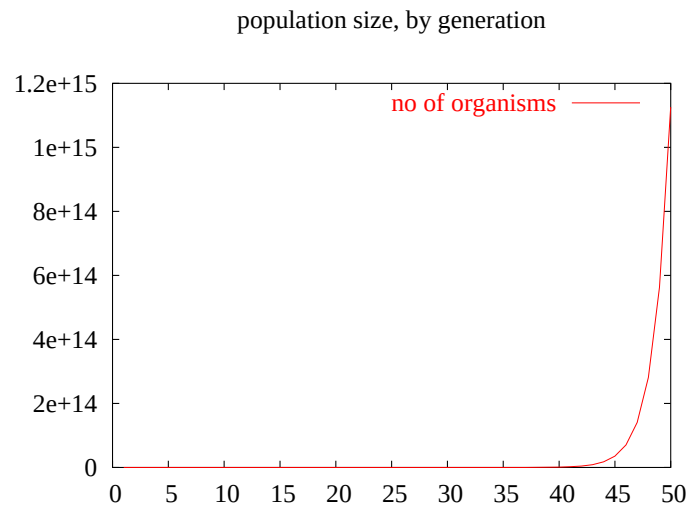
1.1 The geometric increase in populations


Malthus

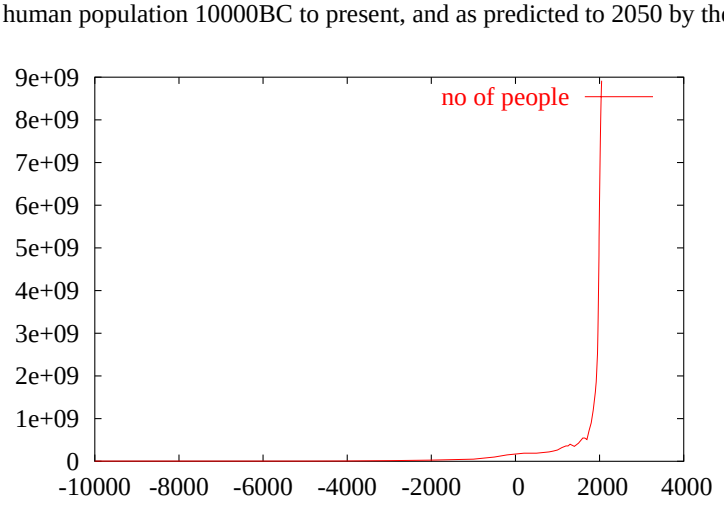
Darwin's third axiom, the hypothesis of natural selection, was motivated in part by his reading of Thomas Malthus's work on populations. Malthus pointed out that poverty and famine would be the result of unbridled population growth, an idea which is obviously true but still not so obvious that it could not be misunderstood and misused by politicians. Notice that if each organism produces two offspring, and each of those produce two offspring, and so on, the population size will increase **geometrically**. The population sizes form a **geometric series**, increasing by a factor of n every generation, where n is the number of offspring that survive to reproduce. The following table shows the increase when each organism produces 2 offspring:

generation number	number of individuals
0	$1 = 2^0$
1	$2 = 2^1$
2	$4 = 2^2$
3	$8 = 2^3$
…	…
10	$1,024 = 2^{10}$
20	$1,048,576 = 2^{20}$
…	…
100	$1,267,650,600,228,229,401,496,703,205,376 = 2^{100}$
…	…

Drawn in a graph, this is an exponential curve: the rate of increase is constantly increasing (remember: e+14 means $\times 10^{14}$, that is: move the decimal point 14 places to the right):

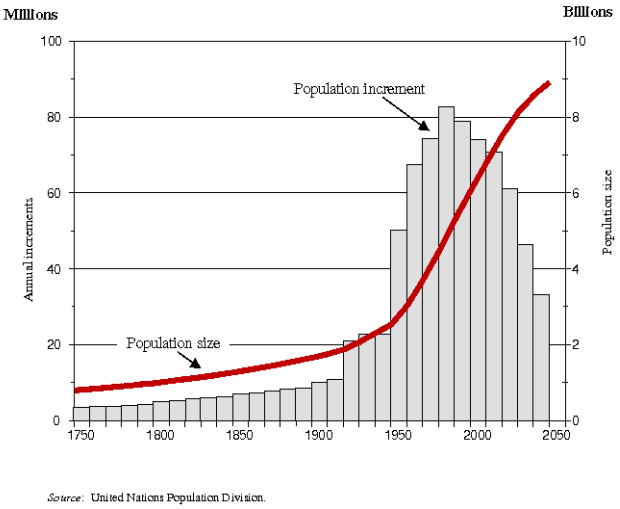


Applied to populations, this looks quite serious. Consider fruit flies (*Drosophila*), for example. They live only several weeks, and the time between generations is less than 2 weeks (depending on temperature), instead of the roughly 20 years between human generations. A fruit fly will produce about 500 eggs. If they all survived and reproduced, then after 100 generations – less than 4 years – there would be there would be 500^{100} of them. Since there are only about 10^{57} electrons in the sun, 500^{100} fruit flies would require vastly more matter than there is in the whole solar system. The easy conclusion: no matter what, most of them cannot survive, and all of us other organisms are in the same boat. Remarkably, with just a few small glitches, the human population has been increasing roughly exponentially throughout our history:[1]



[1] Sources: up to 1950, McEvedy and Jones (1978); 1950-2050, United Nations Secretariat (2001).

But it is obvious that the exponential increase can continue for only a short time. At some point, various factors will intervene to keep the population size from increasing so rapidly. The census figures collected by the UN show that population growth has slowed dramatically in developed countries, and they are projecting that the growth will level off, as we see in this graph spanning just a few hundred years:



Many other species are facing extinction. In any case, the competition among organisms that inevitably results from natural increases in population was one of the fundamental ideas behind Darwin's third axiom, natural selection. To reason about effects of selection, we need probabilities!

1.2 Two basic laws of probability

We say two events e_1 and e_2 are independent if the probability of one has no influence on the probability of the other. In this case, the probability that both e_1 and e_2 happen is the product of their separate probabilities:

$$p(e_1 \text{ and } e_2) = p(e_1)p(e_2).$$

This is the **product rule** for **independent events**. So for example, with a fair coin, the probability of getting a head on each toss is $\frac{1}{2}$, and each toss is independent, so the probability of two heads in a row is

$$\begin{aligned} p(\text{heads on toss 1 and heads on toss 2}) &= p(\text{heads on toss 1})p(\text{heads on toss 2}) \\ &= \tfrac{1}{2}\,\tfrac{1}{2} \\ &= \tfrac{1}{4} \end{aligned}$$

Obviously, this is the same a the chances of getting heads on the first toss and tails on the second:

$$\begin{aligned} p(\text{heads on toss 1 and tails on toss 2}) &= p(\text{heads on toss 1})p(\text{tails on toss 2}) \\ &= \tfrac{1}{2}\ \tfrac{1}{2} \\ &= \tfrac{1}{4} \end{aligned}$$

Since the probability of getting heads is $\frac{1}{2}$, and the probability of getting tails is $\frac{1}{2}$ with a fair coin, the probability of getting heads or tails is $\frac{1}{2} + \frac{1}{2} = 1$. Notice that these two events are dependent on each other: the events are **disjoint** in the sense that if you get heads on a toss, you cannot get tails. In general, if e_1 and e_2 are two of the possible outcomes of some event, (and no case of e_1 can be a case of e_2), then

$$p(e_1 \text{ or } e_2) = p(e_1) + p(e_2).$$

This is sometimes called the **sum rule** for **disjoint events**.

What is the probability of getting heads on toss 1 or tails on toss 2 (or both)? We can use the product rule and the sum rule to figure this out, but it takes a little calculation to do it. Notice that these two events are independent, but we are not asking for the probability of that both occur, but the probability that at least one of them occurs. And the events are not disjoint: in a sequence of two roles, both could occur.

So the way we calculate this is to consider all the possible outcomes of two roles. Let h_1 mean heads on the first toss, and t_1 mean tails on the first toss, and similarly for the second toss, h_2 and t_2. Then there are four possible outcomes of two tosses, and since each toss is independent we can use the product rule to calculate the probability of each outcome:

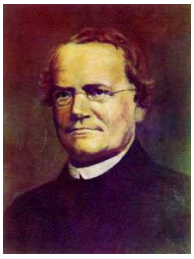
toss 1	toss 2	probability=$p_1 p_2$	case
h_1	h_2	$\frac{1}{2}\ \frac{1}{2} = \frac{1}{4}$	i
h_1	t_2	$\frac{1}{2}\ \frac{1}{2} = \frac{1}{4}$	ii
t_1	h_2	$\frac{1}{2}\ \frac{1}{2} = \frac{1}{4}$	iii
t_1	t_2	$\frac{1}{2}\ \frac{1}{2} = \frac{1}{4}$	iv

Now, we can see that each of these 4 cases is disjoint from the others: if one case happens, the others cannot happen. And since the question is, what is the probability of getting heads on toss 1 or tails on toss 2 (or both), we are asking the probability of case i or ii or iv. Only case iii is a loser. So we calculate

$$\begin{aligned} \text{p(heads on toss 1 or tails on toss 2)} &= (\text{sum cases i, ii, iv}) \\ &= \tfrac{1}{4} + \tfrac{1}{4} + \tfrac{1}{4} \\ &= \tfrac{3}{4} \end{aligned}$$

How can you tell if a coin is fair? You can tell by taking a large enough sample of flips. If you flip the coin repeatedly (and the coin does not wear out), in the limit, the **relative frequency** of heads, that is, the number of heads divided by the total number of flips, will approach $\frac{1}{2}$ if the coin is fair. This idea is called the **law of large numbers**. So we use probability theory to reason about relative frequencies.

1.3 Genetic atoms do not blend



Mendel

The Swiss monk Gregor Mendel describes 7 years of careful pea cultivation in a 1865 report entitled "Experiments with Plant Hybrids." He read this report to a conference on natural history and mailed it to various people, including the famous Swiss botanist Karl von Nägeli. He even sent Nägeli more than 100 seed packets with instructions for replicating the experiments, but Nägeli dismissed the work and returned it to Mendel. (There is speculation that Nägeli was put off by the math.) Mendel's work was not rediscovered until around 1900, when it was realized that it provided an important correction to Darwin's ideas about how the traits inherited by sexual reproduction combine. His work introduces the notion of hereditary atoms, "genes," that are discrete and do not "blend."

Mendel's famous contribution is quite simple. The first experiment he reports is a study of the seeds produced by his pea plants: some seeds are "round or roundish" while other plants have "angular wrinkled" seeds. Two "purebred" or "homozygotic" round peas will always produce a round pea; and a purebred wrinkled pea will always produce a wrinkled pea. But when Mendel cross-pollinated these two kinds of plants to get **hybrids**, he found that the resulting plants did not produce seeds with a blending of the parent characters, but that all the new peas of these hybrid plants were smooth and round. How to explain this?

A clue comes from self-pollinating the hybrids to see what happens in the first generation of hybrid offspring. Mendel discovered that from the hybrid parents, almost exactly $\frac{3}{4}$ of the first generation of offspring were smooth and round, and $\frac{1}{4}$ were wrinkled. Taking self-pollinated seeds from all those plants to produce another generation, he found the proportions $\frac{5}{8}$ and $\frac{3}{8}$ respectively in the second generation. Continuing in this way:

generation	# round	# angular	total # of plants
0	2	0	2
1	3	1	4
2	5	3	8
3	9	7	16
4	17	15	32
5	33	31	64
generation n	$2^n + 1$	$2^n - 1$	2^{n+1}

Here we're imagining that Mendel is keeping all these plants alive, so the total number of plants is growing geometrically, increasing by a factor of 2 from each generation to the next.

The proportion of round vs. angular seeds is also changing! In the 1st generation, there are 3 times as many round seeds as angular ones, but in the 5th generation, the numbers are getting close together. We can see this by plotting the proportions instead of the total numbers, this way:

generation	proportion round/total	proportion angular/total
0	$\frac{1}{1}$	0
1	$\frac{3}{4}$	$\frac{1}{4}$
2	$\frac{5}{8}$	$\frac{3}{8}$
3	$\frac{9}{16}$	$\frac{7}{16}$
4	$\frac{17}{32}$	$\frac{15}{32}$
5	$\frac{33}{64}$	$\frac{31}{64}$

How to explain this? Mendel's explanation is simple and ingenious, and it accounts not only for the proportions of smooth and wrinkled peas in this first generation but for later generations. He proposed that each plant has a pair of genes each corresponding to a "constant character," where the gene for round seeds is **dominant** and the gene for wrinkled seeds is **recessive**:

A: the dominant gene for round smooth seeds
a: the recessive gene for angular wrinkled seeds

In modern terms, we say that A and a are **alleles**, alternative forms of the gene for the smooth/angular trait. Since each pea plant has two alleles, there are 3 kinds of plants: AA, Aa, or aa. But since A (round smooth) is dominant, both the AA and Aa plants will produce A round seeds. This is the origin of the distinction between the **genotype**, the genetic endowment, and the **phenotype**, the resulting physical properties of the organism itself. Here there are 3 genotypes (AA, Aa, aa), but just two phenotypes (round or angular seed production).

When we cross AA with aa parents, where each parent provides one of its genes (at random) to the offspring, assuming each parent contributes one gene, we predict that the next generation will all be Aa, hybrids. And because A (round smooth) is dominant, this generation of hybrids will have only round smooth seeds.

What can happen when we cross 2 hybrid Aa parents? If we watch who contributes which gene, there are 4 cases i-iv:

$parent_1$	$parent_2$		result	case
a_1A_1	a_2A_2	$\Rightarrow$	A_1A_2	i
a_1A_1	a_2A_2	$\Rightarrow$	a_1A_2	ii
a_1A_1	a_2A_2	$\Rightarrow$	a_2A_1	iii
a_1A_1	a_2A_2	$\Rightarrow$	a_1a_2	iv

Notice that two of these cases, ii and iii produce aA offspring, so if each of the four cases is equally likely, then on average we expect a 1:2:1 ratio among these genotypes. And since A is dominant, we find a 3:1 ratio among the phenotypes, which is what Mendel observed.

We can put this first step in a slightly more general form. Assuming that the probability of getting each gene is equally likely, then the probability of getting AA is the probability of getting A_2 from the seed and A_1 from the pollen, that is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. And aa has the same probability. The probability of a_1A_2 and a_2A_1 are each $\frac{1}{4}$ too, and since we have aA in both of these cases, the probability of aA is $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

That is, the probability of each result in this simple situation is calculated as follows. The probability that parent$_1$=a_1A_1=1, the probability that parent$_2$=a_2A_2=1, and the probability that these combine to yield A_1A_2=$\frac{1}{4}$, so the probability of the result A_1A_2 is the product of these 3 independent events $1 \times 1 \times \frac{1}{4} = \frac{1}{4}$. We can fill out our table of possibilities accordingly:

parent$_1$(p_1)	parent$_2$(p_2)	(p_3)	result	probability(result)=$p_1p_2p_3$	case
a_1A_1(1)	a_2A_2(1)	$\Rightarrow$ ($\frac{1}{4}$)	A_1A_2	$\frac{1}{4}$	i
a_1A_1(1)	a_2A_2(1)	$\Rightarrow$ ($\frac{1}{4}$)	a_1A_2	$\frac{1}{4}$	ii
a_1A_1(1)	a_2A_2(1)	$\Rightarrow$ ($\frac{1}{4}$)	a_2A_1	$\frac{1}{4}$	iii
a_1A_1(1)	a_2A_2(1)	$\Rightarrow$ ($\frac{1}{4}$)	a_1a_2	$\frac{1}{4}$	iv

Notice that the sum of the probabilities of cases i-iv is 1, as required. Furthermore, we can calculate that

$$\begin{aligned} \text{p(AA)} &= (\text{case i}) = \tfrac{1}{4} \\ \text{p(aA)} &= (\text{sum cases ii, iii}) = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2} \\ \text{p(aa)} &= (\text{case iv}) = \tfrac{1}{4} \end{aligned}$$

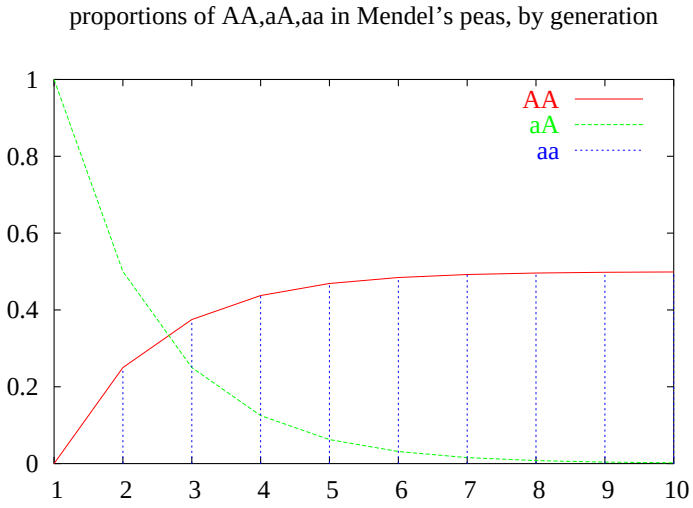
Now suppose we take all the hybrids and self-pollinate them. Then all the AA plants will produce AA offspring, all the aa plants will produce aa offspring, and all the aA plants will produce offspring in the proportion we found in the hybrids themselves. So if there are x AA's, y Aa's and z aa's, where $x + y + z = 1$, then in the next generation there will be $x' = x + \frac{y}{4}$ AA's, $y' = y - \frac{y}{2}$ aA's, and $z' = z + \frac{y}{4}$ aa's, where $x' + y' + z' = 1$. We can specify all three calculations at once, like this:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \Rightarrow \begin{pmatrix} x + \frac{y}{4} \\ y - \frac{y}{2} \\ z + \frac{y}{4} \end{pmatrix}$$

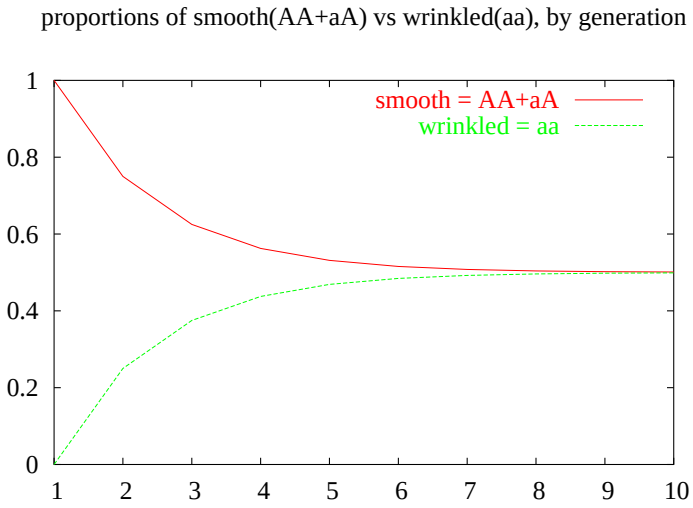
With this idea, the results for the genotypes are these:

$$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \\ \frac{1}{4} \end{pmatrix} \Rightarrow \begin{pmatrix} \frac{3}{8} \\ \frac{1}{4} \\ \frac{3}{8} \end{pmatrix} \Rightarrow \begin{pmatrix} \frac{7}{16} \\ \frac{1}{8} \\ \frac{7}{16} \end{pmatrix} \Rightarrow \ldots$$

We can graph these numbers to see that the hybrids are going to decrease quickly to zero, because in every generation some of their offspring are purebred, homozygotic:



So we see that Mendel's method of self-pollination is a good way to get the pure AA and aa genotypes to dominate in the population. It took Mendel to get to this idea from careful observation of the phenotypes! If we watch the proportions of the phenotypes only, as Mendel did, what we get is this:

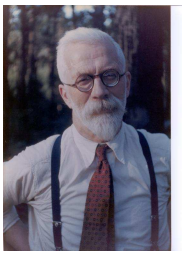


Now we understand why this happens.

The most important discoveries that Mendel made are these 3 very basic ones:

1. there are atoms of heredity that combine without blending;
2. these atoms can be dominant or recessive; and
3. pairs of these atoms in the genotype can determine the resulting phenotype.

(Later, in section §3.2, we take a look at some complications hiding behind these claims.)



Fisher

Of course, things get more complicated when we watch the inheritance of multiple traits, particularly because different traits sometimes "linked:" if you have one, you always or probably have the other. The appearance of blending can occur when many genes in combination are responsible for a trait – e.g. it can happen that offspring are intermediate in size between the parents, when size is a result of several genes: some combinations may yield intermediate results. This is the rule rather than the exception: most of the traits we usually think about are specified by many genes, and some of them can be linked to each other in various ways. Ronald Fisher was one of the researchers who studied the result of combining Mendelian genetics with Darwinian theory, showing how natural selection can progress in small changes that have a large cumulative effect.

Digression: Optional stuff for hackers

In case there are any hackers in the group, or "wanna-be" hackers, it is easy to draw the graphs shown in this section yourselves, and then you can tinker with the equations a little and really see how they work. This is strictly optional, but if you want to try it, I would be glad to help out if you get stuck.

I did the calculations using a free program called **octave**, and I plotted the graphs with **gnuplot** If you use Linux, these are standardly included in the distributions. If you use Mac OS X, you can get them using `http://fink.sourceforge.net/`, and if you use windows, you can get instructions at `http://octave.sourceforge.net/Octave_Windows.htm`. Once you have these programs, you can download my octave scripts. If you download my script `mendel.m` and then just start octave and type: `mendel` at the prompt, it should do the calculations and draw the graphs above. I will produce octave scripts for calculations coming later in the course too, and put them all on the class web page.

1.4 The Hardy-Weinberg equilibrium

Hardy

In Mendel's study, we saw that the relative frequencies of the genotypes AA, aA, and aa changed dramatically over the period of just a few generations, and this is shown in the graphs. But in that study, the plants were all self-pollinated. We did not look yet at what would happen under any other conditions. In a more natural setting, we might expect each plant to get pollen from a parent chosen according to the relative frequency of the different types of parents around. Godfrey Hardy and Wilhelm Weinberg independently noticed in the same year, 1908, that in this mathematically simple condition, the relative proportions of genotypes in a population will remain unchanged: it will go to a stable point, an equilibrium, and stay there.

The first step of Mendel's study was just to produce hybrids, and then all later generations were produced by self-pollination. Suppose instead that in the second generation, we had cross-fertilization with parents of all 3 kinds, AA, aA, and aa, in 1:2:1 proportions. We can divide the all the possibilities up into cases, as we did for Mendel's study above. (This is a little bit tedious, but a good exercise. If you are lazy, you can skip ahead to the shortcut that Hardy and Weinberg point out, at the end of this section.)

parents ⟨aA,aA⟩ The chances of an aA parent in this generation is $\frac{1}{2}$, so:

$\text{parent}_1(p_1)$	$\text{parent}_2(p_2)$	(p_3)	result	probability=$p_1p_2p_3$	case
$a_1A_1(\frac{1}{2})$	$a_2A_2(\frac{1}{2})$	$\Rightarrow(\frac{1}{4})$	A_1A_2	$\frac{1}{16}$	i
$a_1A_1(\frac{1}{2})$	$a_2A_2(\frac{1}{2})$	$\Rightarrow(\frac{1}{4})$	a_1A_2	$\frac{1}{16}$	ii
$a_1A_1(\frac{1}{2})$	$a_2A_2(\frac{1}{2})$	$\Rightarrow(\frac{1}{4})$	a_2A_1	$\frac{1}{16}$	iii
$a_1A_1(\frac{1}{2})$	$a_2A_2(\frac{1}{2})$	$\Rightarrow(\frac{1}{4})$	a_1a_2	$\frac{1}{16}$	iv

parents ⟨AA,AA⟩ This case is simple because there is only one possible outcome:

$\text{parent}_1(p_1)$	$\text{parent}_2(p_2)$	(p_3)	result	probability=$p_1p_2p_3$	case
$AA(\frac{1}{4})$	$AA(\frac{1}{4})$	$\Rightarrow(1)$	AA	$\frac{1}{16}$	v

parents ⟨aa,aa⟩ And this case is like the last one:

$\text{parent}_1(p_1)$	$\text{parent}_2(p_2)$	(p_3)	result	probability=$p_1p_2p_3$	case
$aa(\frac{1}{4})$	$aa(\frac{1}{4})$	$\Rightarrow(1)$	aa	$\frac{1}{16}$	vi

parents ⟨AA,aA⟩+⟨aA,AA⟩ This happens 2 ways:

$\text{parent}_1(p_1)$	$\text{parent}_2(p_2)$	(p_3)	result	probability=$2(p_1p_2p_3)$	case
$A_aA_b(\frac{1}{4})$	$a_2A_2(\frac{1}{2})$	$\Rightarrow(\frac{1}{4})$	A_aA_2	$\frac{1}{16}$	vii
$A_aA_b(\frac{1}{4})$	$a_2A_2(\frac{1}{2})$	$\Rightarrow(\frac{1}{4})$	A_bA_2	$\frac{1}{16}$	viii
$A_aA_b(\frac{1}{4})$	$a_2A_2(\frac{1}{2})$	$\Rightarrow(\frac{1}{4})$	a_2A_a	$\frac{1}{16}$	ix
$A_aA_b(\frac{1}{4})$	$a_2A_2(\frac{1}{2})$	$\Rightarrow(\frac{1}{4})$	a_1A_b	$\frac{1}{16}$	x

parents ⟨aa,aA⟩+⟨aA,aa⟩ This happens 2 ways:

$\text{parent}_1(p_1)$	$\text{parent}_2(p_2)$	(p_3)	result	probability=$2(p_1p_2p_3)$	case
$a_aa_b(\frac{1}{4})$	$a_2A_2(\frac{1}{2})$	$\Rightarrow(\frac{1}{4})$	a_aA_2	$\frac{1}{16}$	xi
$a_aa_b(\frac{1}{4})$	$a_2A_2(\frac{1}{2})$	$\Rightarrow(\frac{1}{4})$	a_bA_2	$\frac{1}{16}$	xii
$a_aa_b(\frac{1}{4})$	$a_2A_2(\frac{1}{2})$	$\Rightarrow(\frac{1}{4})$	a_2a_a	$\frac{1}{16}$	xiii
$a_aa_b(\frac{1}{4})$	$a_2A_2(\frac{1}{2})$	$\Rightarrow(\frac{1}{4})$	a_1a_b	$\frac{1}{16}$	xiv

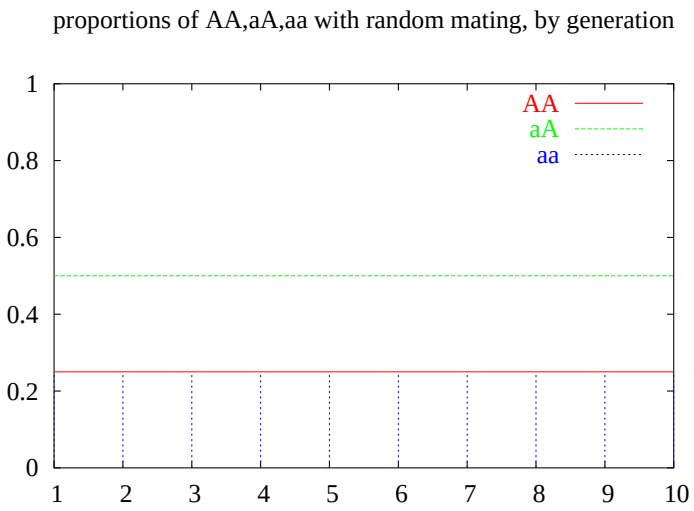
parents ⟨aa,AA⟩+⟨AA,aa⟩ This happens 2 ways:

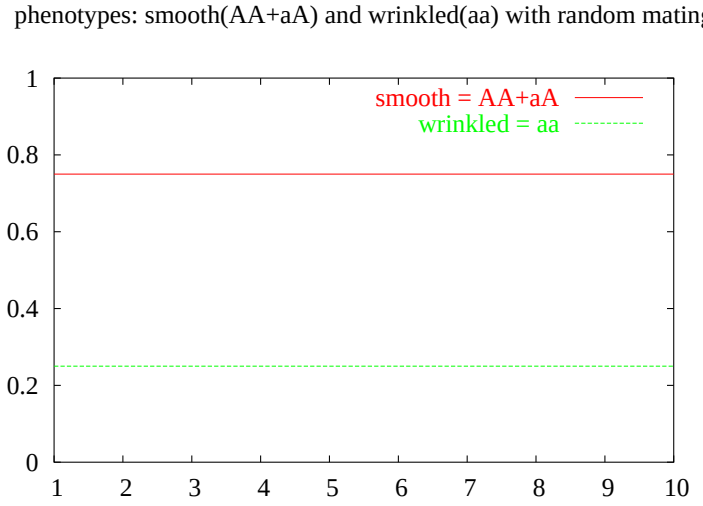
$\text{parent}_1(p_1)$	$\text{parent}_2(p_2)$	(p_3)	result	probability=$p_1p_2p_3$	case
$a_aa_b(\frac{1}{4})$	$A_cA_d(\frac{1}{4})$	$\Rightarrow(\frac{1}{4})$	a_aA_d	$\frac{1}{32}$	xv
$a_aa_b(\frac{1}{4})$	$A_cA_d(\frac{1}{4})$	$\Rightarrow(\frac{1}{4})$	a_bA_d	$\frac{1}{32}$	xvi
$a_aa_b(\frac{1}{4})$	$A_cA_d(\frac{1}{4})$	$\Rightarrow(\frac{1}{4})$	a_aA_c	$\frac{1}{32}$	xvii
$a_aa_b(\frac{1}{4})$	$A_cA_d(\frac{1}{4})$	$\Rightarrow(\frac{1}{4})$	a_bA_c	$\frac{1}{32}$	xviii

Notice that the sum of the probabilities of cases i-xviii is 1, as required. Furthermore, we can calculate that the relative proportion of each genotype is unchanged.

$$\begin{aligned}
p(AA) &= (\text{sum cases i, v, vii, viii}) \\
&= \tfrac{1}{16} + \tfrac{1}{16} + \tfrac{1}{16} + \tfrac{1}{16} = \tfrac{1}{4} \\
p(aA) &= (\text{sum cases ii, iii, ix, x, xi, xii, xv, xvi, xvii, xviii}) \\
&= \tfrac{1}{16} + \tfrac{1}{16} + \tfrac{1}{16} + \tfrac{1}{16} + \tfrac{1}{16} + \tfrac{1}{16} + \tfrac{1}{32} + \tfrac{1}{32} + \tfrac{1}{32} + \tfrac{1}{32} = \tfrac{1}{2} \\
p(aa) &= (\text{sum cases iv, vi, xiii, xiv}) \\
&= \tfrac{1}{16} + \tfrac{1}{16} + \tfrac{1}{16} + \tfrac{1}{16} = \tfrac{1}{4}
\end{aligned}$$

So the proportions of each genotypes does <u>not</u> change when the mating is random. The graph of the proportions of aa, aA, and AA is just 3 constants:





Contrast the graphs of Mendel's results: we see that random breeding is very different from self-breeding. Mendel showed that self-breeding leads to increasing proportions of homozygotes,

but we see now that random breeding leads to **no change**.

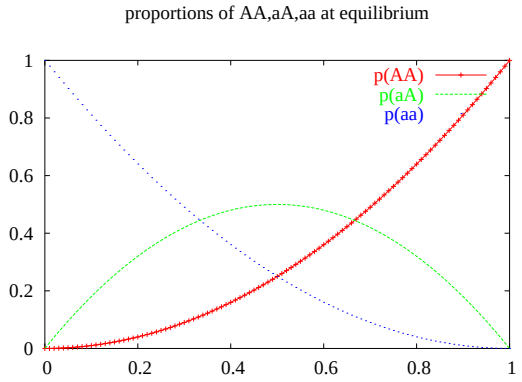
What Hardy and Weinberg noticed is that in cases like this, the probability of having a parent AA is exactly p_A^2, the probability of an aA parent is $2p_a p_A$, and the probability of an aa parent is p_a^2, and when breeding is random, **this holds constant from one generation to the next.** When $p_A = p_a = \frac{1}{2}$, the relative frequencies of the genotypes AA, aA and aa are 1:2:1. What happens when the proportion of A genes is $\frac{3}{4}$? In this case, we have a different equilibrium point. By the product rule, it is not hard to see that in this case the population will have $\frac{3}{4}^2 = \frac{9}{16}$ AA, $\frac{1}{4}^2 = \frac{1}{16}$ aa, and $2(\frac{3}{4})(\frac{1}{4}) = \frac{3}{8}$ aA. So in general, we can determine the relationship with this "shortcut:"

$$\text{since } p_a + p_A = 1, \text{ we have } 1 = (p_a + p_A)^2 = p_a^2 + 2p_a p_A + p_A^2$$

and since then $p_A = 1 - p_a$,

$$p_{AA} = p_A^2 \qquad p_{aA} = 2(1 - p_a)p_a \qquad p_{aa} = p_a^2 = (1 - p_A)^2.$$

We can plot these values, for different values of the relative frequency of the dominant gene A:



When there are 3 alleles for some trait, a, A and $\mathbb{A}$, we have the similar relationship:

$p_a + p_A + p_{\mathbb{A}} = 1$, so we have $1 = (p_a + p_A + p_{\mathbb{A}})^2 = p_a^2 + p_A^2 + p_{\mathbb{A}}^2 + 2p_a p_A + 2p_a p_{\mathbb{A}} + 2p_A p_{\mathbb{A}}$

1.5 Perturbing equilibrium

We saw in the last section that if each organism in each species gets a gene A strictly according to the relative frequency of that gene, the proportions of each genotype goes immediately to a fixed, stable point, an equilibrium, and never changes thereafter. Mendel showed that self-breeding produces not equilibrium, but a continual change. Darwin's point was really that neither of these is what happens in real life. In real life, various kinds of selection are acting. Your chances of surviving to reproduce depends in part on which genes you have, and this can affect the relative frequencies of genes. In fact, there are many ways to affect these relative frequencies. We review a few of the important ones here.

1.5.1 selection: predators, disease, food supply,...

Suppose that, initially, there are the same number of A and a alleles in our garden, as in Mendel's study. But then a family of rodents moves in that likes wrinkled peas but not the smooth ones. Suppose these rodents eat $\frac{1}{4}$ of the wrinkled peas in each generation. What happens then? Then in generation 1, instead of

$$\frac{1}{4}aa + \frac{2}{4}aA + \frac{1}{4}AA$$

we have:

$$\frac{1}{8}eaten(aa) + \frac{1}{8}aa + \frac{4}{8}aA + \frac{2}{8}AA.$$

In this generation, the proportion of A to a has changed. If there were 8 seeds, 1 was eaten, and so if we count the non-eaten alleles here, we get

$$(1a + 1a) + (4a + 4A) + (2A + 2A) = 6a + 8A,$$

so $p_A = \frac{8}{14}$ and $p_a = \frac{6}{14}$. So, using the equation from the previous section, we have:

$$p_{AA} = p_A^2 = \frac{64}{196} \qquad p_{aA} = 2p_Ap_a = 2\frac{6}{14}\frac{8}{14} = \frac{96}{196} \qquad p_{aa} = p_a^2 = \frac{36}{196}$$

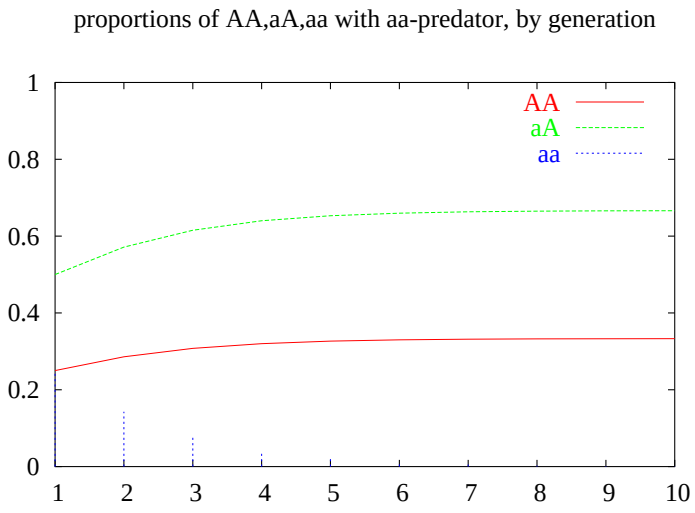
Assuming 196 seeds, when half of the aa seeds are eaten now, in generation 2 we are left with

$$p_{AA} = \frac{64}{(196-18)} = \frac{64}{178} \qquad p_{aA} = \frac{96}{(196-18)} = \frac{96}{178} \qquad p_{aa} = \frac{(36-18)}{(196-18)} = \frac{18}{178}$$

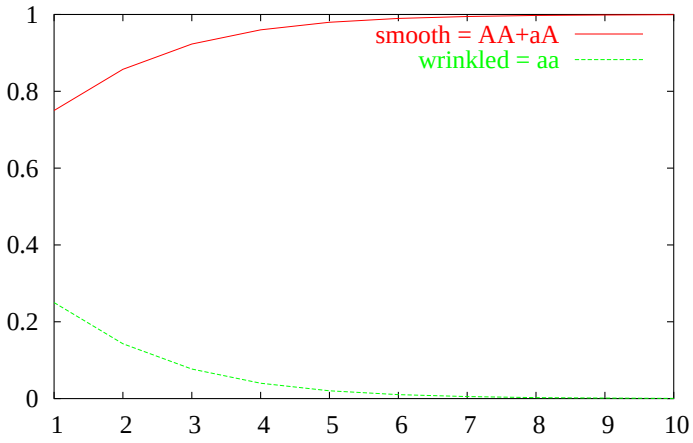
Counting the non-eaten alleles here, we have

$$(64A + 64A) + (96a + 96A) + (18a + 18a) = 224A + 132a$$

a ratio of 56A:33a. So $p_a = \frac{33}{89}$ and $p_A = \frac{56}{89}$. We can see that the proportion of a's is decreasing rapidly! The poor rodents will starve!

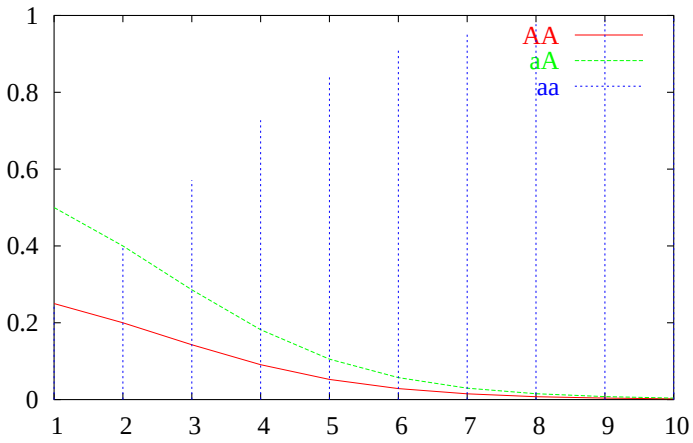


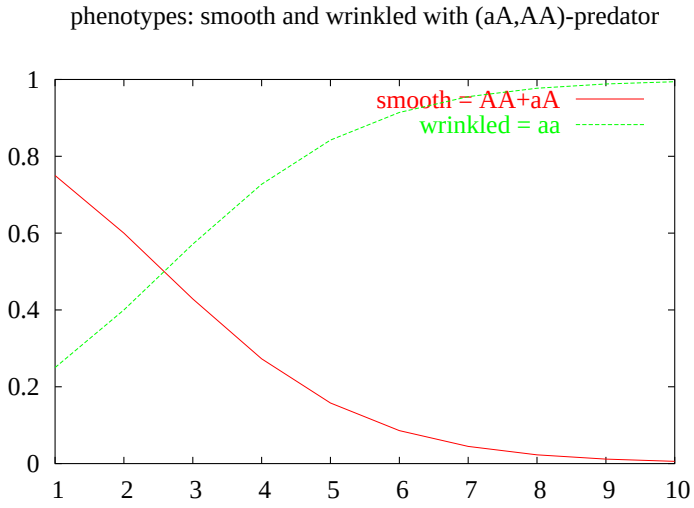
phenotypes: smooth and wrinkled with aa-predator



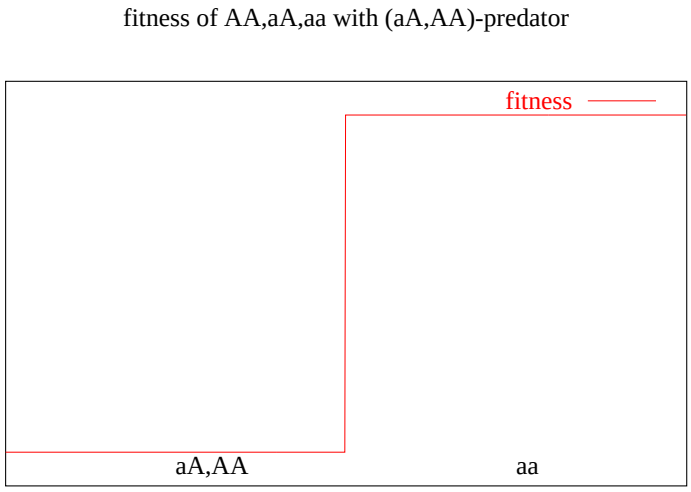
Compare a rodent that prefers smooth peas: this preference does not impact the proportion of A's as quickly, because both the aA and the AA plants are smooth. This spreads the impact. Still after a while, the (aA,AA)-eating rodents will get hungry, and the wrinkled peas will take over!

proportions of AA,aA,aa with (aA,AA)-predator, by generation

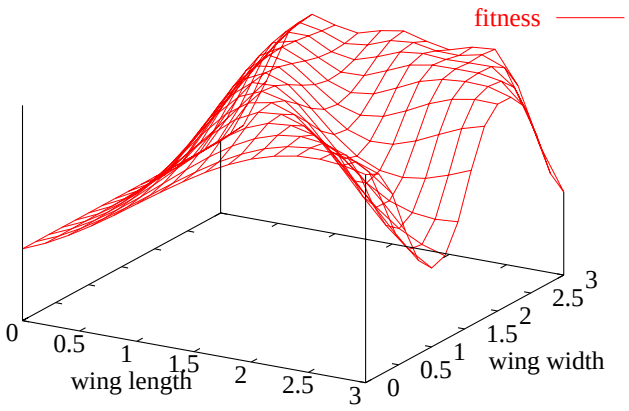




Another way to visualize what's going on is in terms of **fitness**, where fitness is a measure of how well each genotype will fare in the environment: how many offspring it will produce on average. When there is a smooth pea predator, an (aA,AA)-eating rodent that eats half the smooth peas, we see that the smooth peas have low fitness, while the wrinkled peas have high fitness:



As noted on page 26 above, at the end of section 1.3, there can be many genes determining typical traits, which can yield rather smooth variation. For example, the variations in the size and shape and structure of a bird's beak or a fruit fly's wing can vary rather smoothly. And the fitness of a particular size or shape is often highest in the middle and low at the extremes, with possibly complex interactions determining exactly what shape is best. When we plot wing length and width against fitness, for example, we might end up with a graph that looks like this:



These graphs are sometimes called **fitness landscapes**.

1.5.2 non-random mating

In the last section, we saw how the effect of a garden pest could quickly change the population of peas into the "pest-resistant" smooth variety. Another factor which can have a similar influence in the genetic endowment of animals that choose their mates, is mating preference. If the healthy, reproducing females prefer the non-aa males, for some recessive allele a, this will obviously have an effect on the population similar to the garden pest, leading to an increase in the dominant A allele.

1.5.3 drift

Because of the Hardy-Weinberg equilibrium effect, small random changes in a population will tend to stay there unless selection or some other non-random force acts on them. This means that the genes in a population will not stay the same, because over the short term there will be small changes, and these will persist unless they introduce some disadvantage. It is now widely agreed this factor is a very important one, since many genetic changes appear to be unrelated to any selective influences, at least not immediately. Drift can introduce lots of variation that may play a role in selection only many generations later.

1.6 Mutations for new variation

Ronald Fisher describes seeing "a very large number" of visible mutations in laboratory-raised fruit flies (*Drosophila*), and he observes that most of them are not adaptive:

> *The great majority of these are very obvious defects and deformities affecting principally the wings and eyes, but sometimes the body and legs. Beyond these, however, there are a much larger number that are completely lethal, in that the mutant fly is incapable of developing beyond the larval stage, and often even beyond the egg…we must admit that the greater part of the mutational activity to be observed is completely ineffectual in regard to evolutionary consequences. In asserting this, however, there is no reason to deny that a minority of mutations, especially those that are rarer and slighter in their effects, may contain the ingredients out of which adaptive improvements will, in the future, be built up.* (Fisher, 1934, p.254)

With modern methods and an understanding of the chemical basis of inheritance, it is possible to study the frequency of mutations. (Drake et al., 1998) describe a molecular study of fruit fly (*Drosophila*) DNA: looking at 13 places in 490,118 chromosomes, they found 51 mutations, for a rate of

$$\frac{51}{490118 \times 13} = 8.0044 \times 10^{-6} \text{ mutations per position per generation}$$

and a molecular study of six positions in mouse DNA, detecting 69 mutations out of 1,485,036 samples for a rate of

$$\frac{69}{1485036 \times 6} == 7.7439 \times 10^{-6} \text{ mutations per position per generation}$$

These rates may seem very low, but these animals have very complex DNA, with many positions, as we will discuss next week.

1.6.1 migration, isolation

When a subpart of the population migrates to a new area, far enough from the original that it is genetically isolated, it can change the gene pool in several ways:

- the migrating population may have been a biased sample from the original population (e.g. the brave, strong ones),
- once in the new place, this population may be under different selective pressures than the original population
- once in the new place, genetic drift in this population is likely to differ from the original population

Darwin thought he saw evidence of this in the birds of the Galapagos, and on other islands (as we noted on page 5). Biologists are now trying to see this process actually happening.

1.6.2 The power of gradual change

The contemporary zoologist Richard Dawkins is well known for a funny illustration of the power of gradual change. It is now thought that whales evolved from a dog-like animal, but these animals do not look at all similar. We have seen in mathematically simple cases how the

shaping influences of selection, mate preference, and mutation can have rapid, dramatic effects, but could the whale really have come from a dog? This kind of puzzle reminded Dawkins of the following passage from Shakespeare, where Hamlet is talking with the very agreeable Polonius:

Hamlet: Do you see yonder cloud that's almost in shape of a camel?
Polonius: By the mass, and 'tis like a camel, indeed.
Hamlet: Methinks it is like a weasel.
Polonius: It is backed like a weasel.
Hamlet: Or like a whale?
Polonius: Very like a whale.

So Dawkins considers the following similar puzzle in *The Blind Watchmaker*.

Suppose we take an arbitrary string of 28 characters, and make arbitrary changes in it. How long will it take to get to "Methinks it is like a weasel"? Suppose the possible characters are just the 26 letters and 1 space. Then there are 27 possible characters that could appear in each position. We could think of this as a problem of rolling 27-sided dice. Suppose we roll once for the first character, again for the second and so on. What are the chances of rolling all the right characters in order? The chances are $\frac{1}{27}^{28} = 8.3525 \times 10^{-41}$. This is a very small number! It has 41 zeroes after the decimal point!

But now suppose we begin with a random string of 28 characters, and reproduce it 100 times, but imperfectly, with mutations. Suppose that, on each character, there is 1 chance in 10 that it will be changed. Then suppose we adopt the following selection rule: choose the offspring that agrees with "methinks it is like a weasel" on the most characters (choosing an arbitrary one if there is a tie), and kill off all the other offspring before reproducing the next generation. (This is an unrealistically strong selection effect, but that makes it easy to study! And it is interesting to consider whether we are still in the range of effects that show up only once in 10^{46} times.) How many generations will it take to get to "methinks it is like a weasel" in this simple setup? This experiment is easy to program (a program that does this appears on the webpage). Trying this program a few times, the desired string was obtained at generation 156, 127, 181,...So the moral is the same as we saw above: with a non-random shaping influence, the genetic population can change quickly and dramatically.

```
1:  uvk fgigposzldoxyazwibxyqjdp
2:  uvkbogigposzldoxynzbibxyqjdp
3:  uvkbogiyposzld xynzbfbxyqjdp
4:  udrbogiyposzld xynzbfbxhkjdp
5:  udqbogifxgszld xynzbf xhkjdp
6:  udqzogif gszld xynzbf xhkjdp
7:  udqzogif gs ld xynzbf xhkjdp
8:  udqzogif gs ld xynz f xnkjdy
9:  udqzogif gs ld pynp f xlksde
10:  udqzopif gs ld phnp f xlksde
11:  udqzopit gs ld phnp f xlksde
12:  udqmopit gs ld phnp f xbksde
13:  udqmoplt gs ld phog f xbksde
14:  udqmoplt gs lw phog f xbasde
15:  udqmoplt gs lw phog f xbasde
16:  udqmoplt gs lw phog a xbasdl
17:  udqgoplt gs lw phog a xbasdl
18:  pdqgoplt gs fw phog a xbasdl
19:  pdqgoplt gs fr phog a xbasdl
20:  pdqgoplt gs fr phmg a xbasdl
21:  pdqgcplt gs fr nhmg a xbasdl
22:  pdsgcplt gs fr nhmg a xbasdl
23:  pdsgcplt gs hr nhmg a xhasdl
24:  pdsgcplt gs hr nhkg a xhasdl
25:  pdsglplt gs jr nhkg a xhasdl
```

```
26:  ndsglplt gs jr nhkg a xhasdl
27:  ndsgkplt gs jr nhkg a xhasdl
28:  nesgkplt gs jr nhkg a xhasdl
29:  nesgkplt gs jr nhkg a xhasdl
30:  netgkplt gs jr nhkg a xhasdl
31:  netgkplt gs jr nhkg a xhasdl
32:  netgkplt gs js nhkg a xhasdl
33:  netgkplt gs js nhke a xhasdl
34:  netgkplt gs js nhke a xhasdl
35:  netgkolt gs js nhke a xhasdl
36:  netgkolt gs js nhke a xhasdl
37:  netgkolt gs js nhke a xhasdl
38:  netgkolt gs js nhke a xbasdl
39:  netgkolt gs js nhke a xbasdl
40:  netgkolt gs js nhke a xbasdl
41:  netgkolt gs js mhke a xbasdl
42:  netgkolt gs js mhke a xbasdl
43:  netgkolt gs js mhke a xbasdl
44:  netgkolt gs js mhke a xbasdl
45:  netgkolt gs js mhke a xbasdl
46:  netgkokt gs js mhke a xbasdl
47:  netgkokt gs js mhke a xbasel
48:  netgkokt gs js mhke a xbasel
49:  netgkokt gs js mhke a wbasel
50:  netgkokt gs js mhke a wbasel
51:  netgjokt gs js mhke a wbasel
52:  netgjokt is js mhke a wbasel
53:  netgjokt is js mhke a wbasel
54:  netgjokt is js mhke a wcasel
55:  netgjokt is js mhke a wcasel
56:  netgjokt is js mhke a wcasel
57:  netgjokr is js mhke a wcasel
58:  netgjoks is js mhke a wcasel
59:  netgjoks is js mhke a wcasel
60:  netgjoks is js mhke a wcasel
61:  netgjoks is js mhke a wcasel
62:  netgjoks is js mhke a wcasel
63:  netgjoks is js mhke a wcasel
64:  netgjoks is js mhke a wcasel
65:  netgjoks is js mhke a wcasel
66:  netgjoks is js mhke a wcasel
67:  netgjoks is js mhke a wcasel
68:  netgjoks is js mhke a wcasel
69:  netgjoks is js mhke a wcasel
70:  netgjoks is js mhke a wcasel
71:  netgjoks is js mhke a wcasel
72:  netgioks is js mhke a wcasel
73:  netgioks is js mhke a wcasel
74:  netginks is js mhke a wcasel
75:  netginks is js mhke a wfasel
76:  netginks is js mhke a wfasel
77:  netginks is js mhke a wfasel
78:  netginks is js mhke a wfasel
79:  netginks is js mhke a wfasel
80:  netginks is js mhke a wfasel
81:  netginks is js mhke a wfasel
82:  netginks is js mhke a wdasel
83:  netginks is js mhke a wdasel
84:  netginks is js mhke a wdasel
85:  netginks is js mhke a wdasel
86:  netginks is js mhke a wdasel
87:  netginks is is mhke a wdasel
88:  netginks is is mhke a wdasel
89:  netginks is is mhke a wdasel
90:  netginks is is mhke a wdasel
91:  netginks is is mhke a wdasel
92:  netginks is is mhke a wdasel
93:  netginks is is mhke a wdasel
94:  netginks is is mhke a wdasel
95:  netginks is is mhke a wdasel
96:  netginks is is mhke a wdasel
97:  netginks is is mhke a weasel
98:  netginks is is mhke a weasel
99:  netginks is is mhke a weasel
100:  netginks is is mhke a weasel
101:  netginks is is mhke a weasel
102:  netginks is is mhke a weasel
103:  netginks is is mhke a weasel
104:  nethinks is is mhke a weasel
105:  nethinks is is mhke a weasel
106:  nethinks is is mhke a weasel
107:  nethinks is is mhke a weasel
108:  nethinks is is mhke a weasel
109:  nethinks is is mhke a weasel
110:  nethinks is is mhke a weasel
111:  nethinks is is mhke a weasel
```

```
112:  nethinks is is mhke a weasel
113:  nethinks is is khke a weasel
114:  nethinks is is khke a weasel
115:  nethinks is is khke a weasel
116:  nethinks is is khke a weasel
117:  nethinks is is mhke a weasel
118:  nethinks is is mhke a weasel
119:  nethinks is is mhke a weasel
120:  nethinks is is mhke a weasel
121:  nethinks is is mhke a weasel
122:  nethinks is is mhke a weasel
123:  nethinks is is mhke a weasel
124:  nethinks is is mhke a weasel
125:  nethinks is is mhke a weasel
126:  nethinks it is mhke a weasel
127:  nethinks it is mhke a weasel
128:  nethinks it is mhke a weasel
129:  nethinks it is mhke a weasel
130:  nethinks it is mhke a weasel
131:  nethinks it is mhke a weasel
132:  nethinks it is mhke a weasel
133:  nethinks it is mike a weasel
134:  nethinks it is mike a weasel
135:  nethinks it is mike a weasel
136:  nethinks it is mike a weasel
137:  nethinks it is mike a weasel
138:  nethinks it is mike a weasel
139:  nethinks it is mike a weasel
140:  nethinks it is mike a weasel
141:  nethinks it is mike a weasel
142:  nethinks it is mike a weasel
143:  nethinks it is like a weasel
144:  nethinks it is like a weasel
145:  nethinks it is like a weasel
146:  nethinks it is like a weasel
147:  nethinks it is like a weasel
148:  nethinks it is like a weasel
149:  nethinks it is like a weasel
150:  nethinks it is like a weasel
151:  nethinks it is like a weasel
152:  nethinks it is like a weasel
153:  nethinks it is like a weasel
154:  nethinks it is like a weasel
155:  nethinks it is like a weasel
156:  methinks it is like a weasel
```

Looking at this run of the program, we can see that the effect of variation ans selection is similar to organismic evolution in another respect too. As Fisher notes in the passage quoted on page 34, most mutations are not beneficial. Even with our very strict selection rule, the program had to consider 15,501 candidates to get this sequence of 156 improved ones.

1.7 Summary

In this lecture we briefly look at natural history as a science of relative frequencies, relative frequencies of genes and traits in a population. After quickly reminding ourselves how populations grow geometrically, we saw how Mendel's observations of pea plants gave him some ideas about genetic transmission which have proved to be largely correct, even though the physical realization of the transmission was not understood. Probabilities are related to relative frequencies (by the laws of large numbers), and we saw that Mendel's results can be understood with two simple laws of probability: the product rule for independent events and the sum rule for disjoint events. Hardy and Weinberg showed how this works in simpler and more general form when mating is random, but when there is any non-random influence, we saw how this can have immediate and dramatic consequences for the relative frequencies of genes. Several sources of non-random influence were considered, several of which are instances of Darwin's natural selection.

Exercises

1. **Darwin:** One of the things Darwin worries about in the last chapter of *The Origin of Species* is the emergence of sterile offspring. For example, mules are sterile, and so are worker ants. Why does this seem like a serious objection to his proposals about evolution? What is Darwin's brief answer? Does his answer persuade you? (A short answer to this one is fine: a short paragraph should be enough.)

2. **Population growth:** Suppose we start building a population from 1 female fruit fly. Suppose every female fruit fly lays about 500 eggs when it is 2 weeks old, and that 250 of these are females that survive to reproduce in the next generation. So then, the population is growing by a factor of 250 in each generation. If nothing slows this rate of growth, how many years before the total mass of fruit flies is larger than the mass of the earth? (Show your work)
   (Assume that a fruit fly weighs 1 milligram, the earth weighs 6×10^{27} grams, and there are 52 weeks per year.)

3. **Mendel:** Suppose you observe that some pea plants have some trait X. How can you tell whether X is genetically determined? (Again: a short answer about how you would do this)

4. **Hardy:** Suppose that the relative proportions of genotypes in a population are these:

   $$\text{a } \tfrac{1}{5} \qquad \text{A } \tfrac{4}{5}$$

   What are the proportions of aa, aA, and AA genotypes after random mating? (Show your work)

5. **Hardy:**[2] Suppose that there are two alleles R and r for the Rh factor in humans, where r is recessive and R is dominant. The rr individuals are called Rh negative; the rR and RR individuals are called Rh positive. The offspring of an Rh negative female and an Rh positive male may suffer serious anemia while still in the uterus. Assuming that the relative frequency of r is $\frac{1}{8}$, and assuming random mating, what proportion of each new generation will be at risk of this anemia? (Show your work)

6. **Darwin:** In the last chapter of *Origin of Species* Darwin says,

   > As natural selection acts solely by accumulating slight, successive, favourable variations, it can produce no great or sudden modification; it can act only by very short and slow steps.

   Two questions: (i) Why do you think he believes this? (ii) Do the basic mechanisms of evolution (the 3 main components we discussed) require this?

7. **Darwin:** It is natural to assume that a complicated organ like the human eye evolved its particular characteristics because vision is so valuable for survival and reproduction, including the light sensitive proteins (opsins) in the retinal cells, the lens, the eyelids and eyelashes. (We mentioned the opsins in the introduction, and even showed a picture, on pages 12-13.)
   But the very same proteins that sense the light in human eyes are present in most light-sensing organisms, all the way down to light-sensitive bacteria. (i) Does this provide any reason to doubt that the opsin genes were selected because of their value in the eye?

[2] This question is adapted from (Boyd and Silk, 2000, p92).

Furthermore, the cells in the lens of the human eye are modified epithelial (surface) cells (found in all vertebrates), containing a special soluble proteins ('crystallins'), that are also present in most vertebrates, including non-sighted ones. (ii) Does this provide any reason to doubt that the genes for the skin cells and the crystallins for the lens were selected for the eye? (briefly defend your answers to both i and ii)

8. **Population growth:** A desert locust can hatch and grow to maturity, in ideal conditions, in about 4 weeks A female desert locust lays 5 pods of approximately 80 eggs each. Suppose half of these eggs produce females that survive, and that each adult locust weighs 10 grams. If the locusts could reproduce and mature at this rate indefinitely, how long before they would weigh as much as the earth (6×10^{27} grams)? (Show your work)

9. **Hardy:** Suppose that the relative proportions of genotypes in a population are these:

   $$\text{a } \tfrac{1}{3} \qquad \text{A } \tfrac{2}{3}$$

   Assuming that the phenotype is always determined by the dominant gene (as in Mendel's study), what proportion of the population would show the dominant phenotype after random mating? (Show your work)

10. **Selection:** Many biology books describe a report by Kettlewell (1958), who studied how a dark-winged variety of "peppered moth" came to predominate over the light-winged variety in late 1800's Britain. He proposed that this happened because the industrialization of Britain darkened tree trunks, making the lighter variety of moth more visible on tree trunks where it could be eaten by birds. A field study reported that the light moths on the dark trees were, in fact, more often eaten by birds.

    Even this classic study of selection has some puzzling features though. Lewontin (2002) points out that the moths in the field study were tethered to the trees, raising questions about whether the differential predation (of the wild, untethered moths) by birds was really the selective force at work here. And furthermore, it was discovered that even in the caterpillar stage (before the light or dark wings grow), the caterpillars of the dark-winged moths survive more often than the caterpillars of the light-winged moths. Considering these facts, do you think we should still accept Kettlewell's hypothesis that the dark-wings came to predominate because they were selected? (Briefly explain why you are still inclined to believe Kettlewell, or not.)

Solutions to Selected Exercises

1. **Darwin: (short answer!)** Natural selection favors the traits that lead to the most descendants, so the appearance of sterile offspring after centuries of natural selection might lead you to doubt that natural selection is really what determines which traits are propagated through a population. Darwin's brief response to this points out that when an organism is thrown into an unusual environment, it would be no surprise, and no contradiction of natural selection, to find that it cannot reproduce (like a fish out of water!). But this response does not immediately handle the mule or the worker ant. With respect to the mule, the offspring of a horse and donkey, this too is an unusual circumstance. Darwin says "their constitutions can hardly fail to have been disturbed from being compounded of two distinct organisations." When horses breed with horses and donkeys with donkeys, as is the usual case, the offspring are fertile, and obviously that is how the species survive. But what about the worker ant? In an ant colony, there is the reproducing queen and many sterile worker ants. How does this fit Darwin's theory? Darwin does not provide an explanation here in the last chapter, but we can see what he might have said. The activities of the sterile worker ant lead to the success of the genetically very similar queen, and that queen will produce other genetically similar queen ants to found other colonies.

2. **Population growth:** If a fruit fly weighs 1 milligram and the earth weighs 6×10^{27} grams, then 6×10^{30} flies weigh as much of the earth. So the question is, how many generations are needed to produce that many flies?

   If the population grows by a factor of 250 in each generation, then in generation n, there are 250^n flies. So the question is, for what n does $250^n = 6 \times 10^{30}$. One way to do this is by calculating 250^n for $n = 1, 2, 3, \ldots$ until you exceed 6×10^{30}. On my calculator, this looks like this (I typed the stuff after the » prompts, the ** is exponentiation, and e+23 means $\times 10^{23}$):

   ```
   » 10**2
   ans = 100
   » 250**10
   ans = 9.5367e+23
   » 250**11
   ans = 2.3842e+26
   » 250**12
   ans = 5.9605e+28
   » 250**13
   ans = 1.4901e+31
   ```

   So 13 generations is more than enough. Since each generation takes 2 weeks, 26 weeks, or half a year, is more than enough!

There is an easier and more exact way to calculate the n such that $250^n = 6.0 \times 10^{30}$. (Since some people asked for a review of this, I have added some notes about how to figure this kind of problem out, on pages 46-48.) This formula can be expressed as $n = \log_{250} 6.0 \times 10^{30}$, which in turn is $n = \frac{\log_{10} 6.0 \times 10^{30}}{\log_{10} 250}$, so

```
» log10(100)
ans = 2
» log10(6e+30)/log10(250)
ans = 12.835
```

So 12.835 generations is exactly enough. Since each generation takes 2 weeks, 25.670 weeks is enough, and so $\frac{25.670}{52} = 0.49366$ year is enough. That's $0.49366 \times \frac{365 \text{days}}{\text{year}} = 180.18$ days.

3. **Mendel: (short answer!)** Figuring out whether a trait X in pea plants is genetically determined can be challenging. When X is a relatively simple trait, as in Mendel's study, we can look at the proportions of descendants of plants with trait X and see if you find the proportions Mendel predicts. But that works only when X is determined by 2 alleles. Most genetically determined traits are determined by a certain combination of many different genes, and these genes may be linked in various ways. But even in these complex cases you can get evidence that X is genetically determined by seeing whether descendants of plants with X tend to have X too, when variations in growing conditions are controlled for.

   We will return to this problem again later! (see, e.g. page 165)

4. **Hardy:** With random mating, the proportions are these

$$\begin{aligned} p_{aa} &= p_a^2 = \tfrac{1}{25} \\ p_{aA} &= 2p_a p_A = 2 \times \tfrac{1}{5} \times \tfrac{4}{5} = \tfrac{8}{25} \\ p_{AA} &= p_A^2 = \tfrac{16}{25} \end{aligned}$$

5. **Hardy:** With random mating, the proportions are these

$$\begin{aligned} p_{rr} &= p_r^2 = \tfrac{1}{64} \\ p_{rR} &= 2p_r p_R = 2 \times \tfrac{1}{8} \times \tfrac{7}{8} = \tfrac{14}{64} \\ p_{RR} &= p_R^2 = \tfrac{49}{64} \end{aligned}$$

   The probability of an Rh negative (rr) female is $\frac{1}{64}$ and the probability of an Rh positive (rR or RR) male is $\frac{49}{64} + \frac{14}{64} = \frac{63}{64}$. By the product rule the probability of this combination is $\frac{1}{64} \times \frac{63}{64} = \frac{63}{64 \times 64} = \frac{63}{4096} = 0.015381$. That is, about 1.5%.

Lecture 2

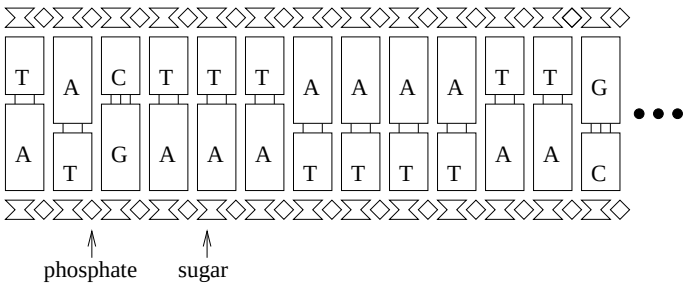
The chemical realization of heredity

2.1 Molecular mechanisms

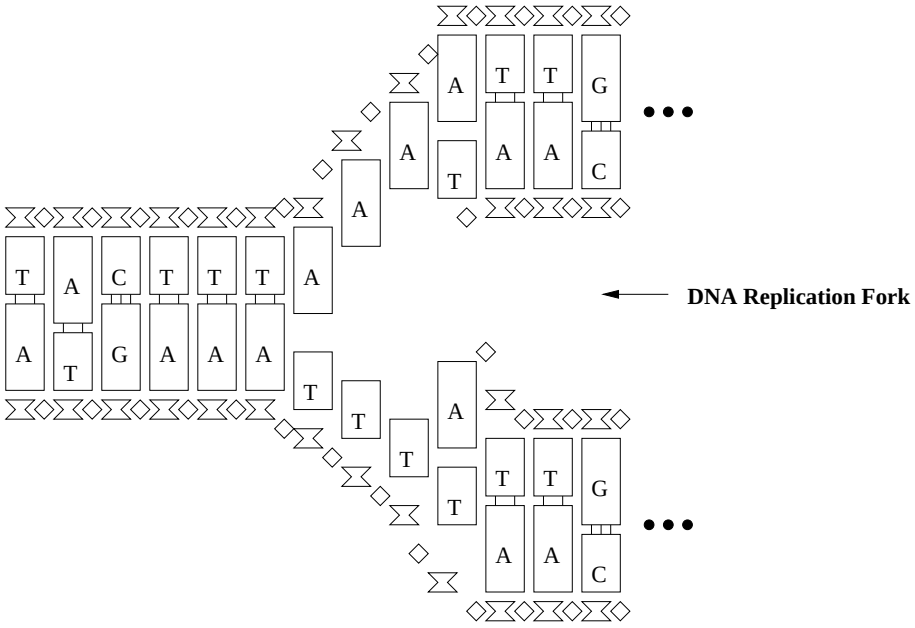
Watson & Crick

Proteins control cellular processes by enabling certain chemical reactions to happen easily, **catalyzing** those reactions, and inhibiting other reactions. As mentioned in section §0.4.2, the structure of proteins is determined by the sequence of amino acids from which they are built. And the sequence of amino acids in a protein is determined by the basic genetic materials DNA and RNA, discovered by Francis Crick and James Watson in 1953, with the assistance of Maurice Wilkins and Rosalind Franklin, and based on the earlier discovery of helical proteins by Linus Pauling and others.

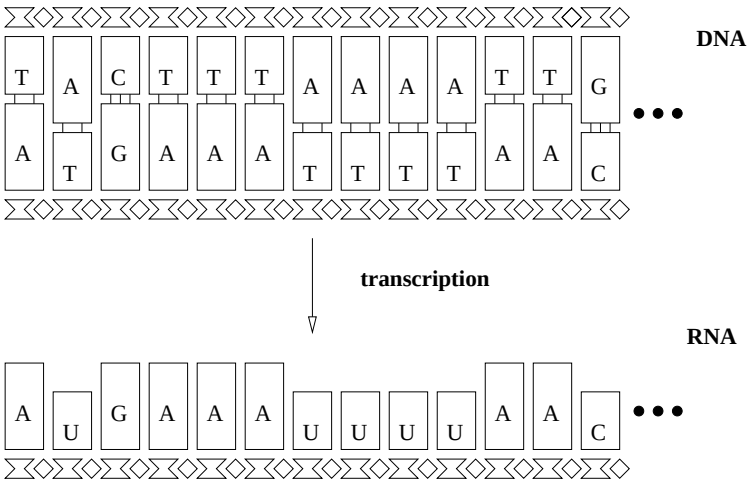
Each DNA molecule has two long coiled strands of **nucleotides**: each nucleotide is a phosphate, a sugar and one of the **bases** Adenine, Thymine, Guanine, and Cytosine. We abbreviate each of these nucleotides with the first letter of the base it contains: a, t, g, c. In their pioneering studies, Watson and Crick noticed that the concentrations of a and t seemed to be nearly identical, up to the limits of measurement accuracy, as did the concentrations of g and c. We now know that the a and t bind together, as do the g and c, and so a sequence like $tactttaaaattg$ will be bound to the complementary sequence $atgaaattttaac$:



Watson and Crick remark in their 1953 paper that this structure suggests a method for replication. A DNA molecule can replicate by simply splitting the complementary chain apart and letting each base on each strand combine with its complement.



The sequences of bases in DNA can also specify complementary sequences in 1-stranded RNA molecules, by a process called **transcription** which are like DNA strands except Thymine (t) is replaced by Uracil (u).



RNA can thus serve as a kind of intermediary, a "messenger" between the DNA itself and the mechanics of protein synthesis.

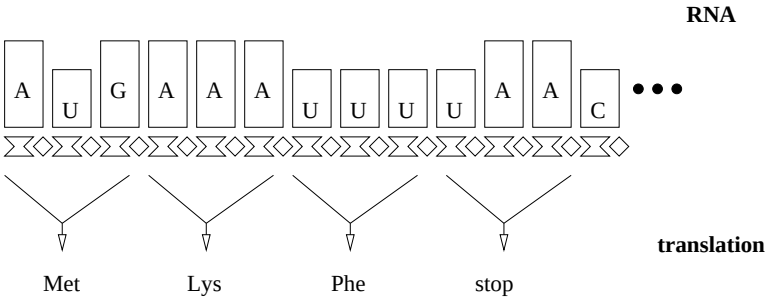
As mentioned in §0.4.2, a protein is a sequence of amino acids, and each of the 20 different amino acids found in organisms is specified by a triple of bases, a **codon**. Since there are 4

different bases, how many different triples are there? Just as there are $10^3 = 1000$ different triples of the 10 digits, so there are $4^3 = 64$ triples of 4 bases. 61 of these triples specify one of the 20 different amino acids, and 3 of them specify the ends of sequences: these "punctuation" marks are called **stop codons**.

The triples of bases in each amino acid, and the stop codons, are shown in this table, using the 3-letter abbreviations for amino acids shown on page 12:

	u	c	a	g	
u	Phe	Ser	Tyr	Cys	u
	Phe	Ser	Tyr	Cys	c
	Leu	Ser	Stop	Stop	a
	Leu	Ser	Stop	Trp	g
c	Leu	Pro	His	Arg	u
	Leu	Pro	His	Arg	c
	Leu	Pro	Gln	Arg	a
	Leu	Pro	Gln	Arg	g
a	Ile	Thr	Asn	Ser	u
	Ile	Thr	Asn	Ser	c
	Ile	Thr	Lys	Arg	a
	Met	Thr	Lys	Arg	g
g	Val	Ala	Asp	Gly	u
	Val	Ala	Asp	Gly	c
	Val	Ala	Glu	Gly	a
	Val	Ala	Glu	Gly	g

For example, Gly (Glycine) is coded by the triplets ggu, ggc, gga, and ggg. So RNA can be **translated** into amino acid sequences to form proteins, polypeptides:



Why are there so many ways of coding each amino acid? Let's consider this question carefully because we will see other versions of the same idea later in the class.

How many bases are needed to name 20 **different amino acids?** We can label each of 10 things with a single decimal digit, because there are 10 of those:

$$0, 1, 2, 3, 4, 5, 6, 7, 8, 9$$

If we want to label 11 things, a single decimal digit is not enough. 2 digits is plenty, since with two digits we have 100 different names:

$$\begin{array}{c} 00, 01, 02, 03, 04, 05, 06, 07, 08, 09 \\ 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 \\ 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 \\ 30, 31, 32, 33, 34, 35, 36, 37, 38, 39 \\ 40, 41, 42, 43, 44, 45, 46, 47, 48, 49 \\ 50, 51, 52, 53, 54, 55, 56, 57, 58, 59 \\ 60, 61, 62, 63, 64, 65, 66, 67, 68, 69 \\ 70, 71, 72, 73, 74, 75, 76, 77, 78, 79 \\ 80, 81, 82, 83, 84, 85, 86, 87, 88, 89 \\ 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 \end{array}$$

So 1 decimal digit suffices to name 10^1 things, 2 decimal digits suffice to name 10^2 things, and in general:

Decimal naming rule: n decimal digits suffice to name 10^n things.

Now suppose that you want to name k different things, for some other k. How many digits do you need? Well what we want is

$$\text{the smallest integer } n \text{ such that } 10^n \geq k.$$

One way to figure this out is to figure out which n is such that

$$10^n = k$$

and then "round it up" to the first greater integer value. The n that satisfies this equation is called the **logarithm base 10** of k. That is,

$$10^n = k \text{ is the same as } n = \log_{10} k.$$

And we use little corners to mean rounding the symbol for rounding up, so

$$\text{the smallest integer } n \text{ such that } 10^n \geq k \text{ is the same as } n = \lceil \log_{10} k \rceil.$$

So if we want to name 20 things using decimal numbers, how many digits do we need? Now $k = 20$ and so

$$\text{the smallest integer } n \text{ such that } 10^n \geq 20 \text{ is the same as } n = \lceil \log_{10} 20 \rceil.$$

On my calculator, I can calculate $\log_{10} 20$ and I find that

$$\log_{10} 20 = 1.301.$$

Then, rounding up, we find

$$\lceil \log_{10} 20 \rceil = \lceil 1.301 \rceil = 2.$$

The question we want to ask involves naming the amino acids with bases, though, where there are not 10 bases, but only 4! So if we want to name 20 things with 4 different bases, how many "digits" do we need? If we just replace the 10 in the equation above by 4, we can figure this out. Since $k = 20$,

$$\text{the smallest integer } n \text{ such that } 4^n \geq 20 \text{ is the same as } n = \lceil \log_4 20 \rceil.$$

If your calculator lets you calculate $\log_4 20$ you will find that

$$\log_4 20 = 2.161$$

(If your calculator cannot calculate $\log_4$, see the trick below.) Then, rounding up, we find

$$\lceil \log_4 20 \rceil = \lceil 2.161 \rceil = 3.$$

To check this, notice that if we have sequences of three bases, we can name $4^3 = 64$ different things, with these names:

uuu uuc uua uug	cuu cuc cua cug	auu auc aua aug	guu guc gua gug
ucu ucc uca ucg	ccu ccc cca ccg	acu acc aca acg	gcu gcc gca gcg
uau uac uaa uag	cau cac caa cag	aau aac aaa aag	gau gac gaa gag
ugu ugc uga ugg	cgu cgc cga cgg	agu agc aga agg	ggu ggc gga ggg

With sequences of 2 bases, we could only name $4^2 = 16$ different things – not enough for the 20 different amino acids.

If we think of the 4 bases as a kind of "vocabulary" for unambiguously naming the amino acids, the general rule we use is this:

Naming rule: n digits from a vocabulary of size b suffices to name b^n things.
So to name k things using a vocabulary of size b, you need sequences of length $n = \lceil \log_b k \rceil$

So now we see that "codons," sequences of 3 bases, are enough to name the 20 different amino acids, while a sequence of 2 bases would not be enough. But 3 bases is enough to name 64 things, so what should be done with all the extra names? What happens is that each amino acid gets a few different names, and **the different names of each amino acid are all similar to one another**. Looking at the table on page 45, we can see that the names of each amino acid are all near each other in the table. And we can see that changing the third element of a codon has no effect at all in 8 out of the 16 cases. Why would the naming scheme be arranged like this? One idea is that this has happened because it reduces the impact of point mutations on protein specification. (We will say more about mutations later.)

The trick for calculating $\log_n k$ for weird n's. Many calculators let you calculate $\log_{10} k$ or $\log_e k$ but not $\log_4 k$. (As I mentioned on page 27, I use **octave** as my calculator, and it has $\log_{10}$, $\log_2$, and $\log_e$, but not $\log_4$!) What can you do when this happens? The thing to do is to use the following important fact. For any number b:

$$\log_n k = \frac{\log_b k}{\log_b n}$$

So if your calculator does only $\log_{10}$, the way to calculate $\log_4 20$ is this:

$$\log_4 20 = \frac{\log_{10} 20}{\log_{10} 4}$$

With my calculator, "octave," the calculation looks like this:

```
» log10 (20) / log10 (4)
ans = 2.1610
```

Or you can do it with $\log_2$ or $\log_e$ too (in octave, $\log_e$ is just log):

```
» log2 (20) / log2 (4)
ans = 2.1610
» log (20) / log (4)
ans = 2.1610
```

On this calculator, the command for "rounding up" is `ceil`, so we can calculate $\lceil \log_4 20 \rceil$ this way:

```
» ceil( log10 (20) / log10 (4) )
ans = 3
```

Departures from the 'universal' genetic code. The coding of amino acids specified by the table above is almost universal – presumably established quite early in the evolution of life – but there are a couple minor variations that occur in mitochondria of most species, and also in certain very small bacteria (Mycoplasma), come ciliated protozoans, and certain other organisms (Osawa et al., 1992):

where	codon	unusual code	'universal' code
in *Mycoplasma*, mitochondria of many species	UGA	Trp	Stop
in mitochonria in yeasts	CUG	Thr	Leu
in *Acetabularia, Tetrahymena*, paramecium, etc	UAA,UAG	Gln	Stop
in *Euplotes*	UGA	Cys	Stop

It is widely thought that these exceptions developed after the 'universal' code was already well-established, but the question is still actively studied.

In modern genetics, a **gene** is a segment of DNA that codes a polypeptide, a protein. Since, each polypeptide folds into a characteristic shape as discussed in in §0.4.2, and these determine how cells grow and reproduce and metabolize their nutrients, this sense of "gene" corresponds closely to Mendel's.

In eukaryotes, genes often are specified by parts of a nucleotide sequence that are interspersed with apparently non-coding sequences, sequences that are pruned away before the translation into proteins. The coding sequences are called **introns**, and the non-coding sequences are **extrons**. In the human genome, the genes comprise about 2% of the DNA; in other organisms like the pufferfish, much more of the DNA is devoted to genes. The average human gene is about 3000 bases, with the largest known human gene, dystrophin, having 2.4 million bases. For more than half of the genes we know about, the functions are completely unknown.

Genomes vary from individual to individual, and species to species, but much less than you might think. The human genome sequence is almost exactly the same in all people, with better than 99.9% identical base pairs. In fact, the human genome is remarkably similar to the genome of the mouse, and other higher vertebrates.

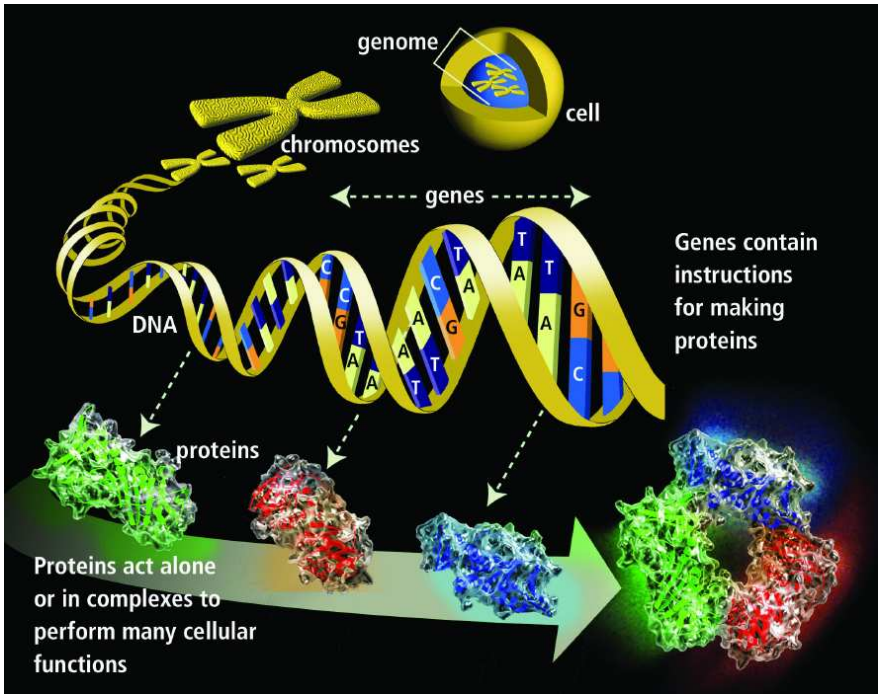
There are enormous variations in the amounts of genetic material in various organisms. The fruitfly genome *Drosophila melanogaster* has about 1.8×10^8 base pairs, while the human has 3.1×10^9 base pairs, and the amphibian called a mudpuppy has more than 130×10^9 base pairs. The range of plant genome sizes is enormous too, ranging from approximately the same size as small animals to more than five times as large as the human. The following table shows some genome sizes (haploid size):

organism	genome mass in picograms ≈no. of base pairs in billions	approx. no. of genes
HIV1	0.000009750	9
Haemophilus influenzae	.00183	1740
Mycobacterium tuberculosis	.004397	
brewer's yeast *Saccharomyces cerevisiae*	0.012	6034
pufferfish *Fugu rubripes*	0.4	>30000
red fire ant *Solenopsis invicta*	0.6	
burrowing frog *Limnodynastes ornatus*	0.95	
cut-throat weaver (bird) *Amadina fasciata*	1.0	
American rattlesnake *Crotalus durissus terrificus*	1.3	
fruitfly genome *Drosophila melanogaster*	1.8	13601
ostrich *Struthio camelus*	2.2	
domestic corn	2.5	25000
coyote *Canis latrans* and domestic dogs	2.8	
African mole-rat *Georhychus capensis*	3.2	
human *Homo sapiens*	3.1647	>35000
chimpanzee *Pan troglodytes*	3.7	
domestic cattle *Bos taurus*	3.7	
jumping spiders *Habronattus*	5.7	
tubificid worm *Spirosperma ferox*	7.6	
mountain grasshopper *Podisma pedestris*	16.9	
deep-sea shrimp *Hymenodora*	38	
Easter lily *Lilium longiflorum*	90	
marbled lungfish *Protopterus aethiopicus*	133	
amoeba dubia	670	

To help get a perspective on these large numbers of base pairs, consider that each base occupies about 3.4 angstroms along the sugar-phosphate backbone, so if all 3 billion of the human bases were stretched out in a straight line, they would be over a meter long. All this material is tightly coiled and twisted in the nucleus of every cell of the human organism (except for the gametes, the sex cells, which have half the genome). And if we put the whole sequence of bases from your genome on a computer disk, using one character per base, in order, the file would require about 3 Gigabytes.

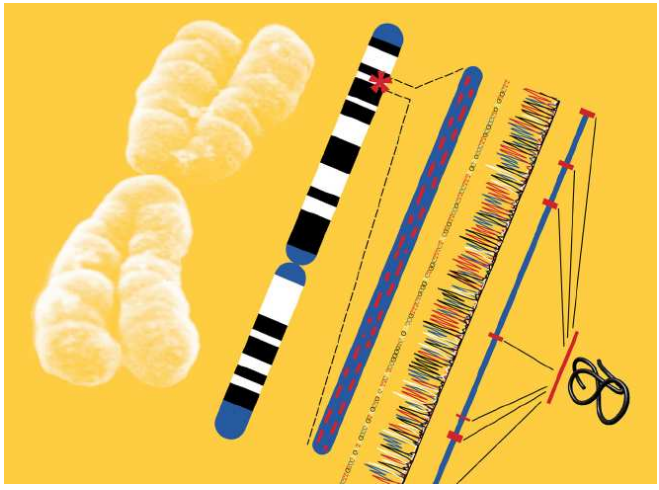
DNA in humans and other eukaryotes is arranged into distinct chromosomes – physically separate molecules that range in length from about 50 million to 250 million base pairs. The human has 23 chromosomes. Genes that share a chromosome will show "linkage" effects, which were mentioned briefly on page 27. We will set these complications in transmission aside for the moment.

The basic genetic processes are pictured on these images produced by the US Department of Energy Genome Project:



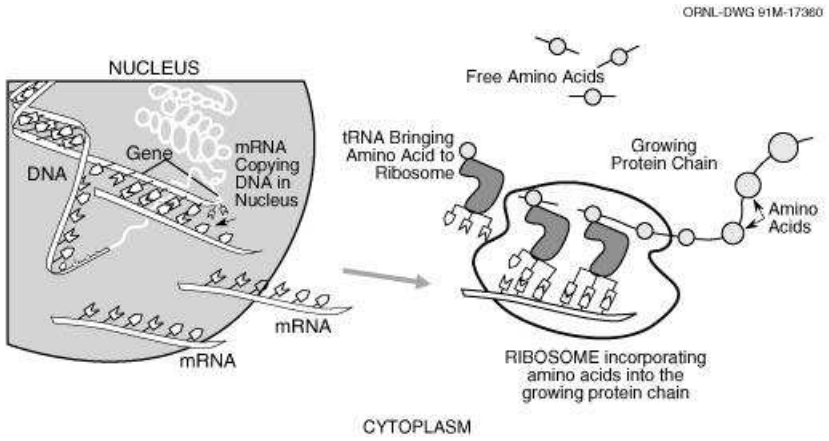
US Department of Energy, Human Genome Program, http://www.ornl.gov/hgmis

The following image indicates how the coding segments of a gene are distributed along the DNA and pruned away before translation into the protein:



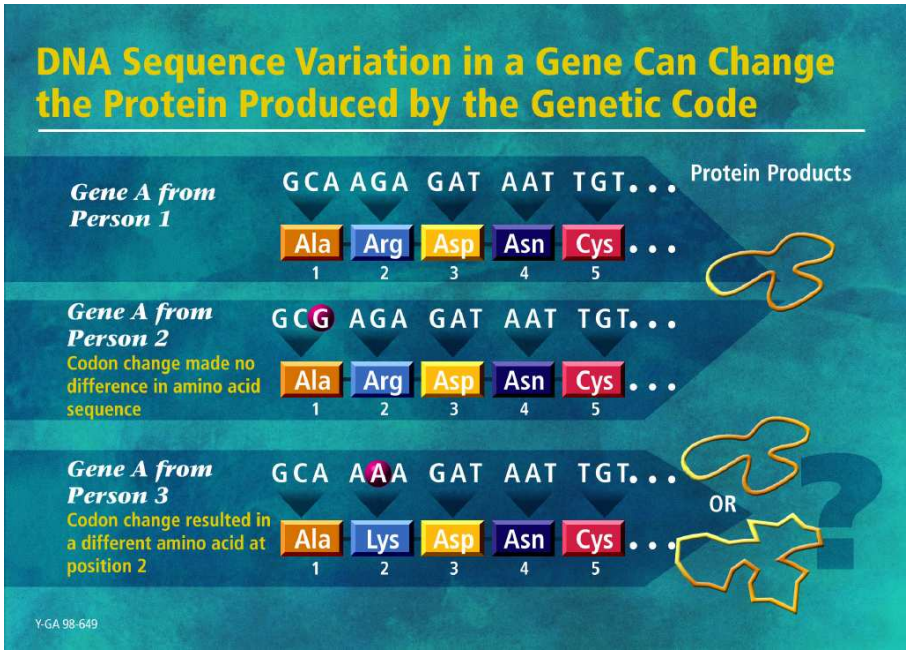
US Department of Energy, Human Genome Program, http://www.ornl.gov/hgmis

And the following image indicates a little more detail: the DNA is transcribed into messenger RNA (mRNA) and transfer RNA (tRNA) for use in the synthesis of a chain of amino acids:



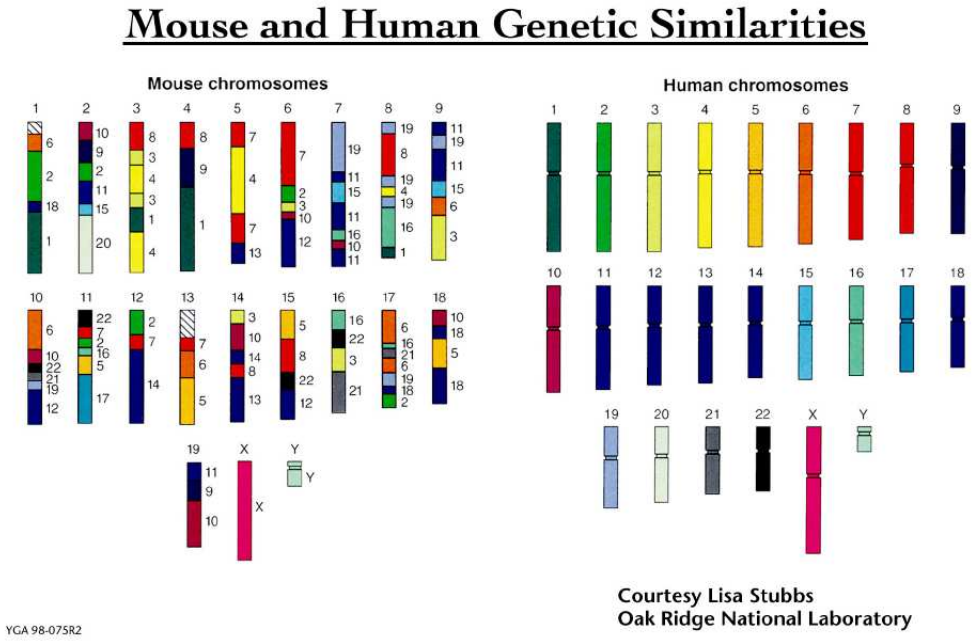
US Department of Energy, Human Genome Program, http://www.ornl.gov/hgmis

It is no surprise that a change in a single base can affect the protein synthesized:



US Department of Energy, Human Genome Program, http://www.ornl.gov/hgmis

And finally, as mentioned above, large parts of human DNA are virtually identical to parts of mouse DNA, but the arrangement is different:



US DOE `http://www.ornl.gov/hgmis`, and cf. (Gunter and Dhand, 2002; Dermitzakis et al., 2002)

2.2 Molecular change: mutation

The process of DNA replication is usually perfect, but occasionally errors occur. Given the enormous size of these delicate molecules, errors are not a surprise, but the nature and extent of changes are surprising. The common kinds of errors can be classified into various types:

Point mutations: these are the most common

- Often they produce no change in the organism, the phenotype
- A point mutation can change a single amino acid, which might or might not matter
- It can change an amino acid codon into a stop codon, blocking the production of some protein
- These changes can sometimes be catastrophic

Deletions, insertions are most often caused by part of a chromosome breaking off

- As we can see from the transcription process above, an insertion or deletion can potentially change many proteins, because it can cause a "frame shift:" that is, the alignment of codons can be affected
- In the special case of a 3-base-pair insertion or deletion, at least many of the same amino acids can be specified

Duplication: this type of "insertion" is some 10 times less likely than point changes

- a base pair can be copied more than once

Inversion, translocation: a broken off piece of a chromosome can attach in inverted order, or in the wrong place, so that

- one or more base pairs can have their order inverted, or
- a subsequence of bases is spliced into the wrong position

Change in number of chromosomes: fortunately, this is also rather rare

- a failure to separate properly in replication can result in too many or too few chromosomes
- in animals, this is usually fatal

Single nucleotide substitutions are roughly 10 times more frequent than length mutations, and mutations are more frequent in some positions than others, but on average the mutation rate in humans may be 2.5×10^{-8} mutations per nucleotide site or even more (Eyre-Walker and Keightly, 1999; Nachman and Crowell, 2000). Since there are approximately 3.2×10^9 base pairs in the human genome, we expect there are something more than

$$\frac{3.2 \times 10^9 \text{ base pair}}{\text{genome}} \times \frac{2.5 \times 10^{-8} \text{ mutations}}{\text{base pair}} \approx 80 \text{ mutations per genome per generation}$$

(Compare the mutation rates for fruitflies and mice mentioned in §1.6 on page 34.) Since most mutations are selectively neutral, and the rate of mutation is so high, this becomes a significant factor in shaping the genome, as emphasized by biologist Motoo Kimura's "neutral theory" of molecular evolution.

2.3 Molecular phylogeny

The sequence of bases in the human genome, more than 3 billion of them, is nearly mapped out now. Once a change is introduced, it is preserved unless the organism fails to reproduce or else the site of change is affected by another change. This means that the chances of two different people having a long sequence of identical base pairs is extremely low.

This fact is used in genetic methods for forensic identification. Not only identity, but relatedness can be established in this way, so these methods are often used to resolve questions about paternity. It is not feasible to determine an individual's entire genetic sequence yet, so the usual methods involve using "marker" chemicals that will bind to certain parts of the DNA, forming a different pattern for different individuals.

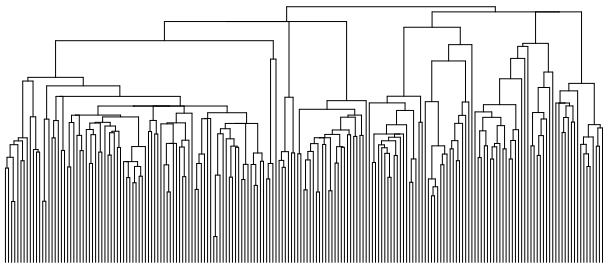
What is more remarkable is that the relatedness of different organisms can be assessed in this way too. If we can identify nearly common DNA sequences in different organisms, the number of small differences provides a rigorous way to assess how far back the most recent common ancestors were. This method has reshaped our picture of the phylogeny of life. The modern picture shown on page 3 in chapter 0 is based on this kind of data (from Woese and others). This modern phylogeny is significantly different from the phylogenies we had even just 15 or 20 years ago. Notice how closely related all the higher animals, the "metazoa" are, compared to their simpler ancestors, and we see that the primitive bacteria-like organisms

divide into two groups, with the the primitive “archea” or “archaebacteria” more closely related to us than the other bacteria.

2.4 Digression: HIV and why AZT fails

The December 2004 UN Aids Epidemic Update estimates the number of people infected with the human immunodeficiency virus (HIV) at that time to be 39,400,000, with more people newly infected in 2003 (the most recent year for which statistics are available) than in any other year. This virus causes the acquired immune deficiency syndrome (AIDS) that is usually fatal within 2 years. HIV is transmitted as a roughly spherical particle that has two strands of RNA together with an enzyme (‘transcriptase’) that transcribes this RNA, all packaged in a protein envelope. This particle bonds to other proteins on the surface of certain cells in the human body, specifically white blood cells (T cells) that attack infections, foreign cells, and cancer cells. T cells also produce other substances that regulate the immune response. The HIV envelope then fuses with the T cell membrane and releases its RNA and its enzymes. These produce DNA in the T cell that is integrated into the cell’s own. The RNA transcribed from this DNA is infected with the HIV genetic material, and transcribed into RNA that is distributed throughout the cell. An infected cell can have 400,000 to 2,500,000 copies of infected RNA (Hutchinson, 2001), which form new particles, budding off thousands of new copies of the HIV particles into the subject’s bloodstream. Advanced HIV infections trigger the collapse of the immune system known as AIDS.

One of the first drugs for HIV/AIDS patients was azidothymidine (AZT), which works by blocking the action of the HIV enzyme transcriptase. This drug appeared to be very successful when it was first introduced, but after just a few years, patients stopped responding to it. Why did the drug stop working? Tests on patients taking AZT showed that the RNA sequences populating the bloodstream were changing *even in a single patient*. That is, the virus was responding to the selective pressure of AZT even in the span of a single infection in a single organism. This extremely rapid evolution could occur because the error rate in the transcription of this RNA is unusually high, making 1-10 errors per replication of its 9750 bases (see the table of genome sizes on page 49). Sampling the HIV population in an infected subject, we find not one strain of HIV, but clusters of many related strains. By comparing the genetic sequences on the molecular level, a branching phylogeny can be drawn, as described in section §2.3:



It is no wonder that the strains resistant to AZT (or any other single drug) will quickly be found and come to dominate the population (Casado et al., 2001; Hutchinson, 2001; Palumbi, 2001).

2.5 Three languages of life

It is easy to summarize what we have said about the languages of DNA, RNA and proteins so far (since we have not said very much yet!).

2.5.1 The language of DNA: first step

DNA is built from a sequence of nucleotides, where each nucleotide consists of a phosphate, a sugar called deoxyribose, and one of the four bases A, T, G, C (Adenine, Thymine, Guanine, and Cytosine). The two ends of a DNA molecule are not the same. One end always has a 5' carbon atom of the deoxyribose exposed and the other end has a 3' carbon, so we will use these names to "bracket" the sequence of bases, always going from the 5' end to the 3' end. (See "deoxyribose" in the glossary.)

A simple mechanism for building these sequences can be defined this way. First, we use the following notation to indicate that each single base can form a DNA molecule by itself.

DNA parts:	5'	3'	a	t	g	c
	Start	End	Base	Base	Base	Base

To build any longer sequence, let's indicate that we can extend any start sequence x with any base y to get the longer molecule xy:

DNA-rule0:	x	y	$\mapsto$	xy
	Start	Base		Start

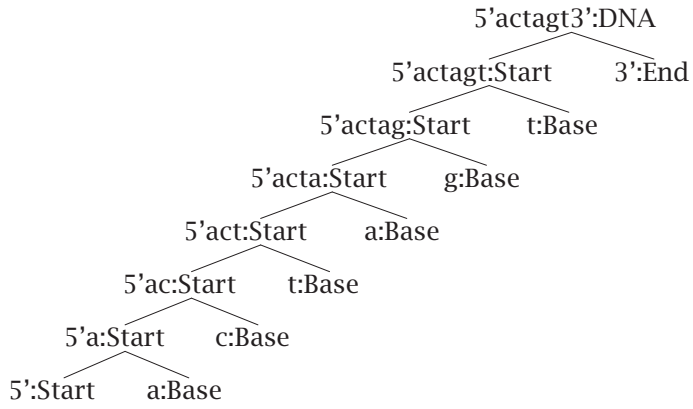
Notice that this structure building rule takes the Start of a molecule to build a longer Start. Rules like this are called **recursive**, because they build complexes that have parts of the same kind inside of them. Here, a Start complex appears on the left side of the rule and the right side too, so the rule is recursive.

We also allow the molecule to be ended at any point, completing a DNA molecule, with the 5' carbon:

DNA-rule1:	x	y	$\mapsto$	xy
	Start	End		DNA

Notice that this rule, rule1, is not recursive, because it completes a DNA molecule, and no DNA molecule has a complete DNA molecule as a part. So DNA appears on the right side of the rule but not the left side.

With the 6 basic parts and the rules DNA-Rule0 and DNA-Rule1, we can build any DNA sequence. The sequence of steps involved in building any sequence, like $5'actagt3'$, can be indicated with a **derivation tree**, like this:



Notice that we start at the bottom, by first combining the 5' Start and the Base a (rule0), to get the Start 5'a. Then we add the Base c to get the new Start 5'ac (using rule0), and so on until we finish with rule1. (A DNA molecule having this sequence will also have a paired, complementary strand: 5'$tgatca$3'.)

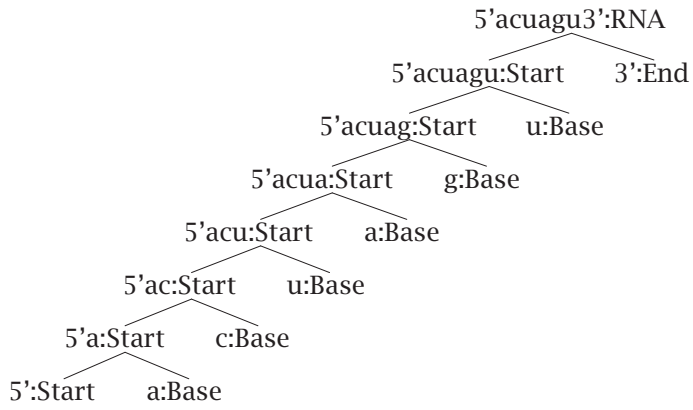
2.5.2 The language of RNA: first step

RNA sequences are like DNA sequences, except we have Uracil (U) instead of Thymine (T). So all the RNA sequences can be constructed from the following basic elements and rules of combination:

RNA Parts:	5'	3'	a	u	g	c
	Start	End	Base	Base	Base	Base

RNA-rule0:	x	y	$\mapsto$	xy
	Start	Base		Start
RNA-rule1:	x	y	$\mapsto$	xy
	Start	End		RNA

This works exactly like the mechanism for defining DNA sequences, except that we have u instead of t:



2.5.3 The language of proteins: first step

Proteins are polypeptides (PP), built from the 20 different amino acids (AA). The amino acids were listed on page 12. Every polypeptide has is bracketed by a free "amino group" called the "N-terminus" (N-) and a free "carboxyl group" called the "C-terminus" (-C) so we specify how to build a protein (PP) as follows

PP Parts:

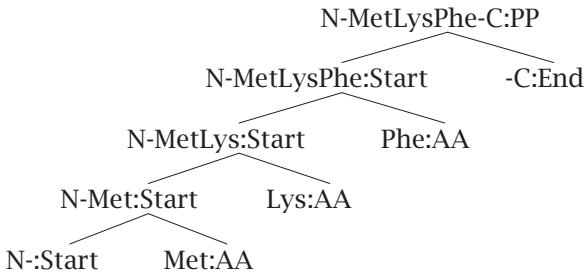
N-	-C	Ala	Asp	Phe	His	Lys	Met	Pro	Arg	Thr	Trp
Start	End	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA

Cys	Glu	Gly	Ile	Leu	Asn	Gln	Ser	Val	Tyr
AA	AA	AA	AA	AA	AA	AA	AA	AA	AA

PP-rule0: $\begin{matrix} x & y \\ \text{Start} & \text{AA} \end{matrix} \mapsto \begin{matrix} xy \\ \text{Start} \end{matrix}$

PP-rule1: $\begin{matrix} x & y \\ \text{Start} & \text{End} \end{matrix} \mapsto \begin{matrix} xy \\ \text{PP} \end{matrix}$

On the first pages of this chapter, we looked at the RNA that specifies the sequence of proteins *Met Lys Phe*. Putting the N- at the beginning and the -C at the end, the sequence is N- Met Lys Phe -C. We can derive this using the rules given, and show the derivation in a tree, like this:



As in the earlier trees, the construction of the protein begins at the bottom, with the first combination of the Start N- and the amino acid Met (rule0). The result is N-Met, and this gets extended with the amino acid Lys (rule0), and so on until we attach the End -C (rule1).

2.5.4 Structures in the language of DNA, RNA, Proteins: copies and secondary structure

Here is part of the genetic sequence from chromosome 16 of the human genome, called the human alpha globin gene cluster. This listing puts a space in after every 10 bases, and it puts a line break and a number after every 60 characters, just to make it easier to refer to particular elements of this sequence. Obviously, the sequence is in the DNA language defined in section §2.5.1 above, but does it have any other special properties?

```
LOCUS       HUMHBA4               12847 bp    DNA     linear   PRI 13-APR-2001
DEFINITION  Human alpha globin psi-alpha-1, alpha-2 and alpha-1 genes, complete
VERSION     J00153.1  GI:183793
SEGMENT     4 of 4
SOURCE      Homo sapiens (human)
ORIGIN
```

```
     1  ggatccccgg  ggctctgggc  ggtgtgggcg  tagtgaagcc  ccacgcagcc  gccctcctcc
    61  ccggtcactg  actggtcctg  caggctcttc  acggtgtacc  ccagcaccaa  ggtctacttc
   121  ccgcacctga  gcgcctgcca  ggacgacgca  gctgctgagc  cacgggagcg  catctgcggc
   181  tgtggcgcgg  cggtgcagca  cgtggacaac  ctgcgcgcct  gagcccgctg  gcggacctga
   241  cgctcgttgc  gcgtggaccc  agccaacttt  ccggtgaggc  ctttccggcc  ggggcaatgg
   301  tgcatcgcct  agccgggatg  gggggctct   gggggtccct  agcggggcag  accccgtctc
   361  accggccccct tctcctgcag  ctgctaatcc  agtgtttcca  cgtcgtgctg  gcctcccacc
   421  tgcaggacga  gttcaccgtg  caaatgcaag  cggcgtggga  caagttcctg  actggtgtgg
   481  ccgtggtgct  gaccgaaaaa  tacgctgagc  cctgtgctgc  gaggccttgg  tctgtgcatg
   541  tcaataaaca  gaggcccgaa  ccatctgccc  ctgcctgtgt  ggtctttggg  gagctagcaa
   601  agcgaggtca  ctattgttgg  ccagtaagct  cagggaccta  aagggagcct  cctagaactc
   661  tcaaatgcgc  cccacccccg  gaggtttgtc  ctcccatggc  gaggagtgcg  atggggcaga
   721  gggagcagtg  tgatatggcg  ggggtagaga  gggtggcctt  cgacttcaaa  cccttgactc
   781  gggcttcgaa  ccatactcgt  tcgcaaagca  gttccccatt  catgcattta  ttcagttcat
   841  tccttccctc  catccccatt  tcctgctggg  acctgtagat  gctaatcctg  gccctttttg
   901  cagagagatg  cagaaactga  ggtcccagag  ccaaatgtgc  aacctaattc  gttggccaga
   961  gcagagggcc  gcagacctgt  tcctttcccc  ttccttcccc  catggacact  tcctcagtgg
  1021  caaacctgcg  ctagcctggt  tagccctccc  tgtgaccctg  cagccctggg  gatgaggtcg
  1081  ggaggaagac  ctcagtggcc  acaatttggc  agacagagag  gtttagtctt  ccagcctgct
  1141  caatgacaag  ctgtgcgacc  ctgggctgtc  ccagagctct  aggcctttac  ctatcgaata
  1201  gaaaaacagc  gtccaactca  tgagattttt  gaaataattt  ttgaaatcat  aacacagggt
  1261  gggtgcctgc  agggacgttg  ccaccccacc  cctccaccca  gccccagctg  ccgtgtctca
  1321  atctctgcag  gtgcccaggc  caaggcattc  ccttccccag  gctccctctt  ctccctcccc
  1381  aaggattggg  aagggaatct  tagggctcca  ccccaggctt  ttcagacaaa  gaataggggc
  1441  tcaggaaaga  ttgggacctt  ggagttctcc  aatccctaat  agggttgggt  gtgggttggg
  1501  catcctgggt  gtgtgtgggg  agcacctgga  ccaggcctgg  cacccaggtc  tgacctggca
  1561  gtcagcaatg  aggtctgaag  agagctgctg  gaagtggagc  cctgactgtg  agtcggccaa
  1621  actccccca  gcagtcagtg  ccacagacct  gttgccctgc  actgcctggg  accccagccc
  1681  ggtagtttgg  agaacttggc  ccctcgttat  ctacatcccc  caagtgtttt  tttgtttttg
  1741  ggggtttttt  ttttttttt   tttgctttgt  ttttgttttt  gagataggcc  cttgctctga
  1801  caccccggct  ggagtgcagt  ggcaagtttt  ggctcactgc  agcctcaacc  tcctgggttc
  1861  aagcgattct  cctgcctctg  tctcccgtgt  agctgggatt  acaggcatgg  gccgccattc
  1921  ctggctaatt  tatgtatttt  taatagagac  acagtttcac  catgttgatc  aggctggtct
  1981  caaactcctg  acctcaagtg  atctgccctc  ctggtctccc  aaagtgctgg  gatgacaggc
```

```
2041  gtgagccacc  acacccagcc  cccgcaactg  tttacatgga  taattaacaa  gctttttgtc
2101  ccaggcagag  tttggtgtga  aagcagctta  tgtttcactt  tggaaaaact  gtgctcttct
2161  ccccatccag  gaagctgcct  gggtctgggc  catatgtgga  taccttatgg  gtataagctg
2221  ctcaggaccc  tgtgtggaag  ctcaggacaa  tgccagcggg  aaggctacca  tgtggagagc
2281  tgtctctgtt  tgggcaggac  taagagacgc  agggaacctt  gggaacctgt  ctactctcac
2341  tcactcctcc  tccccctttcc  ttccaggcac  ctctgcaact  tgccagccaa  tgaccctgca
2401  tcccaggcat  aagagctcct  actctccccc  acctttcact  tttgagctta  cacagactca
2461  gaaattaagc  tgccgtggtg  ctgtctcctg  aggacaaggc  taacaccaag  gcggtctggg
2521  agaaagttgg  cgaccacact  gctggctatg  ccacggaggc  cctggagagg  caagaaccct
2581  cctctccctg  ctcacacctt  gggtccaacg  cccactccag  ggctccactg  gccaccccta
2641  actattctta  ccctggaccc  agcccccagc  ccctcactct  ttgcttcccc  ctgaagcatg
2701  ttcctgacct  tcctctcact  tggccctgag  ttatggctca  gcccagatca  agaaacaatg
2761  caagtaggtg  gccgacacgc  tgaccaatgc  cgtggtccac  ttagatgaca  tgcccaatga
2821  tgtgtctgag  gtgaggaagc  tgcatgtcca  cgagctgtgg  gtggacccag  gcaacatcag
2881  ggagagcttt  gggctgggag  gaatctaggg  tgtgggggca  gctggccttc  ctcataggac
2941  agaccctccc  acgcgttcag  ggaggtggag  cacaggtggc  agtagtatct  gcatcccctg
3001  actctctctc  cacagttcct  gggtaaatgc  ctgctggtga  cctaggcctg  ccacaccctt
3061  cccggtttac  ccatgtggtg  cctccatgga  caaattattt  gcttttgtga  gtgctgtgtt
3121  gacctaaaaa  caccattaag  ctagagcatt  ggtggtcatg  ccccctgcct  gctgggcctc
3181  ccaccaggcc  cgcctcccct  ccctgcccca  gcacttcctg  atctttgaat  gaagtccgag
3241  taggcagcag  cctgtgtgtg  cctgggttct  ctctgtcccg  gaatgtgcca  acagtggagg
3301  tgtttacctg  tctcagacca  aggacctctc  tgcagctgca  tggggctggg  gagggagaac
3361  tgcagggagt  atgggagggg  aagctgaggt  gggcctgctc  aagagaaggt  gctgaaccat
3421  cccctgtcct  gagaggtgcc  aggcctgcag  gcagtggctc  agaagctggg  gaggagagag
3481  gcatccaggg  ttctactcag  ggagtcccag  catcgccacc  ctcctttgaa  atctccctgg
3541  ttgaacccag  ttaacatacg  ctctccatca  aaacaaaacg  aaacaaaaca  aactagcaaa
3601  ataggctgtc  cccagtgcaa  gtgcaggtgc  cagaacattt  ctctcattcc  caccccttcc
3661  tgccagaggg  taggtggctg  gagtgagggt  gctggcccta  ctcacacttc  ctgtgtcatg
3721  gtgaccctct  gagagcagcc  cagtcagtgg  ggaaggagga  aggggctggg  atgctcacag
3781  ccggcagccc  acacctaggg  agactcttca  gcagagcacc  ttgcggcctt  actcctgcac
3841  gtctcctgca  gtttgtaagg  tgcattcaga  actcactgtg  tgcccagccc  tgagctccca
3901  gctaattgcc  ccacccaggg  cctctgggac  ctcctggtgc  ttctgcttcc  tgtgctgcca
3961  gcaacttctg  gaaacgtccc  tgtccccggt  gctgaagtcc  tggaatccat  gctgggaagt
4021  tgcacagccc  atctggctct  cagccagcct  aggaacacga  gcagcacttc  cagcccagcc
4081  cctgccccac  agcaagcctc  cccctccaca  ctcacagtac  tgaattgagc  tttgggtagg
4141  gtggagagga  ccctgtcacc  gcttttcttc  tggacatgga  cctctctgaa  ttgttgggga
4201  gttccctccc  cctctccacc  acccactctt  cctgtgcctc  acagcccaga  gcattgttat
4261  ttcaacagaa  acactttaaa  aaataaacta  aaatccgaca  ggcacggtgg  ctcacacctg
4321  taatcccagt  actttgggag  gctgaggcga  gaggatcacc  tgaggtcggg  agtttgagac
4381  cagcctgacc  aatatggaga  aaccccagtt  atactaaaaa  tacaaaatta  gctgggtgtg
4441  gtggcgcatg  cctgtaatcc  tagctactag  gaaggctgag  gcaggagaat  cgcttgaacc
4501  cgggaggtgg  aggttgaggt  gagccgagat  cacgccattg  cactccagcc  tgggcaacaa
4561  gagcaaaact  ccgtctcaaa  aaataaataa  ataaataaat  aaataaacta  aaatctatcc
4621  atgctttcac  acacacacac  acacacacac  acacacacct  tttttgtgtt  actaaagtag
4681  gagagtgtct  ctctttcctg  tctcctcaca  cccacccccа  gaagagacca  aaatgaaggg
4741  tttggaactc  acgccatggg  ccccatccca  tgctgaggga  acacagctac  atctacaact
4801  actgccacag  cgtctctttt  tggacacccc  taccatcata  ctgtagatac  ccgtgtacaa
4861  ccttcctatt  ctcagtgaag  tgtctccccт  gcatcccttt  cagccagttc  attcagctct
```

```
4921  gctcgcccat  tccacagtct  cactgattat  tactatgttt  ccatcatgat  ccccccaaaa
4981  aatcatgact  ttatttttt   atttttatta  ttattattat  ttttttttt   ttttttgaga
5041  cggagtctcg  ctctgtgacc  caggctggag  tgcagtggca  aatctcggct  cactgcaagc
5101  tccacctcgc  aggttcacgc  cattctcctc  cctcagcctc  ccgagtcgct  gagtagctgg
5161  gctacagcgc  cccccactag  tcgtggctaa  ttttttcttt  ttttaataga  gacagagttt
5221  cactgcatta  gcgaggatgg  tctcgatctc  ctgacctcgc  atctgccagc  ctcagccttc
5281  caatgtgctg  ggattacagc  gtgagccaac  gcgcccggcc  ttatatattt  atttttttga
5341  gacagagtct  cgctgtgtcg  tcaggctaga  gtgctgtggc  acgatctcgg  ctcactgcaa
5401  cctccaactc  cctggttcaa  aggattctcc  agcctccacc  tcccgagtag  ctgggattac
5461  aggcgtgcac  caccacacca  gctaattttt  gtatttttag  tagagacggg  gtttctccat
5521  gttggtcagc  ctggtctcga  actcccgacc  acagctgatc  ccacccacct  cggcctccca
5581  aagtgctggg  attccaggcg  tgcgccgagc  ctggccaaac  catcactttt  catgagcagg
5641  gatgcaccca  ctggactcct  ggacctccca  ccctcccccc  cgccaagtcc  acccccttcct
5701  tcctcacccc  acatcccctc  acctacattc  tgcaacacag  gggccttctc  tcccctgtcc
5761  tttccctacc  cagagccagg  tttgtttatc  tgtttacaac  cagtatttac  ctagcaagtc
5821  ttccatcaga  tagcatttgg  agagctgggg  gtgtcacagt  gaaccacgac  ctctaggcca
5881  gtgggagagt  cagtcacaca  aactgtgagt  ccatgacttg  gggcttagcc  agtacccacc
5941  accccacgcg  ccaccccaca  accccgggta  gaggagtctg  aatctggagc  cgccccccagc
6001  ccagccccgt  gctttttgcg  tcctggtgtt  tgttccttcc  cggtgcctgt  cactcaagca
6061  cactagtgac  tatcgccaga  gggaaaggga  gctgcaggaa  gcgaggctgg  agagcaggag
6121  gggctctgcg  cagaaattct  tttgagttcc  tatgggccag  ggcgtccggg  tgcgcgcatt
6181  cctctccgcc  ccaggattgg  gcgaagccct  ccggctcgca  ctcgctcgcc  cgtgtgttcc
6241  ccgatcccgc  tggagtcgat  gcgcgtccag  cgcgtgccag  gccggggcgg  gggtgcgggc
6301  tgactttctc  cctcgctagg  gacgctccgg  cgcccgaaag  gaaagggtgg  cgctgcgctc
6361  cggggtgcac  gagccgacag  cgcccgaccc  caacgggccg  gccccgccag  cgccgctacc
6421  gccctgcccg  ggcgagcggg  atgggcggga  gtggagtggc  gggtggaggg  tggagacgtc
6481  ctggcccccg  ccccgcgtgc  acccccaggg  gaggccgagc  ccgccgcccg  gccccgcgca
6541  ggccccgccc  gggactcccc  tgcggtccag  gccgcgcccc  gggctccgcg  ccagccaatg
6601  agcgccgccc  ggccgggcgt  gcccccgcgc  cccaagcata  aaccctggcg  cgctcgcggc
6661  ccggcactct  tctggtcccc  acagactcag  agagaaccca  ccatggtgct  gtctcctgcc
6721  gacaagacca  acgtcaaggc  cgcctggggt  aaggtcggcg  cgcacgctgg  cgagtatggt
6781  gcggaggccc  tggagaggtg  aggctccctc  ccctgctccg  acccgggctc  ctcgcccgcc
6841  cggacccaca  ggccacccctc  aaccgtcctg  gccccggacc  caaaccccac  ccctcactct
6901  gcttctcccc  gcaggatgtt  cctgtccttc  cccaccacca  agacctactt  cccgcacttc
6961  gacctgagcc  acggctctgc  ccaggttaag  ggccacggca  agaaggtggc  cgacgccctg
7021  accaacgccg  tggcgcacgt  ggacgacatg  cccaacgcgc  tgtccgccct  gagcgacctg
7081  cacgcgcaca  agcttcgggt  ggacccggtc  aacttcaagg  tgagcggcgg  gccgggagcg
7141  atctgggtcg  aggggcgaga  tggcgccttc  ctctcagggc  agaggatcac  gcgggttgcg
7201  ggaggtgtag  cgcaggcggc  ggctgcgggc  ctgggccgca  ctgaccctct  tctctgcaca
7261  gctcctaagc  cactgcctgc  tggtgaccct  ggccgcccac  ctccccgccg  agttcacccc
7321  tgcggtgcac  gcctccctgg  acaagttcct  ggcttctgtg  agcaccgtgc  tgacctccaa
7381  ataccgttaa  gctggagcct  cggtagccgt  tcctcctgcc  cgatgggcct  cccaacgggc
7441  cctcctcccc  tccttgcacc  ggcccttcct  ggtctttgaa  taaagtctga  gtgggcggca
7501  gcctgtgtgt  gcctgggttc  tctctgtccc  ggaatgtgcc  aacaatggag  gtgtttacct
7561  gtctcagacc  aaggacctct  ctgcagctgc  atggggctgg  ggagggagaa  ctgcagggag
7621  tatgggaggg  gaagctgagg  tgggcctgct  caagagaagg  tgctgaacca  tcccctgtcc
7681  tgagaggtgc  caggcctgca  ggcagtggct  cagaagctgg  ggaggagaga  ggcatccagg
7741  gttctactca  gggagtccca  gcatcgccac  cctcctttga  aatctccctg  gttgaaccca
```

```
 7801  gttaacatac  gctctccatc  aaaacaaaac  gaaacaaaac  aaactagcaa  aataggctgt
 7861  ccccagtgca  agtgcaggtg  ccagaacatt  tctctcattc  ccaccccttc  ctgccagagg
 7921  gtaggtggct  ggagtgaggg  tgctggccct  actcacactt  cctgtgtcac  ggtgaccctc
 7981  tgagagcagc  ccagtcagtg  gggaaggagg  aaggggctgg  gatgctcaca  gccggcagcc
 8041  cacacctagg  gagactcttc  agcagagcac  cttgcggcct  tactcctgca  cgtctcctgc
 8101  agtttgtaag  gtgcattcag  aactcactgt  gtgcccagcc  ctgagctccc  agctaattgc
 8161  cccacccagg  gcctctggga  cctcctggtg  cttctgcttc  ctgtgctgcc  agcaacttct
 8221  ggaaacgtcc  ctgtccccgg  tgctgaagtc  ctggaatcca  tgctgggaag  ttgcacagcc
 8281  catctggctc  tcagccagcc  taggaacatg  agcagcactt  ccaacccagt  ccctgcccca
 8341  cagcaagcct  ccccctccac  actcacagta  ctggattgag  ctttggggag  ggtggagagg
 8401  accctgtcac  cgctttcctt  ctggacatgg  acctctctga  attgttgggg  agttccctcc
 8461  ccctctccac  cacccgctct  tcctgcgcct  cacagcccag  agcattgtta  tttcagcaga
 8521  aacactttaa  aaaataaact  aaaatccgac  aggcacggtg  gctcacgcct  gtaatcccag
 8581  cactttggga  ggccgaggtg  ggaggatcac  ctgaggtcgg  gagtttgaga  ccaccctgat
 8641  caacatgtag  aaaccccatc  tatactaaaa  atacaaaatc  agccgggcat  ggtggcccat
 8701  gcctgtaaac  ccacctactc  cggaggctga  ggcaggagaa  tcattttaac  caaggaggca
 8761  gaggttgcag  tgagctaaga  tcacaccatt  gcactccagc  ctggaaaaca  acagcgaaac
 8821  tccgcctcaa  aaaaaaaaaa  gcccccacat  cttatctttt  ttttttcctt  caggctgtgg
 8881  gcagagtcag  aaagtcagaa  gagggtggca  gacagggagg  ggaaatgaga  agatccaacg
 8941  ggggaagcat  tgctaagctg  gtcggagcta  cttccttctc  tgcccaaggc  agcttaccct
 9001  ggcttgctcc  tggacaccca  gggcagggcc  tgagtaaggg  cctggggaga  cagggcaggg
 9061  agcaggctga  agggtgctga  cctgatgcac  tcctcaaagc  agatcttctg  ccagaccccc
 9121  aggaaatgac  ttatcagtga  tttctcaggc  tgttttctcc  tcagtaccat  ccccccaaaa
 9181  aacatcactt  ttcatgcaca  gggatgcacc  cactggcact  cctgcacctc  ccacccttcc
 9241  ccagaagtcc  acccccttcct  tcctcaccct  gcaggagctg  gccagcctca  tcaccccaac
 9301  atctccccac  ctccattctc  caaccacagg  gcccttgtct  cctctgtcct  ttcccctccc
 9361  cgagccaagc  ctcctccctc  ctccacctcc  tccacctaat  acatatcctt  aagtctcacc
 9421  tcctccagga  agccctcaga  ctaaccctgg  tccccttgaa  tgcctcgtcc  acacctccag
 9481  acttcctcag  ggcctgtgat  gaggtctgca  cctctgtgtg  tacttgtgtg  atggttagag
 9541  gactgcctac  ctcccagagg  aggttgaatg  ctccagccgg  ttccagctat  tgctttgttt
 9601  acctgtttaa  ccagtattta  cctagcaagt  cttccatcag  atagcatttg  gagagctggg
 9661  ggtgtcacag  tgaaccacga  cctctaggcc  agtgggagag  tcagtcacac  aaactgtgag
 9721  tccatgactt  ggggcttagc  cagcacccac  caccccacgc  gccaccccac  aaccccgggt
 9781  agaggagtct  gaatctggag  ccgcccccag  cccagccccg  tgctttttgc  gtcctggtgt
 9841  ttattccttc  ccggtgcctg  tcactcaagc  acactagtga  ctatcgccag  agggaaaggg
 9901  agctgcagga  agcgaggctg  gagagcagga  ggggctctgc  gcagaaattc  ttttgagttc
 9961  ctatgggcca  gggcgtccgg  gtgcgcgcat  tcctctccgc  cccaggattg  ggcgaagccc
10021  tccggctcgc  actcgctcgc  ccgtgtgttc  cccgatcccg  ctggagtcga  tgcgcgtcca
10081  gcgcgtgcca  ggccggggcg  ggggtgcggg  ctgactttct  ccctcgctag  ggacgctccg
10141  gcgcccgaaa  ggaaagggtg  gcgctgcgct  ccggggtgca  cgagccgaca  gcgcccgacc
10201  ccaacgggcc  ggccccgcca  gcgccgctac  cgccctgccc  gggcgagcgg  gatgggcggg
10261  agtggagtgg  cgggtggagg  gtggagacgt  cctggccccc  gccccgcgtg  cacccccagg
10321  ggaggccgag  cccgccgccc  ggccccgcgc  aggccccgcc  cgggactccc  ctgcggtcca
10381  ggccgcgccc  cgggctccgc  gccagccaat  gagcgccgcc  cggccgggcg  tgcccccgcg
10441  ccccaagcat  aaaccctggc  gcgctcgcgg  cccggcactc  ttctggtccc  cacagactca
10501  gagagaaccc  accatggtgc  tgtctcctgc  cgacaagacc  aacgtcaagg  ccgcctgggg
10561  taaggtcggc  gcgcacgctg  gcgagtatgg  tgcggaggcc  ctggagaggt  gaggctccct
10621  cccctgctcc  gacccgggct  cctcgcccgc  ccggacccac  aggccaccct  caaccgtcct
10681  ggccccggac  ccaaacccca  cccctcactc  tgcttctccc  cgcaggatgt  tcctgtcctt
```

```
10741  ccccaccacc  aagacctact  tcccgcactt  cgacctgagc  cacggctctg  cccaggttaa
10801  gggccacggc  aagaaggtgg  ccgacgccct  gaccaacgcc  gtggcgcacg  tggacgacat
10861  gcccaacgcg  ctgtccgccc  tgagcgacct  gcacgcgcac  aagcttcggg  tggacccggt
10921  caacttcaag  gtgagcggcg  ggccgggagc  gatctgggtc  gaggggcgag  atggcgcctt
10981  cctcgcaggg  cagaggatca  cgcgggttgc  gggaggtgta  gcgcaggcgg  cggctgcggg
11041  cctgggccct  cggccccact  gaccctcttc  tctgcacagc  tcctaagcca  ctgcctgctg
11101  gtgaccctgg  ccgcccacct  ccccgccgag  ttcacccctg  cggtgcacgc  ctccctggac
11161  aagttcctgg  cttctgtgag  caccgtgctg  acctccaaat  accgttaagc  tggagcctcg
11221  gtggccatgc  ttcttgcccc  ttgggcctcc  ccccagcccc  tcctcccctt  cctgcacccg
11281  taccccctg   gtctttgaat  aaagtctgag  tgggcggcag  cctgtgtgtg  cctgagtttt
11341  ttccctcaga  aacgtgccag  catgggcgtg  gacagcagct  gggacacaca  tggctagaac
11401  ctctctgcag  ctggataggg  taggaaaagg  caggggcggg  aggaggggat  ggaggaggga
11461  aagtggagcc  accgcgaagt  ccagctggaa  aaacgctgga  ccctagagtg  ctttgaggat
11521  gcatttgctc  tttcccgagt  tttattccca  gacttttcag  attcaatgca  ggtttgctga
11581  aataatgaat  ttatccatct  ttacgtttct  gggcactctt  gtgccaagaa  ctggctggct
11641  ttctgcctgg  gacgtcactg  gtttcccaga  ggtcctccca  catatgggtg  gtgggtaggt
11701  cagagaagtc  ccactccagc  atggctgcat  tgatccccca  tcgttcccac  tagtctccgt
11761  aaaacctccc  agatacaggc  acagtctaga  tgaaatcagg  ggtgcggggt  gcaactgcag
11821  gccccaggca  attcaatagg  ggctctactt  tcacccccag  gtcaccccag  aatgctcaca
11881  caccagacac  tgacgccctg  gggctgtcaa  gatcaggcgt  ttgtctctgg  gcccagctca
11941  gggcccagct  cagcacccac  tcagctcccc  tgaggctggg  gagcctgtcc  cattgcgact
12001  ggagaggaga  gcggggccac  agaggcctgg  ctagaaggtc  ccttctccct  ggtgtgtgtt
12061  ttctctctgc  tgagcaggct  tgcagtgcct  ggggtatcag  agggagggtt  cccggagctg
12121  gtagccataa  agccctggcc  ctcaactgat  aggaatatct  tttattccct  gagcccatga
12181  atcacccttg  gtaaacacct  atggcaggcc  ctctgcctgc  gtttgtgatg  tccttcccgc
12241  agcctgtggg  tacagtatca  actgtcagga  agacggtgtc  ttcgttattt  catcaggaag
12301  aatggaggtc  tgacctaaag  gtagaaatat  gtcaaatgta  cagcagaggg  ctggttggag
12361  tgcagcgctt  tttacaatta  attgatcaga  accagttata  aatttatcat  ttccttctcc
12421  actcctgctg  cttcagttga  ctaagcctaa  gaaaaaatta  taaaaattgg  ccgggcgcgg
12481  tggctcacac  ctgtaattgc  agcactttgc  caggcttagg  caggtggatc  acctgaagtc
12541  aggggttcga  gaccagccta  gccaacatag  tgaaaccctg  tctctactaa  aaagacaaaa
12601  attgtccagg  tgtgatgact  catgcctgta  aacctggcac  tttgggaggc  ggaggttgta
12661  gtgagtcaag  atcgcgccat  cgcactccag  cttgggcaac  aagagcgaaa  ctctgtctca
12721  aaaaaaaatt  taatctaatt  taatttaatt  taaaaattag  cacggtggtt  gggcacagtg
12781  gcctcacgcc  tgtaatccca  gcactttggg  aagccaaggt  gggcagatca  caaggtcagg
12841  ggaattc
```

This sequence is 12847 characters long. With a little study, it becomes clear that this section of DNA is not just a random sequence! It has some quite remarkable properties. There are many structures and redundancies in here.

Reduplication

The most striking kind of non-random structure in this DNA sequence is repetition. We can find in this sequence (and in most parts of the genetic code) many repetitions of the same (or almost the same) sequence. In the sequence above, we find two occurrences of the same 1141-base sequence which we have bracketed here:

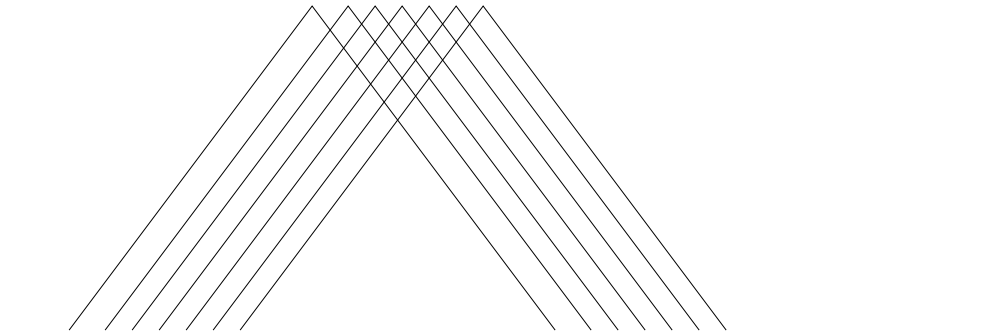
```
 6001  ccagccccgt  gctttttgcg  tcctggtgtt  tg[ttccttcc  cggtgcctgt  cactcaagca
 6061  cactagtgac  tatcgccaga  gggaaaggga  gctgcaggaa   gcgaggctgg  agagcaggag
 6121  gggctctgcg  cagaaattct  tttgagttcc  tatgggccag  ggcgtccggg  tgcgcgcatt
 6181  cctctccgcc  ccaggattgg  gcgaagccct  ccggctcgca  ctcgctcgcc  cgtgtgttcc
 6241  ccgatcccgc  tggagtcgat  gcgcgtccag  cgcgtgccag  gccggggcgg  gggtgcgggc
 6301  tgactttctc  cctcgctagg  gacgctccgg  cgcccgaaag  gaaagggtgg  cgctgcgctc
 6361  cggggtgcac  gagccgacag  cgcccgaccc  caacgggccg  gccccgccag  cgccgctacc
 6421  gccctgcccg  ggcgagcggg  atgggcggga  gtggagtggc  gggtggaggg  tggagacgtc
 6481  ctggcccccg  ccccgcgtgc  acccccaggg  gaggccgagc  ccgccgcccg  gccccgcgca
 6541  ggccccgccc  gggactcccc  tgcggtccag  gccgcgcccc  gggctccgcg  ccagccaatg
 6601  agcgccgccc  ggccgggcgt  gcccccgcgc  cccaagcata  aaccctggcg  cgctcgcggc
 6661  ccggcactct  tctggtcccc  acagactcag  agagaaccca  ccatggtgct  gtctcctgcc
 6721  gacaagacca  acgtcaaggc  cgcctggggt  aaggtcggcg  cgcacgctgg  cgagtatggt
 6781  gcggaggccc  tggagaggtg  aggctccctc  ccctgctccg  acccgggctc  ctcgcccgcc
 6841  cggacccaca  ggccaccctc  aaccgtcctg  gccccggacc  caaaccccac  ccctcactct
 6901  gcttctcccc  gcaggatgtt  cctgtccttc  cccaccacca  agacctactt  cccgcacttc
 6961  gacctgagcc  acggctctgc  ccaggttaag  ggccacggca  agaaggtggc  cgacgccctg
 7021  accaacgccg  tggcgcacgt  ggacgacatg  cccaacgcgc  tgtccgccct  gagcgacctg
 7081  cacgcgcaca  agcttcgggt  ggacccggtc  aacttcaagg  tgagcggcgg  gccgggagcg
 7141  atctgggtcg  aggggcgaga  tggcgccttc  ctc]tcagggc  agaggatcac  gcgggttgcg

 9841  tta[ttccttc  ccggtgcctg  tcactcaagc  acactagtga  ctatcgccag  agggaaaggg
 9901  agctgcagga  agcgaggctg  gagagcagga  ggggctctgc  gcagaaattc  ttttgagttc
 9961  ctatgggcca  gggcgtccgg  gtgcgcgcat  tcctctccgc  cccaggattg  ggcgaagccc
10021  tccggctcgc  actcgctcgc  ccgtgtgttc  cccgatcccg  ctggagtcga  tgcgcgtcca
10081  gcgcgtgcca  ggccggggcg  ggggtgcggg  ctgactttct  ccctcgctag  ggacgctccg
10141  gcgcccgaaa  ggaaagggtg  gcgctgcgct  ccggggtgca  cgagccgaca  gcgcccgacc
10201  ccaacgggcc  ggccccgcca  gcgccgctac  cgccctgccc  gggcgagcgg  gatgggcggg
10261  agtggagtgg  cgggtggagg  gtggagacgt  cctggccccc  gccccgcgtg  cacccccagg
10321  ggaggccgag  cccgccgccc  ggccccgcgc  aggccccgcc  cgggactccc  ctgcggtcca
10381  ggccgcgccc  cgggctccgc  gccagccaat  gagcgccgcc  cggccgggcg  tgcccccgcg
10441  ccccaagcat  aaaccctggc  gcgctcgcgg  cccggcactc  ttctggtccc  cacagactca
10501  gagagaaccc  accatggtgc  tgtctcctgc  cgacaagacc  aacgtcaagg  ccgcctgggg
10561  taaggtcggc  gcgcacgctg  gcgagtatgg  tgcggaggcc  ctggagaggt  gaggctccct
10621  cccctgctcc  gacccgggct  cctcgcccgc  ccggacccac  aggccaccct  caaccgtcct
10681  ggccccggac  ccaaacccca  cccctcactc  tgcttctccc  cgcaggatgt  tcctgtcctt
10741  ccccaccacc  aagacctact  tcccgcactt  cgacctgagc  cacggctctg  cccaggttaa
10801  gggccacggc  aagaaggtgg  ccgacgccct  gaccaacgcc  gtggcgcacg  tggacgacat
10861  gcccaacgcg  ctgtccgccc  tgagcgacct  gcacgcgcac  aagcttcggg  tggacccggt
10921  caacttcaag  gtgagcggcg  ggccgggagc  gatctgggtc  gaggggcgag  atggcgcctt
10981  cctc]gcaggg  cagaggatca  cgcgggttgc  gggaggtgta  gcgcaggcgg  cggctgcggg
```

Clearly, this long identical duplication is not here by chance! And there are other duplications besides this one too. What explains their presence?

The reasons for all the duplicated sequences in the genetic code are not fully understood. One well-supported proposal is that often long sequences all evolve at once, even when not all of the sequences are actually being used in protein synthesis. This is sometimes called

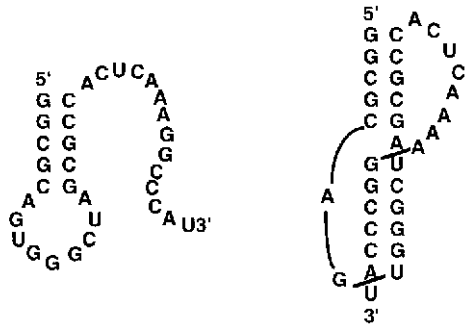
"concerted evolution." When there are repair mechanisms that tend to keep the sequences similar or identical, their presence can make sequences less susceptible to mutation.

Notice that a copied sequence involves "crossing dependencies", because the dependency between the first symbols in each sequence is crossed by the dependence for the second symbols, and so on:



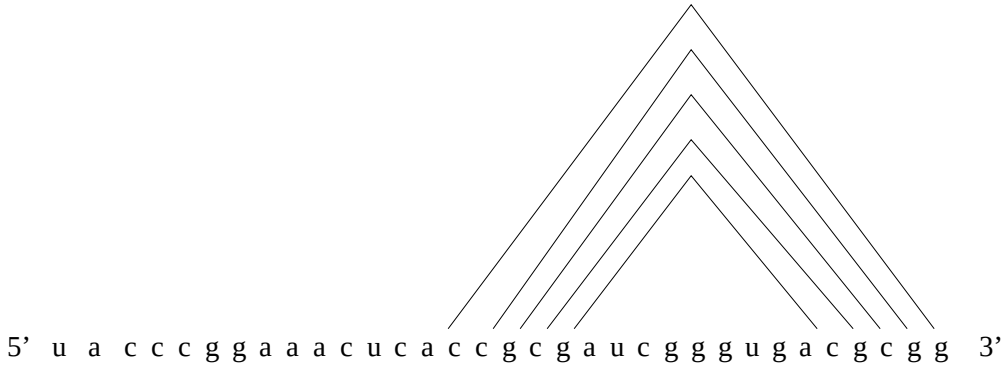
Nested sequences ("stem-loop" motifs)

There is another kind of non-randomness in the genetic sequence that is important, though it does not occur on the huge scale we see in the repetitions. Long RNA molecules fold up in distinctive ways that depend, in part, on attractions between the bases:



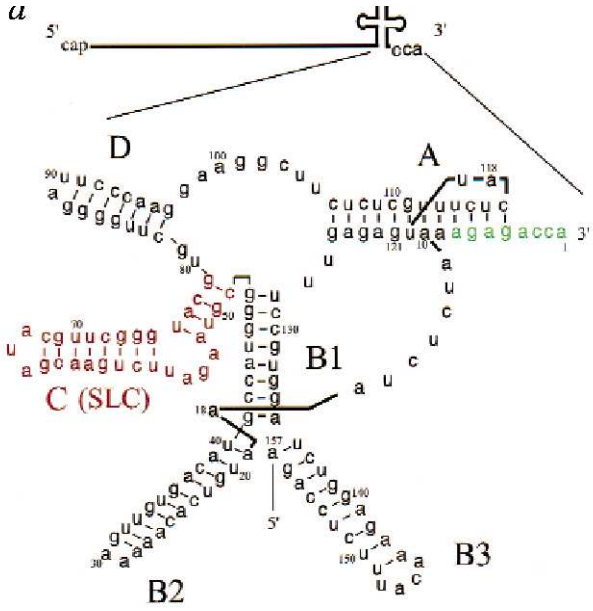
from (Tinoco and Bustamante, 2000)

In the diagram on the left, for example, we see that the attractions *g-c* have caused the molecule to curl up in a particular way. And in the more complex RNA on the right, we also see attractions between other elements: *g-u, u-a.* When this kind of "looping" occurs in the RNA molecules, one subsequence, like the *ccgcg* in the diagram on the left, predicts the *cgcgg* that comes later (going from the 5' end to the 3' end of the molecule). If we list the bases in a line, we can indicate these attracting pairs like this:



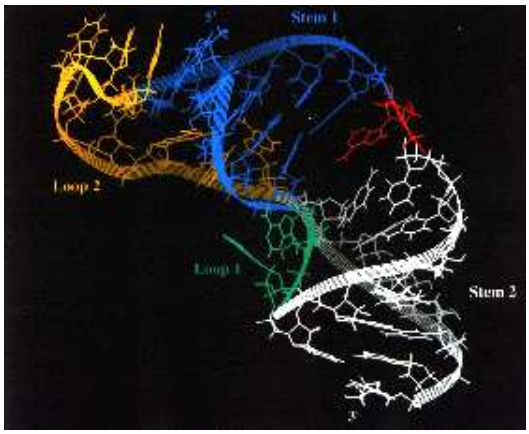
Notice that the attractions are “nested:” unlike the dependencies in duplications, each of these dependencies occurs properly inside another.

A single molecule can have many loops, as we see in this structure for example:



from (Kim, Kao, and Tinoco, 2000)

Each loop corresponds to a nested dependency in the sequence of bases. The actual 3D structure of these molecules is more complex than these diagrams indicate, of course. Here is a slightly more realistic image:

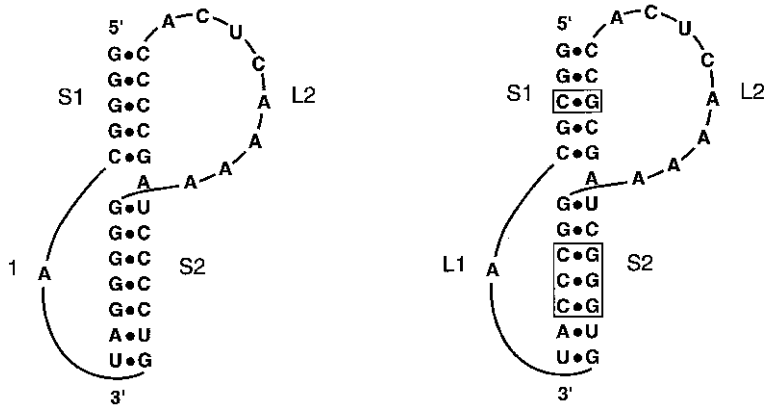


from (Shen and Tinoco, 1995)

These complex structures seem to have important effects on the timing and expression of effects of RNA on cellular processes.

Pseudoknots

More complicated non-random patterns can occur, too. One thing that happens quite commonly is that two or more nested sequences can be interleaved. This kind of pattern is called a pseudoknot:



from (Shen and Tinoco, 1995)

If we mark the two sequences with nested dependencies in the molecule on the left, one with brackets and one with parentheses, we see that the dependencies cross each other:

3'[uaggggg]aaaacuca(ccccg)a[uccccug]a(cgggg)5'

This kind of folding pattern is more complex than simple loops.

2.5.5 Defining nested dependencies

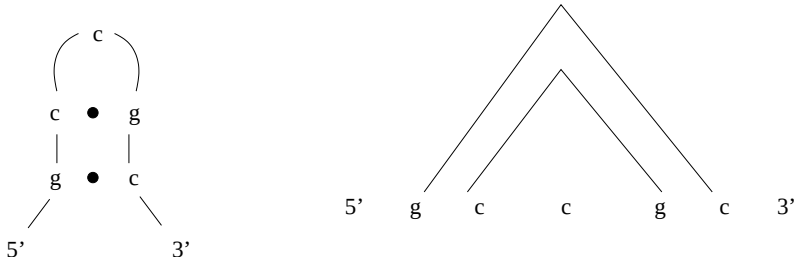
We can modify the simple mechanisms proposed above to provide a way to indicate the repetitions, loop sequences and knot sequences that we find in DNA and RNA. It is worth taking the time to do this here, partly because we will see a similar range of patterns in human languages later.[1]

We have seen that nested dependencies sometimes cause certain folding patterns in RNA, so it would be useful to have a description of the language of RNA and DNA which allows us to notice when these nested dependencies occur. The definitions of the languages of DNA and RNA given earlier are very simple: any beginning sequence can be extended to the right with any other base. When you look at the tree representation of the derivation of any molecule, we see that the molecule is tree only "branches on the left," since the molecule gets extended by adding bases on the right.

It turns out that there is no way to extend that kind of language to the analysis of nested dependencies. For nested dependencies, we want to be able to represent the attractions between pairs of bases that can be arbitrarily far apart in the base sequence. We will have to reorganize our grammar for RNA to get these.

The linguist Noam Chomsky noticed in 1956 that we can describe that kind of relation if we allow derivations that "branch in the middle," with rules that put a base on either side of the molecule that has been built up so far.

Suppose for example that we wanted to get the following nested loop (a simplified version of the one we looked at earlier), with the nested dependencies indicated by the dots on the left, and by the lines on the right:



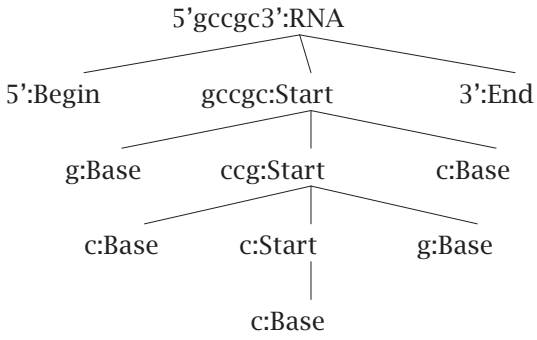
The following rules will let us derive this structure, with the assumption that the attracting

[1] Grammars for stem-loops are often proposed (since they are context free, as we will see later), but it is rarer to see grammars that can analyze pseudoknots and repetitions. See for example Sakakibara et al. (1994), Leung, Mellish, and Robertson (2001), Joshi (2002), and the overview in Searles (2002). Here I use multiple context free grammars(Seki et al., 1991), because they are so simple. (And MCFGs will be used again later.) Greg Kobele points out to me that Mirror Theory grammars, with their slightly more complex structure building mechanisms, apparently provide more elegant derivations of "pseudoknots" than MCFGs do (Brody, 2000; Kobele, 2002).

pairs are *g-c, g-u, u-a*:

RNA Parts:	5’	3’	a	u	g	c		
	Begin	End	Base	Base	Base	Base		
RNA-rule0:	x			$\mapsto$	x			**‘start with any base’**
	Base				Start			
RNA-rule1:	x	y		$\mapsto$	xy			**‘extend to the right’**
	Start	Base			Start			
RNA-rule2:	x	y		$\mapsto$	xy			**‘extend to the left’**
	Base	Start			Start			
RNA-rule3:	x	y	z	$\mapsto$	xyz			**if x & z are attracting**
	Base	Start	Base		Start			
RNA-rule4:	x	y	z	$\mapsto$	xyz			**‘add Begin & End’**
	Begin	Start	End		RNA			

One way to derive our molecule with these rules is this:



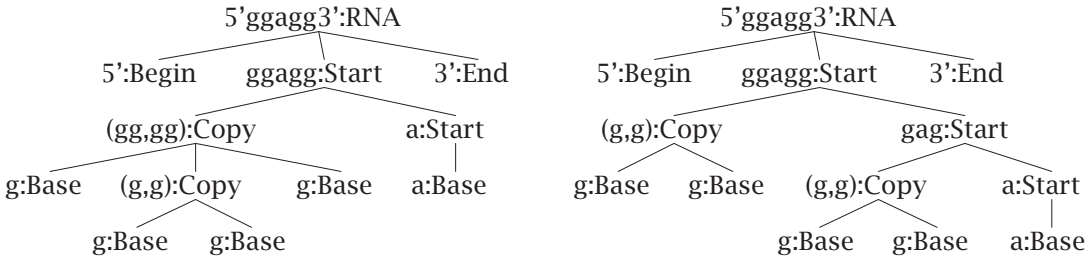
This tree shows the molecule construction beginning at the bottom of the tree too, starting with the Base *c*, changing it to our Start (rule0). Then the attracting pair *g,c* is added on either side of this Start (rule3); then another attracting pair *c,g* is added (rule3), and finally the Begin and End complete the molecule (rule4). This time the shape of the tree is quite different from our earlier derivations: the branching happens at the middle in this tree, so that related pairs can be added on either side.

2.5.6 Defining copy dependencies

Our first grammars for RNA and DNA on pages 56 and 57 were very simple, extending any sequence to the right. To get nested dependencies, we needed to allow extensions on the left and right together, with the definition in the previous section. It turns out that these cannot capture copy dependencies, because they are crossing. In order to capture those, we need to be able to build up the two identical sequences and then splice them together. So we introduce rules that build a pair of sequences x, y that we call a "Copy." The splicing rule, rule6, puts the two sequences of Copy together with any other Start sequence to give us a new Start sequence (with copies in it!): sequence:

RNA-rule5:	x Base	x Base	$\mapsto$	x,x Copy	**'start copies with any base'**
RNA-rule6:	x,x Copy	y Base	y Base	$\mapsto$ xy,xy Copy	**'extend copies with any base'**
RNA-rule7:	x,x Copy	y Start		$\mapsto$ xyx Start	**'splice copies together'**

With these rules, we have derivations like these for a sequence with a repeated *gg*:



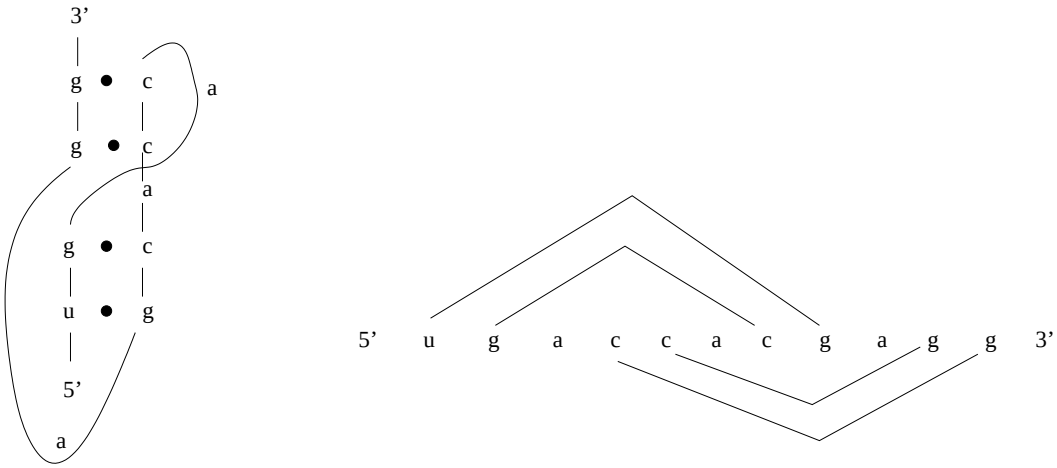
To understand the difference, consider each step of the derivations, beginning at the bottom of each tree. The derivation on the left starts copies with the pair (*g,g*), then extends it to get (*gg,gg*), and then splices them together with *a* in the middle. The derivation on the right starts with one copied sequence (*g,g*) and splices it together with *a* in the middle to form *gag*, and then starts another copied sequence (*g,g*) and splices this one around *gag*.

My intuition is that the first of these derivations, the one on the left, best captures our idea about the copied sequence. If we count each rule as taking a unit of energy, the derivation on the left requires less energy because it uses only one splice step, while the one on the right uses two. For this reason, there are 5 "vertices" or "internal nodes" in the tree on the left (not counting the "leaves"), but there are 6 on the right. The idea is that finding structure in the sequence should allow it to be derived with less energy.

2.5.7 Defining pseudoknots

The last kind of non-randomness we considered was more complex: two nested dependencies interleaved. Here too, the simple kinds of definitions that we used for unstructured sequences or for nesting simply cannot define the desired dependencies. We need patterns that are build pairs of sequences which are then spliced together, like we had for copy dependencies.

Consider the following knot, simplified from the one shown on page 67, with the dependencies indicated by the dots on the left, and by the lines on the right:



This structure can be generated if we add these rules to our mechanisms for RNA generation, and recognize the attracting pairs *u-g, g-c*:

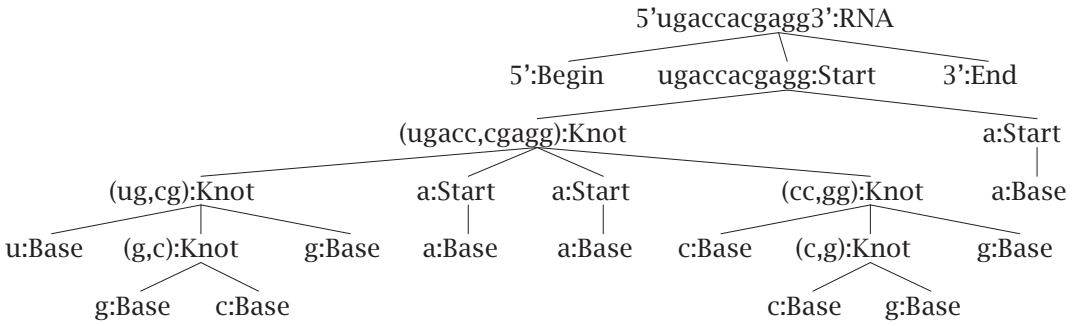
RNA-rule8: $\underset{\text{Base}}{x}\ \underset{\text{Base}}{y} \mapsto \underset{\text{Knot}}{x,y}$ **'start knot with any attracting** x, y**'**

RNA-rule9: $\underset{\text{Knot}}{w,x}\ \underset{\text{Base}}{y}\ \underset{\text{Base}}{z} \mapsto \underset{\text{Knot}}{yw,xz}$ **'extend knot with any attracting** x, z**'**

RNA-rule10: $\underset{\text{Knot}}{u,v}\ \underset{\text{Knot}}{w,x}\ \underset{\text{Start}}{y}\ \underset{\text{Start}}{z} \mapsto \underset{\text{Start}}{uyw,vzx}$ **'interleave 2 knots'**

RNA-rule11: $\underset{\text{Knot}}{x,y}\ \underset{\text{Start}}{z} \mapsto \underset{\text{Start}}{xzy}$ **'splice knots together'**

With these rules, we get derivations like this:



Notice that the dependencies in the knots recognized here are exactly the ones shown in the diagram above.

2.6 Summary

This section introduced some of the basic molecular perspective on heredity and evolution. We considered the basic structure of DNA, built from the bases G,C,T,A in a long sequence and with a complementary strand; the structure of RNA in a single strand, built from G,C,U,A; and the way that RNA specifies amino acid sequences (proteins) with triples of bases called codons. We observed that errors (mutations) can occur in replication, in transcription of DNA to RNA, and in translation from RNA to proteins. In HIV these errors are very frequent, but in higher organisms they are relatively rare.

Interpreting what all these sequences mean is one of the main projects of contemporary biology, and we did not dig into it in any depth. But we examined part of the human genome – a sequence of $\approx$3.1647 billion bases that is written into every cell of our bodies. We noticed certain kinds of structures in this sequence: repetitions, nested dependencies, and pseudoknots. The genetic code of organisms is very far from random!

Finally, we observed that different kinds of rules are needed to define these different non-random sequences. To define any sequence, we can just have a rule that extends any start sequence to the right. But to get nested dependencies, we need to be able to extend to the left and right. And to get duplications or pseudoknots, we need to be able to work on more than one piece at a time, splicing them together later. These ideas sound simple, but they are very important, classifying the different kinds of patterns in languages in a way that is absolutely fundamental. We will return to these last ideas when we consider the structures of human language.

Exercises

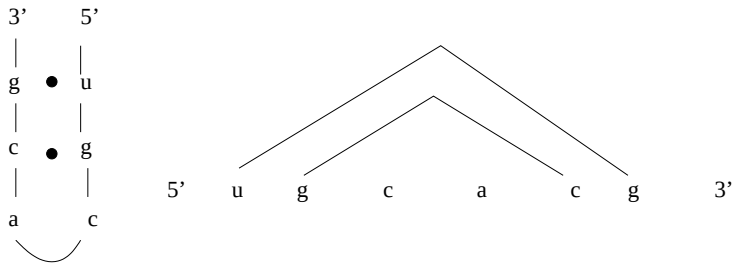
1. **Codon size:** Suppose that instead of 20 amino acids, there were 200. How long would codons have to be to provide a unique name for each of the 200 amino acids? (use the "naming rule")
2. **RNA → protein:** Name one sequence of RNA bases which is translated as the amino acid sequence:

   Leu Pro Val Gly

3. **Generating proteins:** Using the protein rules PP-rules0-1 on page 58, draw a tree representing the derivation of this protein sequence

   N- Leu Pro Val Gly -C

4. **Repetitions in DNA:** If the human genome were totally random, as if each base in the sequence were determined by throwing a fair, 4-sided die, how likely is it that you would get the repeated sequence shown on page 64 in 1141 throws of the die? (use the product rule)
5. **Generating nested dependencies:** Using the RNA-rules0-4 on page 69 draw a tree representing the derivation of this sequence with the indicated dependencies:



6. **HIV:** Two parts to this question. (i) A virus like HIV kills its host quite shortly after the immune system breaks down. Explain why this might seem to cause a problem for Darwinian natural selection theories (remember the first problem of HW1), and explain how a Darwinian might respond. (ii) Why do we find AZT-resistant HIV in patients who have never received AZT?
7. **Codons: RNA ⇒ Protein** The SARS virus begins with the following sequence of nucleotides:

   ATATTAGGTTTTTACCTACCCAGG

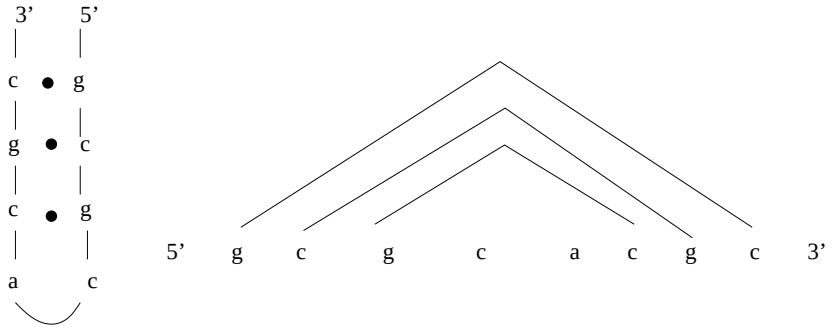
   If this whole sequence is translated into the RNA sequence

   AUAUUAGGUUUUUACCUACCCAGG

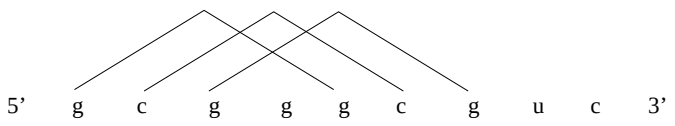
   and then translated into a sequence of amino acids, what amino acid sequence is it?
8. **Mechanisms to define the sequences:** Using the protein rules PP-rules on page 58, draw a tree representing the derivation of this amino acid sequence (the one you figured out in question 1, assuming that it begins with N- and ends with -C):

9. **Repetitions in DNA:** If the human genome were totally random, as if each base in the sequence were determined by throwing a fair, 4-sided die, how likely is it that you would find exactly the same sequence (from question 1) repeated at a particular point? (use the product rule)

10. **Generating nested dependencies:** Using the RNA-rules on page 69 (and in the class handout) draw a tree representing the derivation of the following sequence with the indicated dependencies:



11. **Generating crossing dependencies:** Using the RNA-rules on page 70 (and in the class handout) draw a tree representing the derivation of the following sequence with the indicated dependencies:



12. **Viruses:** One of the puzzles about viruses is why they sometimes are 'dormant' for a period of time. Nowak and May (2000) write:

> Viruses populate the world between the living and the non-living. They themselves are not capable of reproduction, but if put into the right environment, they can manipulate a cell to generate numerous copies of themselves. 'Reproduce me!' is the essence of the virus, the message that the viral genome carries into the head-quarters of a cell – the nucleus. The viral genome manages to attract the attention of the workers in the cell, which are the various enzymes capable of copying and interpreting genetic information. These workers will read the viral message and follow its instruction; they will produce viral proteins and more copies of the viral genome. The cell may devote all its resources to produce new virus particles and die after everything has been turned into a virus. A swarm of new 'Reproduce me!' messages is then leaving the cell, searching for new targets.
>
> But viruses can also have more subtle targets than this. They may tell the host cell to reproduce them, but only at a slow rate not endangering the survival of the cell. They may enter a cell, insert their genetic material into the genome and be very quiet for a long time. Under specific circumstances they may become reactivated and demand their reproduction. Other viruses once inserted into the genome of the cell may induce the cell to divide thereby producing two *infected* daughter cells. Such viruses may drive their host cell into uncontrolled multiplication, and thereby cause cancer.

Is there a problem for Darwin here? In the last chapter of *Origin of Species* Darwin looks at a river bank and sees many organisms that have adapted to surviving with each other. Can you imagine situations in which a virus would be better adapted (=more likely to reproduce more) if it does not reproduce at all immediately, but only months or even years later? What might such a situation be?

Selected Solved Exercises

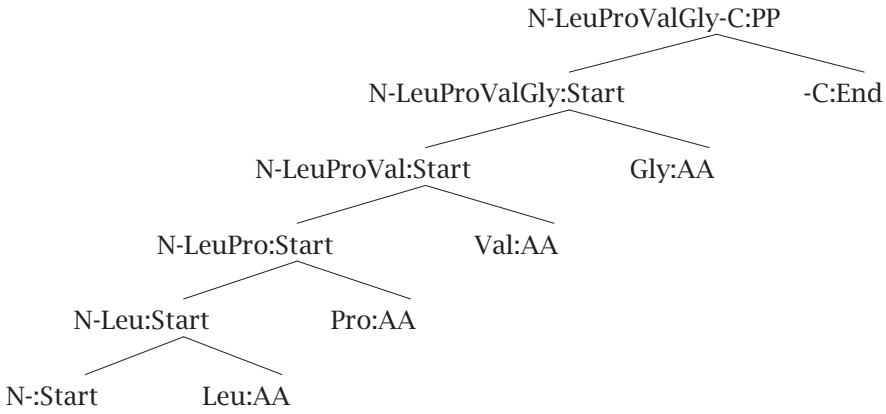
1. To name 200 amino acids with 4 bases, we need n digits where $4^n \geq 200$. Since $4^3 = 64$ and $4^4 = 256$, 4 digits, "quadruplets" of bases would suffice.

   Using logs, we calculate $n = \log_4 200 = \frac{\log_{10} 200}{\log_{10} 4} = 3.8219$, and rounding up we have 4.

2. We can just read these values out of the table in the notes, for example:

Leu	Pro	Val	Gly
cuu	ccg	gua	gga

3. The following derivation tree derives the polypeptide (PP) N-LeuProValGly-C:



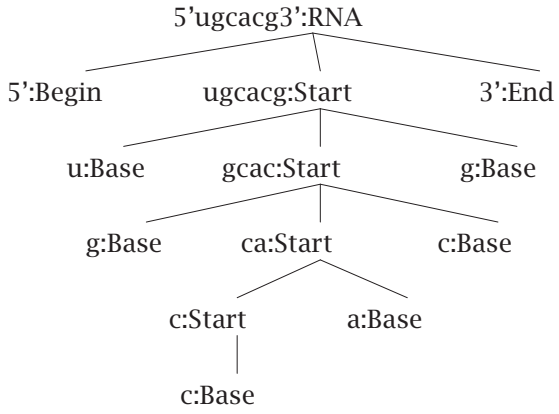
4. The answer is $(\frac{1}{4})^{1141}$. What number is that?

   My usual calculator says $(\frac{1}{4})^{500} = 9.3326 \times 10^{-302}$ but it will not provide a value for $(\frac{1}{4})^{1141}$ because it is too small.

   So I switched to another calculator and found $(\frac{1}{4})^{1141} \approx 1.1209 \times 10^{-687}$

   (There is no way that sequence of 1141 bases is repeated in human DNA by chance!)

5. The sequence 3-ugcacg-5 with the indicated nested dependencies can be generated this way:



6. i. The fact that HIV kills its host so quickly might seem problematic for a Darwinian theory: shouldn't an adaptive organism keep its host alive so it can reproduce more? But this is really not a problem. The HIV, like any other organism, does not need to live forever to survive and propagate. It can kill its host and survive as long as the HIV has replicated and been passed to other hosts before the host is killed. HIV shows that is is sometimes a very successful strategy to replicate quickly and spread to other hosts even though the burden of all the HIV kills the first host. The loss of the first host is not a problem once it has spread to many others!

   (It is no surprise that the evolution of **virulence** – the evolutionary strategy of damaging the host but surviving by spreading quickly – is of interest to epidemiologists!)

   ii. Since HIV spreads, we can assume it spreads from hosts that have already taken AZT. So it is no surprise at all that pretty soon, AZT-resistant HIV is common in the population.

Lecture 3

The neo-Darwinian synthesis

Recent biology has found a reasonably coherent synthesis of the Darwinian theory of natural selection, Mendelian classical genetics, and molecular genetics, briefly reviewed in the preceding pages. The simple, initial proposals of Darwin, Mendel, Watson & Crick have required modifications, but in their simple form they proved especially useful for the development of the field. Let's remind ourselves about some of the complications, many of them mentioned already.

3.1 Modeling populations: from complexity to chaos

As the quote from Gould at the beginning of lecture 1 warns, most of natural history is a science of the relative frequencies of genes and traits, and so it is no surprise that the theory of the genes and traits in populations of various sizes is complicated. The simple models of population growth and the selection pressure of fitness in the last lecture show that genetic populations can change quickly – sometimes with very dramatic changes in just a generation or two – but we did not really take a look at what happens when you try to move to calculate populations more carefully. Almost right away, there are some big surprises! We take a very brief look at some of those surprises now.

Remember the definition of the Fibonacci numbers (which number the petals of many flowers) in Lecture 0: each number is the sum of the previous two. These numbers are similar to population figures in the sense that each element depends on the previous one. This is obviously true in the case of population growth: the size and constitution of each generation depends on the previous one. We calculated populations in section 1.1 on page 19 on the assumption that the population grows at a fixed rate r:

Letting	s = population size at each generation
	r = rate of reproduction
we have:	$s(n) = r \times s(n-1)$
sometimes abbreviated this way:	$\Delta s = rs$
so when $s(0) = 1$ we have:	$s(n) = r^n$

This is completely optional, but if you have had some calculus, it is useful to remember how to relate the two equations just above. We can replace the discrete equation

$$\Delta s = rs$$

by the differential equation

$$\frac{ds}{dt} = rs.$$

where t is now continuous time, not the number of generations. Assuming $s, r \neq 0$, we can easily separate the variables:

$$\frac{ds}{s} = r\ dt.$$

Integrating both sides:

$$\int \frac{ds}{s} = \int r\ dt.$$

Since $\int \frac{dx}{x} = \log_e |x| + C$ for constant C,

$$\log_e |s| = rt + C.$$

Exponentiating both sides:

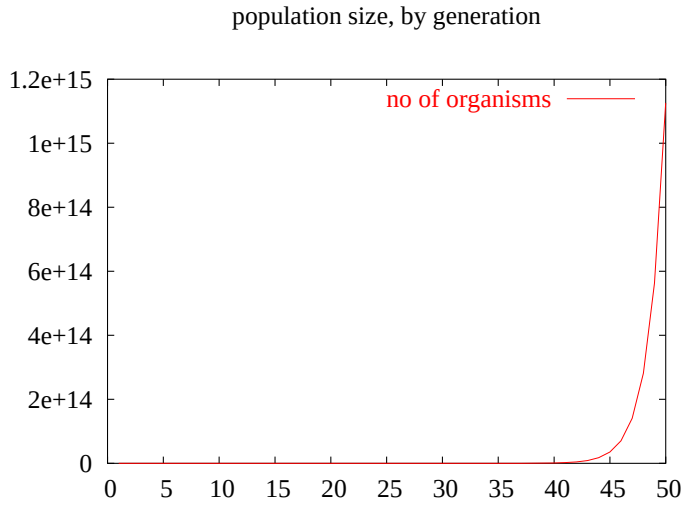
$$s = e^{rt+C} = e^C e^{rt}.$$

The constant e^C corresponds to the initial population size s_0, assumed here to be 1, so

$$s = e^{rt} = (e^r)^t,$$

which is exactly the equation on the previous page, when we just call the constant e^r simply r.

Drawing the geometric increase of population on a graph, we get a curve that increases ever faster, like this:



Obviously, this is not the way populations grow in nature, at least not for long! In any real

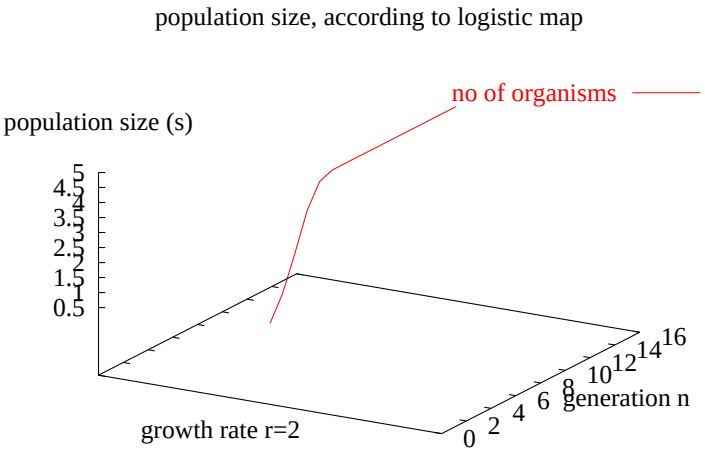
environment, as the population gets larger and larger, the reproductive value of each individual eventually has to decrease.

Suppose we assume that the population size increases with rate r when the population is 0, but as the population approaches some maximum value k, the rate of population growth decreases, getting to 0 when $r = k$. The following equation has this effect:

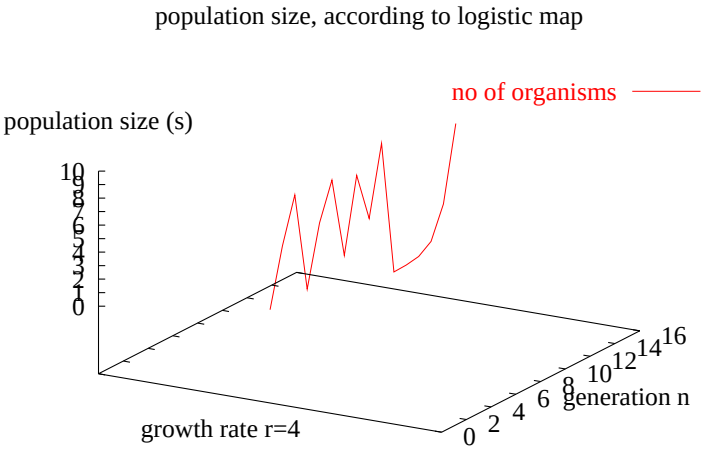
Letting	s = population size at each generation
	r = rate of reproduction
	k = "carrying capacity"
we have:	$s(n) = r \times s(n-1) \times (1 - \frac{s(n-1)}{k})$
sometimes abbreviated this way:	$\Delta s = rs(1 - \frac{s}{k})$

This is a very famous function, called the *discrete logistic function*, or the *logistic map*.

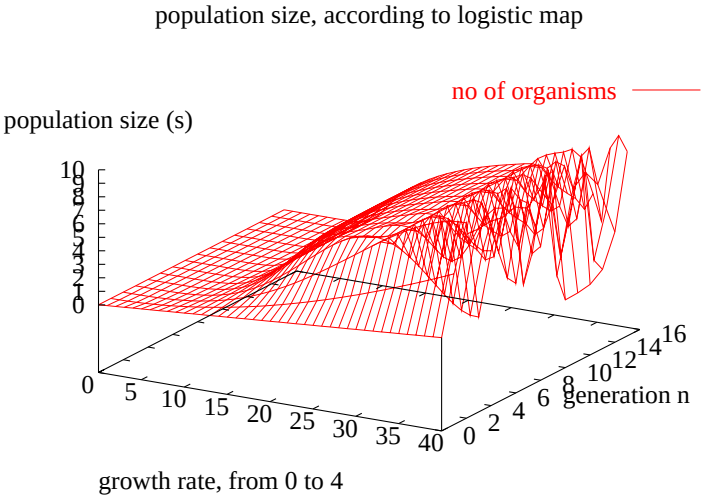
When we draw the population growth with this equation, what do we get? Well, when $k = 10$ and $r = 2$, we see that the population initially grows quite quickly, but then slows down:



But when $k = 10$ and $r = 4$, we see some kind of overshoot and readjustment, leading to oscillating populations size:



Looking at a range of values of r, keeping everything else fixed, we see that overshooting, unstable behavior occurs for many values of r:



The logistic map is one of the simplest functions that exhibits "chaotic" behavior, and it turns out that "chaotic" oscillations like this in population size are found in nature, as we see in the figure on page 83 from Kendall (1991)

We see in the graph above that when growth rates get much higher than mere replacement, the logistic map is extremely sensitive, with tiny variations producing drastically different results. This sensitivity is sometimes thought to lead to "butterfly effects." Can the flutter of a butterfly in Brazil cause a tornado in Texas? The idea, posed by Edward Lorenz in a famous 1976 paper is not as crazy as it might seem at first; relations like the one described by this function would not be at all surprising in the natural world.

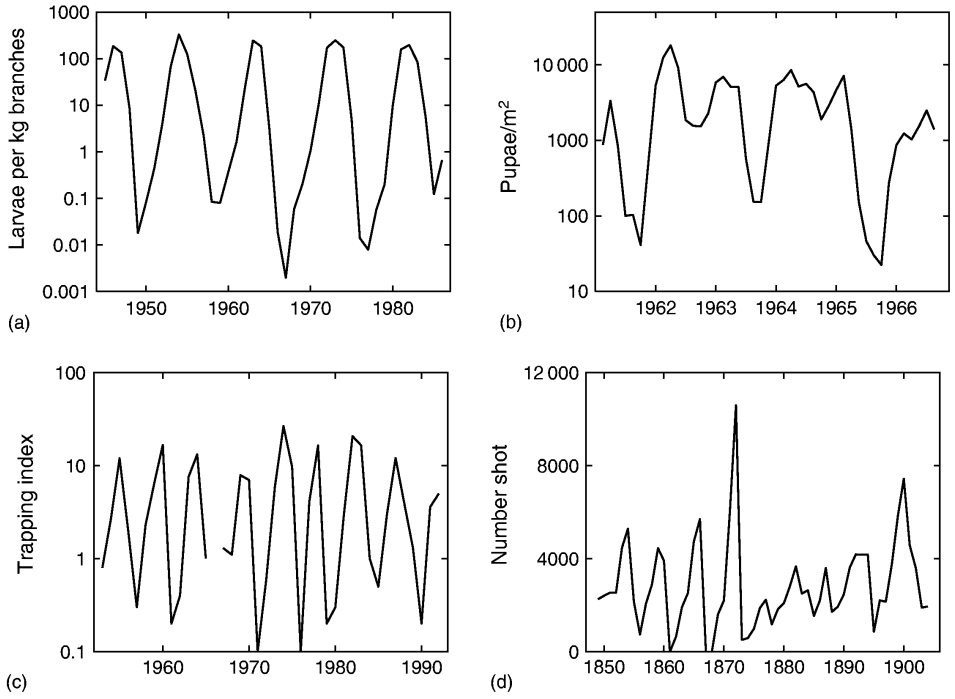


Figure 2 Examples of nonlinear dynamics in field populations. (a) Larch budmoth (*Zeiraphera diniana*) in Switzerland. These cycles are probably caused by interactions with parasitoids, although there may also be feedback through changes in nutritional quality of the larch needles. Tree-ring data indicate that this extraordinarily regular cycle has persisted for at least 150 years. (b) Coffee leaf-miners (*Leucoptera* spp.) in Tanzania. This also appears to be a host–parasitoid system, coupled with seasonal variation in growth rates (there are eight generations per year). (c) Voles (*Microtus* and *Clethrionomys*) in Finland. This is probably driven by predation by least weasels and stoats, which specialize on small rodents. (d) Red Grouse (*Lagopus lagopus scotius*) in Scotland. Both intraspecific competition and interactions with an internal parasite have been proposed as mechanisms for these cycles. (Redrawn from Baltensweiler and Fischlin, 1988; Bigger, 1973; Hanski *et al.*, 1993; Middleton, 1934)

Figure 3.1: from Kendall (2002)

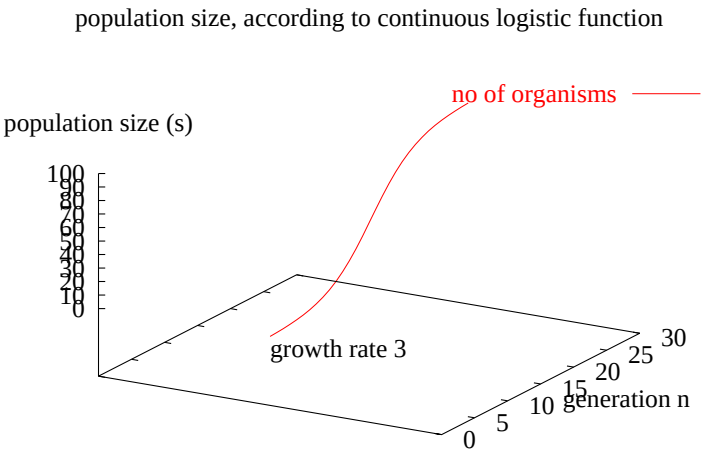
…These studies of the logistic map revolutionized ecologists' understanding of the fluctuations of animal populations – May (2002, p223)

Strangely, the erratic behavior of the logistic map goes away, and things become much simpler, if we let the population adjust itself constantly on a continuous timescale.

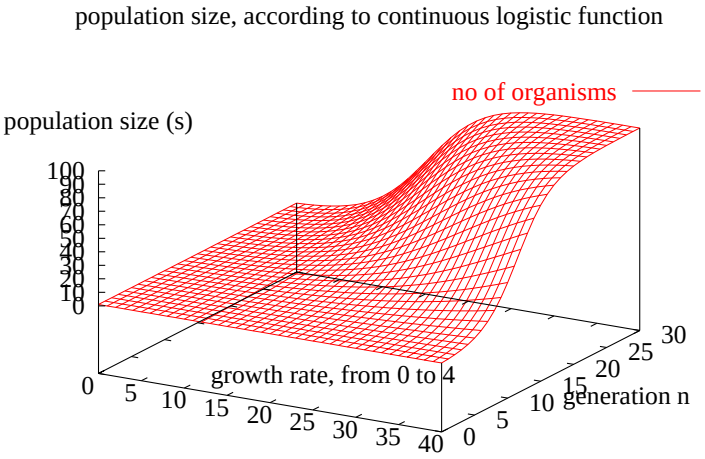
Letting	s = population size at each generation
	r = rate of reproduction
the logistic function is:	$\Delta s = rs(1 - \frac{s}{k})$
when $s(0) = 1$ and time is continuous:	$s(t) = \frac{k}{1+Ae^{-rt}}$
for a constant A	

This last equation looks mysterious. Since the population $s(t)$ at any time t is equal to k divided by something larger than 1, it is clear that there is no way for the population to get larger than k. To get a better idea of exactly how this works, we can look at some graphs.

When we draw the population growth with this new equation, what do we get? The population starts to grow quickly, as in the exponential model, except that it slows down quickly when it gets near the carrying capacity. So when the population starts below the carrying capacity, we get an S-shaped curve like this,

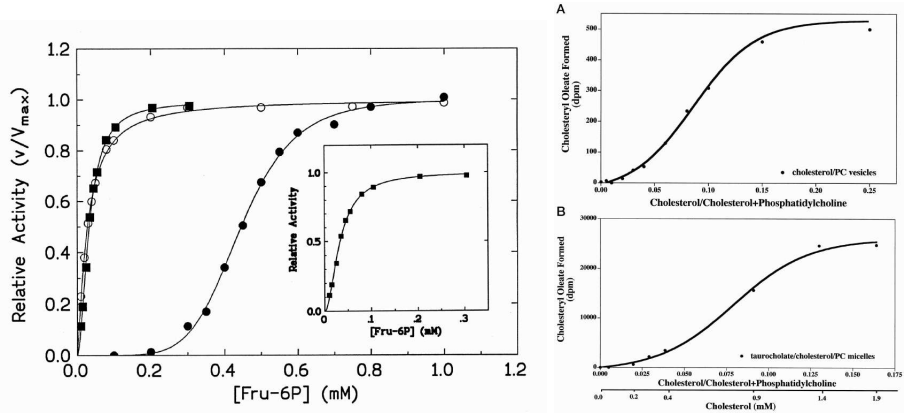


Varying the growth rate (but keeping it positive) varies the steepness of the S:

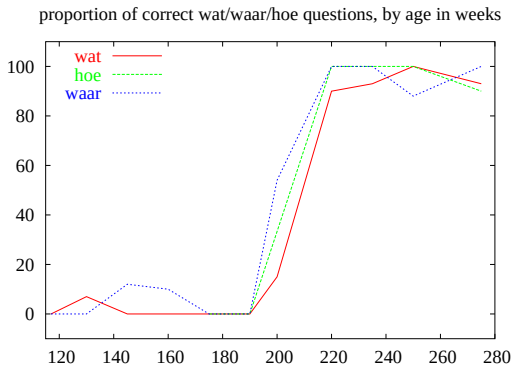


This continuous logistic curve is sometimes called "sigmoid" after the Greek letter sigma: ς. But since we do not all speak Greek, we may as well call it S-shaped. This name reminds us that it is the curve of population size when we do not have overshooting oscillations. This curve and near variants are probably the most important mathematical functions in the life sciences, because they fit very many things. Taking some examples out of the blue, plotting the activity of a catalyst as a function of the concentration of fructose in E. Coli bacteria (Byrnes et al.,

1995, on the left below), or plotting the formation of cholesterol as a function of an enyzme concentration in the cells of a hamster (Chang et al., 1998, on the right below), we have sigmoid curves:



We also seem to get something similar to S-shaped curves plotting the proportion of correct wh-questions (*who, how, where*) in Dutch children by age as in the figure on the left below (van Kampen, 1997), or the number of uses of the auxiliary verb *do* in the development of English from 1400-1700 (Kroch, 1987), as in the figure on the right below from Wonnacott (2000):



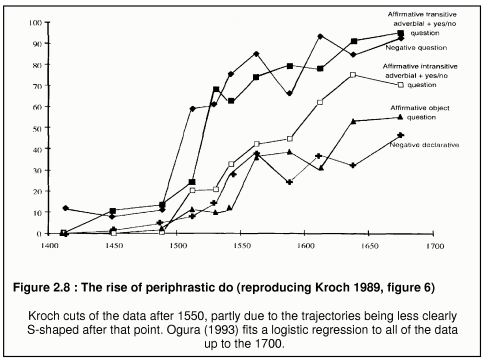


Figure 2.8 : The rise of periphrastic do (reproducing Kroch 1989, figure 6)

Kroch cuts of the data after 1550, partly due to the trajectories being less clearly S-shaped after that point. Ogura (1993) fits a logistic regression to all of the data up to the 1700.

S-shaped functions are also designed into many artificial neural networks, specifying "neuron" response as a function of activation, since the logistic curve starts to pick up suddenly at a certain "threshold" point and then slowing down quickly when it approaches "full activation" – a mathematically simple approximation to the behavior of a real neuron (Mehotra, Mohan, and Ranka, 1997, for example). We will see more logistic functions later.

This is completely optional, but if you have had some calculus, it is useful to remember how the continous logistic equation is derived. Remember that the discrete logistic function says:

$$\Delta s = rs(1 - \frac{s}{k}).$$

We replace this with the continuous equation

$$\frac{ds}{dt} = rs(1 - \frac{s}{k}).$$

We can separate the variables and integrate both sides, as we did for the simple population equation on page 80:

$$\int \frac{ds}{s(1 - \frac{s}{k})} = \int r dt.$$

Since $\frac{1}{s(1-\frac{s}{k})} = \frac{k}{s(k-s)} = \frac{1}{s} + \frac{1}{k-s}$, we can write

$$\int (\frac{1}{s} + \frac{1}{k-s}) ds = \int r dt.$$

Integrating we have, for constant C:

$$\log_e |s| - \log_e |k - s| = rt + C.$$

Now we calculate

$$\log_e |\tfrac{k-s}{s}| = -rt - C$$
$$|\tfrac{k-s}{s}| = e^{-rt-C} = e^{-C}e^{-rt}$$
$$\tfrac{k-s}{s} = Ae^{-rt} \quad \text{for constant} \quad A = e^{-C} \quad (\text{here we let } A = \tfrac{k-s_0}{s_0})$$
$$s = \tfrac{k}{1+Ae^{-rt}}$$

This is exactly the equation we showed on page 83. This derivation is a standard example in many first year calculus texts; careful and explicit discussions can be found, for example in Stewart (2003, §6.5), Neuhauser (2000, §7.1), Loomis (1975, §19.1).

3.2 Wrinkles in classical Mendelian genetics

In section §1.3, we emphasized what Mendel got right, but his "laws" need qualification:

Law of Dominance: Traits are controlled by pairs of genes, only one of which will determine the phenotype in a heterozygote

Law of Segregation: When reproductive cells are formed, the two alleles for a trait separate, and are recombined in the zygote.

Law of Independence: Alleles for different traits are distributed independently.

Observing that the reality is actually more complicated than this, Lewontin (2003) says of Mendel's three laws, "two turn out to be untrue in a large fraction of cases, and the third

has a few very revealing exceptions." He is right.

vs. Dominance: In the first place, most traits that you would ordinarily consider are not controlled by a single pair of genes, but by many genes in which case the effects can be complex, e.g. with the phenotypes genes completely masked by others ("epistasis"). And even for traits determined by a single pair of genes, sometimes both genes are expressed ("codominance," as in human blood groups), and sometimes the contributions of the two genes blend ("incomplete dominance," e.g. in some flower colors). Mendel was lucky to have considered some traits where simple dominance holds, since it made the proportions and the theory very simple.

vs. Segregation: As we have noted, many things can happen between separation and recombination, including all different kinds of mutations. There are also various failures of complete separation and recombination. These are important in the Darwinian synthesis, but they certainly complicate things!

vs. Independence: Genes on the same chromosome can be linked, so that if you get one particular allele of one gene, you are likely to get a particular allele of another gene.

I suspect Lewontin means to say that Dominance and Independence fail in a large fraction of cases, and the failures of Segregation are particularly illuminating. Looking back now, we can see that Mendel was lucky to happen upon traits that were controlled by genes with a simple dominance relation, so that the facts about relative frequency were so simple.

3.3 The missing interpretation: proteomics and beyond

Although it is impressive that we can list the 3.2 billion bases in the human genome, it is remarkable how little we can say about what it all means. We observed that there are various kinds of structures in the DNA, RNA and Protein sequences. We observed codons in the RNA, nested dependencies that are sometimes related to how the RNA folds, and duplicated sequences whose roles are not fully understood. It is clear that these things matter, but a full understanding is a long way off: what is each part of the genetic sequence for, how does it interact with proteins and cellular structures in the life of the organism? The study of how proteins control cellular processes is sometimes called *proteomics*, and this is sometimes regarded as the most important field to "follow" genetics (Tyers and Mann, 2003) at the molecular level, with a "phenome project" at the organismic level to specify how genes are expressed at the macro level.

Some biologists are very impressed by the coding mechanism itself. For example, Barbieri (2003) says:

> *One of the greatest biological achievements of the twentieth century was the discovery that the information of a gene is determined by the order of its nucleotides, pretty much as the information of a word is due to the order of its letters…The linear information of nucleotides is used to assemble a linear sequence of amino acids, and then this polypeptide chain folds on itself (because of electrical forces that exist between*

> *amino acids) and spontaneously assumes a specific three-dimensional structure. It is as if one wrote the word* apple *and then observed the word folding on itself and becoming an apple.*

In his recent book, James Watson (2003) says similarly:

> *a gene producing the biological equivalent of a brick will, left to its own devices, produce a pile of bricks.*

Lewontin (2003) rejects these remarks as simplistic and misleading:

> *…this is Watsonian hype. Genes don't have their "own devices." Left to their own devices, they will just sit there, dead molecules. One might as well say that a set of house plans, "left to its own devices," will build a house. To carry out the synthesis of proteins and other components of a developing organism, the cell uses an elaborate machinery of proteins and a warehouse of small parts, both of which are already in place in the fertilized egg. This machinery transcribes the information in the DNA into a related information-bearing molecule, RNA, which may be assembled from more than one gene. Using this RNA chain as a guide, the cell then assembles chains of amino acids using its stock of small parts. To make an active protein, a chain of amino acids must be folded into exactly the right three-dimensional form, a process that is partly determined by the sequence of amino acids, but is guided by yet other proteins and small molecules. Sometimes, before the folding can occur, pieces of the original amino acid chain are clipped out. This entire manufacturing process, from the original transcription of the DNA information to the finally assembled and folded protein, could not take place without the prior presence in the cell of protein catalysts, enzymes. So much for the gene's "own devices."*

What is this argument about? There does not seem to be a disagreement about the facts here, but only about what is most important. Lewontin is making the obvious point that the genome only functions in the very special conditions provided by the cell, while Barbieri and Watson are emphasizing the importance of the code itself. Does the code, by itself, represent anything, or have any information in it? We will try to develop some tools for making sense of this kind of question in later lectures. For now, a natural resolution is to agree with both sides (especially since we do not need to worry about funding priorities).

Everyone agrees that the organism needs both the code and the very special environment in which the code can be transcribed and translated into proteins. We can also observe that the flexibility of the code – to the extent we can apprehend it – is quite remarkable. The variety of forms of life is astounding. For an analogy, a computer is better than an apple, since with a computer too, you just "write down" an expression like 2×1024 (in the right place, where all the electrical requirements for the computation will be satisfied, etc.) and the physics will apply to change the machine into a state that represents 2048. Or you can "write down" the notes of a song and have them played by the computer's speakers. Barbieri is right that this kind of thing is miraculous, sort of like having linguistic expressions that come to life. What it really is

though, is the discovery that certain states of physical systems (given "natural" conditions of functioning) determine certain other states in a way that makes sense, a way that corresponds to the computation of a sum, the vibration of speakers, or the construction of an organism.

So what is the disagreement? Why is Lewontin pointing to the importance of the cellular mechanisms that enable DNA to play its role? One way to notice that we have missed something here is to compare Lamarck and Darwin again: Lamarck was mocked for suggesting that there is passage of acquired traits from one generation to the next, and Darwin rejected this in favor of selection from pre-existing variation. But notice that the cellular mechanisms enabling DNA must be present before it can be used. Those mechanisms are not part of the genome itself; rather, they must develop and then provided along with the genome. The zygote gets more than the genome; it gets an environmental niche where it can function in a certain way. This same kind of thing happens at larger scales too. Organisms often are born into environmental niches that have been created by their ancestors or other organisms, and in many cases these niches are essential. In a sense, the niches provide a kind of inheritance of acquired traits.

3.4 Self-organization

Having been impressed with the flexibility of the DNA code in specifying an enormous variety of living forms, it is important to avoid the absurd conclusion that it has some kind of "total flexibility." Obviously, the properties of organic systems and the environmental conditions on the planet are restricting the range of possible forms. In a way, this is to make Lewontin's point again: what the DNA can specify is restricted by, in fact, determined by the mechanisms that transcribe and translate the code. The properties of these mechanisms and restrictions from the environment can act to determine even quite elaborate and global properties of the result, as we suggested in section §0.4.

Examining any modern phylogeny, like the one on page 3, we see that the distribution of organisms is not uniform, but clumpy, with bursts of many different forms at certain points, and relatively distant relations with few forms elsewhere. This would be hard to explain on the view that 'anything goes' if only it is well adapted. Rather, it appears that certain points of development represent relatively 'successful' compromises between survival and specification.

Gould (2002, §10) lists some of the prominent sources of order in organisms that are independent from selection, sources of order that restrict and control the extent and nature of adaptation in organisms. Here are a few of them, beginning with the obvious:

Compatibility: In the first place, the various organs cannot each be modified independently to achieve the "perfect" design, since the organs must all work together; they must be compatible.

Mechanical limits: Equally obvious is the existence of various mechanical limits on organisms. For example, there are limits on the size of organisms that can fly, limits on the strength of legs of organisms that run, and so on, which all depend on the particular properties of the materials available for the development of the organism. Gould points out that it is no surprise that "zebra wings" have never emerged in the great phylogeny of life.

Heterochrony: Another kind of restriction on the adaption of organisms by natural selection is historical: the directions that prior adaptations have taken. In effect, once the ball is rolling in a certain direction, later adaptations are also channeled in the same direction. **Heterochrony** is the idea that one of the major ways of shaping an organism's development is by adjustments in the timing of development, particularly in the early, embryonic stages. Even Darwin noticed the striking similarity among the embryos of very different organisms, which fits the more recent idea that, given any population of organisms, the dimensions of variation will be shaped in large part by the kinds of changes that can be induced by changes in the timing in the development of the organism (McKinney and McNamara, 1991; Zelditch, 2001).

Paleontological evidence supports this view: **Cope's Law of the Unspecialized** proposes that "most lineages spring from founding species with generalized anatomies" (Cope, 1987; Gould, 2002, p.902).

Allometry: Another related kind of "channeling" effect of prior developments comes in **allometry**: the growth of one part at a different rate than other parts. Darwin noticed the similarity in the bones of the mole, the horse, the porpoise, the bat, and the human: the same basic parts, but developed in different proportions, with only rather slight variations in their overall architecture. Gould uses, among other things, careful studies of the variations in snail shells (*Cerion uva*) to support the importance of this factor in evolutionary development.

Hoxology: One collection of genes relevant to axis formation (discussed in section §0.4.4) and segmentation in in fruit flies (*Drosophila*), are the so-called *Hox* genes. These include "a 180 base pair unit coding for a 60 amino acid homeodomain with important regulatory action as a DNA binding protein" which was found to have homologs in vertebrates:

> *Not only do Hox genes exist in vertebrates, but homologs for all Drosophila Hox genes have been found, arranged in the same linear order on chromosomes, and acting with the same colinearity in development along the A-P axis of the vertebrate body…Fly Hox genes, expressed in vertebrates, usually broker the same developmental sequences as their vertebrate homologs – and vice versa…As one example among so many, the Drosophila Hox gene Antennapedia promotes leg identity, presumably by repressing previously unknown antennal genes. Cassares and Mann (1998) have now identified two antennal determiners, including homothorax (hth). As one line of evidence, they cloned Meisl, the mouse homolog of Drosophila hth, and expressed it ectopically in the fly's anal primordium, which normally develops without expressing any Hox genes. The anal plates of these flies grew as antennae.* (Gould, 2002, pp1102,1105).

Gould cautions "I need hardly remind my fellow evolutionary biologists that these results, no matter how fascinating and surprising, show only limited and partial homology…," but such homologs could support channeling effects in variation and evolution.

Pax-6 and homologs in homoplastic eyes: As noted on page 9, in considering the evolutionary development of organisms, it is important to distinguish true homologies (traits that really descend from a common ancestor) from mere homoplasies (traits that are similar but arose independently). It is commonly claimed that eyes of various kinds have

evolved independently 40-60 times among animals (Salvini-Plawen and Mayr, 1977), but the story is again more complex than this might suggest. We already observed on page 12 that in organisms with apparently independently developed eyes, we often see the very same photosensitive proteins or near variants. We are beginning to understand how this kind of thing can happen. The so-called *Pax-6* gene found in *Drosophila* exerts some influence on the form and function of the eyes. A *Pax-6* homolog was again found in mice. The mouse homolog can be expressed "ectopically" – in abnormal places – in *Drosophila*. It can induce apparently normal, compound fly eyes, on the antenna, legs and wings of the fly (Gehring, 1996). These eyes have normal photoreceptors, lens, cone and pigment cells, in spite of their abnormal locations. This suggests that we should be more careful about our claims about independent, repeated evolution of the eye: genes controlling many earlier forms of the eye may be present in organisms that emerge much later.

In the present context, this supports the idea mentioned earlier: once the ball is rolling in a certain direction, later adaptations may be channeled in the same direction, because the resources are there already. Variation and speciation are not arbitrary, but channeled in certain ways by prior events.

Our understanding of how the development of organisms is genetically controlled is growing quickly. The picture is complex, but clearly supports the view that variation is "channeled" in certain directions by historical precedents and restrictions on form. This idea is commonsense, and was proposed with some suggestive examples in section §0.4, but now that we can watch events at the molecular level, we see how it can happen to a truly remarkable extent because of the historical traces of early events that each organism carries in its genome.

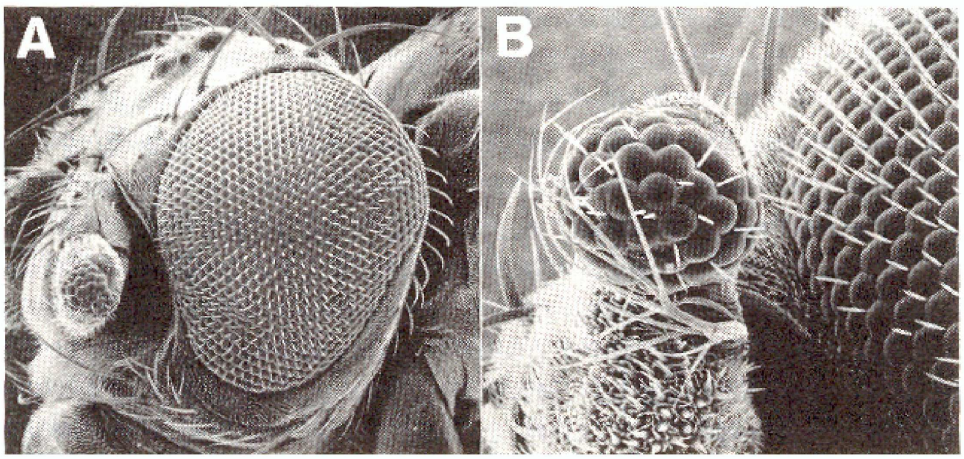


Figure 13.7
Eye induced on the antenna of a fly with the mouse Pax-6 (Small eye) *gene: (A) overview; (B) higher magnification. Scanning electron micrograph by G. Halder and A. Hefti.*

Lewontin (2002, p.13) describes the impact of these surprising findings this way:

It used to be said that the wings of bats and the wings of birds were homologous, but that the wings of birds and the wings of insects were only analogous because they were based on utterly different developmental processes with utterly different genetic bases. The *Hox* gene complex changed all that. By matching DNA sequences between species over an immense range of organisms, it is now possible to discover the common origin and trace the evolution of features of organisms irrespective of the degree of their apparent similarity or difference at any phenotypic level.

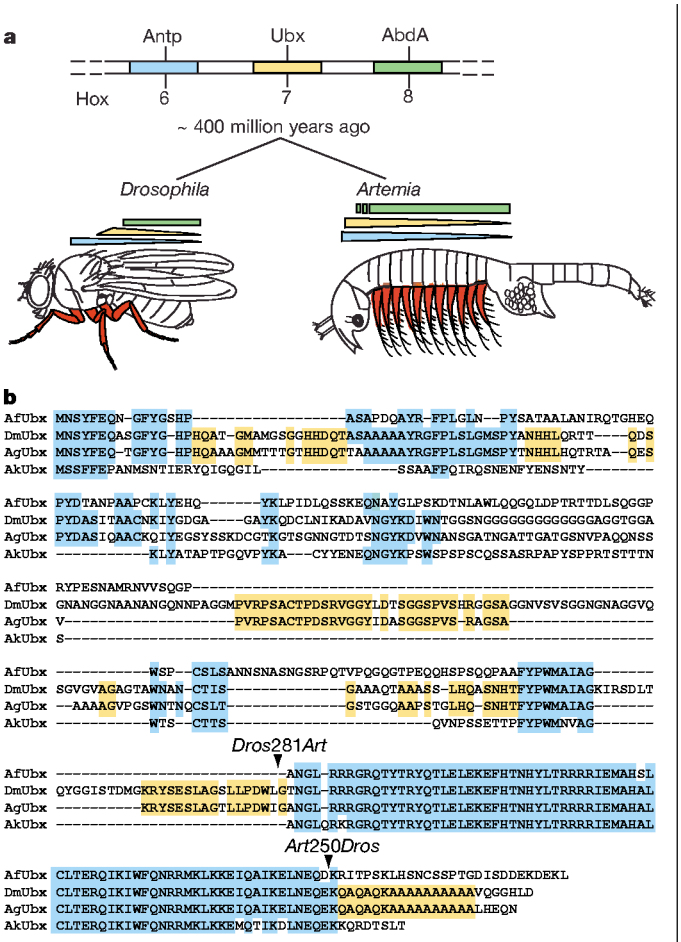
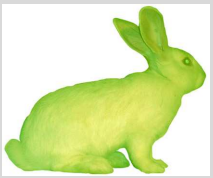


Figure 1 Evolution of trunk Hox gene expression patterns and sequence comparison of arthropod Ubx proteins. **a**, The crustacean lineage (for example *Artemia franciscana*) separated from the insect lineage (for example *Drosophila melanogaster*) about 400 million years ago. Crustaceans retained multiple limbs (red) on the trunk, whereas insect limbs became reduced to three thoracic pairs. At this time in arthropod evolution, the trunk Hox genes (*Antp*, *Ubx* and *Abd-A*) had already duplicated and diverged[23]. **b**, An amino-acid sequence alignment of Ubx protein sequences from the fruit fly *Drosophila* (DmUbx), the mosquito *Anopheles gambiae* (AgUbx), the brine shrimp *Artemia franciscana* (AfUbx) and the velvet worm *Akanthokara kaputensis* (AkUbx). Sequence motifs that are shared to different extents between all of these Ubx homologues are blue; motifs shared only by the hexapods *Drosophila* and *Anopheles* are yellow. The breakpoints of two hybrid proteins shown in Fig. 3 are marked with arrowheads.

Hox expression in crustacean and fruitfly from (Ronshaugen, McGinnis, and McGinnis, 2002)



Since it came up in class: yes, it's true that biologists created a fluorescent rabbit for artistic effect, working with French artist, Eduardo Kac. This was done with an enhanced version of a gene for fluorescence from a jellyfish. The green fluorescent protein produced by this gene is used extensively in medical research. For the artist's account see `http://www.ekac.org/gfpbunny.html`. For some of the genetic engineering details about how this was done, see for example, (Chalfie et al., 1994; Heim, Cubitt, and Tsein, 1995; Haas, Park, and Seed, 1996).

3.5 What is selected?

Darwin had the idea that species were just vaguely indicated collections of related, evolving organisms. The individual organism was the thing that had to survive to reproduce. The biologists Williams (1966) and Dawkins (1976) have argued that Darwin was mistaken about this. They suggest that evolution is really acting only to preserve and propagate each gene. Dawkins distinguishes "replicators" (the genes which replicate and get selected) from their "vehicles" that contain them (the organisms), the "machines" they build in order to propagate. It is "the selfish genes" that are important, not their vehicles, the machines they build to survive.

Critics of this view like David Hull (1980, 1994) point out that to make sense of natural selection, we need both "replicators" and "interactors," since natural selection is defined on the entities that interact with their environments: "Genes are certainly the primary (possibly sole) units of replication, whereas interaction can occur at a variety of levels from genes and cells through organisms to colonies, demes, and possibly entire species." When we see how the beaks of finches have adapted for their feeding requirements, it is natural to regard the whole complex finch itself as surviving to propagate.[1]

Gould agrees that natural selection operates on interactors, things that have a genesis, a history, a coherence, a death, and most importantly, descendants. Evolution works on things that can pass their favorable traits to their descendants, but crucially, this does not require that their descendants are "copies" of themselves. Gould (2002) makes the case this way:

> *The criterion of heredity only demands that the units of selection be able to bias the genetic makeup of the next generation towards features that secured the differential reproductive success of parental individuals…The simple observation of plurifaction – the relative increase an individual's representation in the heredity of subsequent generations – does not suffice to identify the operation of natural selection, for plurifaction can occur by nonselective means…[For example] individuals may plurify by accidents of genetic drift.*
>
> *[Dawkins says] "I am treating a mother as a machine programmed to do everything in its power to propagate copies of the genes which reside inside it" (1976, p.132). Or*

[1] Darwin's interest in finches is well known, but after careful empirical study, the story about finch evolution is suprisingly complex (Grant and Grant, 2002, 2006). Given our interest in communication, recent work on beak size and vocalization in Darwin's finches is especially interesting; see for example Huber and Podos (2006).

> *"A monkey is a machine which preserves genes up trees; a fish is a machine which preserves genes in the water; there is even a small worm which preserves genes in German beer mats. DNA works in mysterious ways" (1976, p.22). These colorful images misstate actual pathways of causality.* Organisms *work in* wondrous *ways, and they operate via emergent properties that invalidate Dawkins's concept of genes as primary agents.*

The **hierarchical theory of evolution** recognizes that there are various kinds of "individuals" that can be selected: genes, cells, organisms, demes (groups of related organisms), and even species. Like a gene or an organism, a species has a genesis ("speciation"), a history, a coherence, a death ("extinction"), and most importantly, descendants. A species has descendants by branching in the phylogeny. Treating species as individuals (and possibly also larger individuals encompassing species) allows us to make evolutionary sense of selection when it favors, for example, a species that has broad geographical distribution. This has been argued to be a factor in certain types of mollusks, for example (Jablonsky, 1987). Obviously, being widely distributed is not a property of any particular organism, but of the species as a whole.

3.6 Another challenge: Phenotypic plasticity

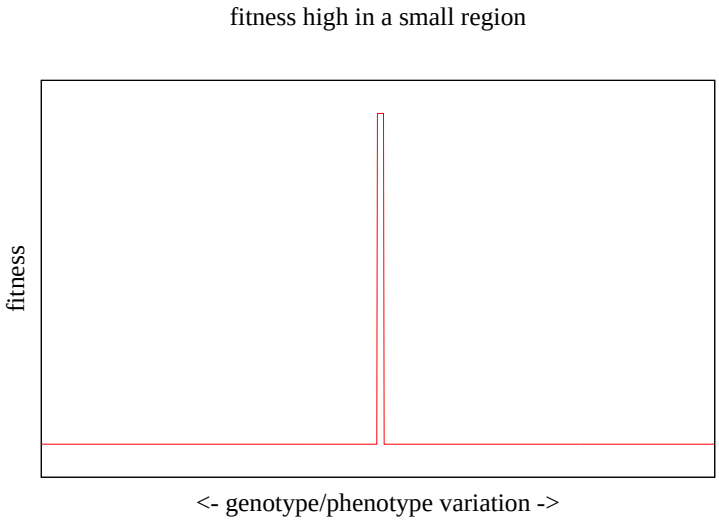
We have been talking about the traits of organisms as if they are fixed, static properties, but obviously this is not accurate. Traits emerge during development and they are, in most cases, shaped by genetic endowment and by the developmental environment. Most traits are "plastic" to some extent. Untangling the relative contributions of genetically determined properties from the environmentally determined ones is often complex: there are many "nature-nurture" controversies in the field, and we will focus later on how this kind of controversy arises in the case of human language.

Environmental effects on traits that are partially determined by genetic endowment are familiar: our skins tan, our hands get calloused, we acquire the language of our community, we learn about food gathering,...In other organisms, traits that are fixed in humans can be variable. For example, in primates, sex is determined by genes on the X and Y chromosomes – in particular the SRY gene on the Y chromosome triggers male (XY) traits, which XX females lack. In some other mammals, the sex determination works differently – some of them have no Y chromosome at all. Many birds and insects use another pair of chromosomes, with WZ females and ZZ males. But in some animals, sex is environmentally determined: in alligators, sex is determined by the temperature at which the egg is incubated. Even more dramatic are the fish who can start with one sex and change to another when it is advantageous for the population to do so (Godwin, Luckenbach, and Borski, 2003). But of course we see the most dramatic plasticity in organisms that can learn about their environment and adapt to it – a plasticity that is effected by chemical changes in the nervous system.

The "plasticity" of phenotypes is clear, but does it change anything in the evolutionary account? Why not just accept that traits are more or less "plastic" and stick to the usual Darwinian account? Naturalist J. Mark Baldwin showed that phenotypic plasticity actually changes the way

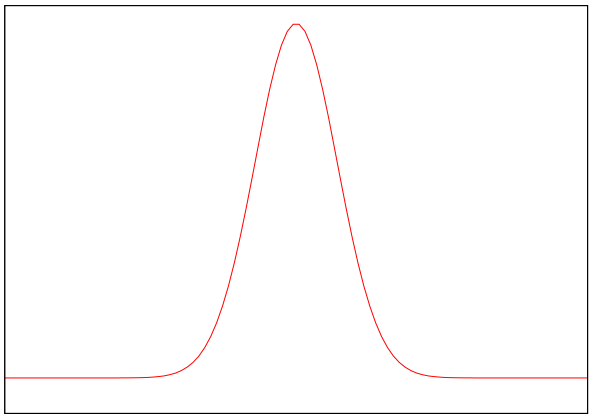
evolution works (Baldwin, 1896). We saw in section §1.5.1 that populations will tend to move toward genotypes that are more fit (where "more fit" = "producing more offspring"), and this will sometimes shape the populations of genes very quickly and dramatically.

Obviously, there will be no shaping influence at all unless there is enough variation in the population so that some individuals are more fit than others. And if most variation between a parent and offspring is small, the shaping influence can take hold here only when some of the offspring are somehow favored, so that reproductive success is not random from one generation to the next. From this perspective, we see that a population will not change much when the fitness landscape is basically flat except for very small region that is better:



The reason that the population will not change very rapidly here is that the descendants of the low fitness organisms are likely to be genetically similar to their parents, and hence unlikely to fall into the range that is extremely fit. Even organisms that are quite near to the ideal phenotype may have descendants that are all equally unfit.

If we imagine a similar situation but where each genotype determines a range of possible phenotypes, the fitness of genotypes that are near the highest fitness "ideal" will have a chance of adapting in such a way that their reproductive success is very high. That means that the genotypes near the ideal will be differentially selected because some of them may adapt. This has the effect of "smoothing" the fitness landscape:



This effect is widely discussed, because it means that acquired properties of organisms can have an effect on their fitness and hence on the inheritance of traits, not quite in the way Lamarck suggested, but in this more subtle way:

Baldwin effect, "organic selection": Phenotypic plasticity smooths the fitness landscape, facilitating the evolution of populations toward fitness peaks.

Notice that this effect happens at the species level – a level recognized by the hierarchical theory introduced in the previous section but not by classical Darwinism. It is the whole population that will tend to move towards the fitness peak, not particular individuals. There is controversy around the question of how much impact this effect has really had, but there are computer simulations showing that the effect can sometimes be dramatic (Hinton and Nolan, 1987; French and Messinger, 1994)

If phenotypic plasticity has this advantage – introducing the possibility of adapting to the environment and accelerating evolutionary development – why aren't all traits "completely" plastic? It is not hard to see that plasticity is sometimes a hazard. When the environment is stable, and an organism can be provided with a fixed trait with high fitness, there is no reason to spend the effort and run the risk of letting adaptation find the optimal phenotype. So in fact there are many trade-offs here, some of which have been proposed in models of leaf size and shape responses to crowding (Donahue and Schmitt, 1999), body size and rate of development in beetles (Guntrip and Sibly, 1998), sex choice in fish (Godwin, Luckenbach, and Borski, 2003) and oysters (Guo et al., 1998), and many other instances of plasticity.

costs
time,energy required for adaptation
risk to survival before adaptation achieved
risk of unreliability in adaptation
complex mechanisms (e.g. brains) needed for certain kinds of plasticity
benefits
genetic change is slow, but adaptation can adjust to quickly changing environment
genetic effects are species-wide, but adaptation can be for individual-specific environment
maintenance of genetic variation in range of adaptability
accelerated evolution towards fitness peaks

The interaction of these factors in real environments is extremely complex, so it has been valuable to explore them in simplified artificial environments, "artificial life" (Hinton and Nolan, 1987; French and Messinger, 1994; Mayley, 1996; Mayley, 1997).

3.7 At the limits, briefly

The neo-Darwinian synthesis has been extended to cover an enormous range of phenomena. Near the limits of this range, we find speculations about how far this account of life should reach.

3.7.1 Evolution before life? after life?

…Life began with the appearance of an autocatalytic (self-replicating) molecule.
– Eigen (1992)

How did the great burst of life on this planet happen, leading to the complex mechanisms of DNA transmission and protein synthesis that we have now? This is still poorly understood, but Oró (1961) made the headlines with his discovery that the base Adenine could be created by a reaction of ammonia and hydrogen cyanide. It appears that the other bases in DNA can also be created from inorganic materials (Voet and Schwartz, 1982), but the step from these to self-replicating polymers remains difficult, and has spurred a great deal of research, still ongoing (Joyce et al., 1997; Egholm et al., 1992).

There is also speculation about the possibilities for the possibility of intelligent entities of some kind after human existence and even after life itself, but this is so far mainly the domain of science fiction writers who imagine societies of robots or life in other parts of the universe.

3.7.2 What is life?

In considering what is selected, and what might precede or follow life, it becomes clear that the boundaries of what should count as "life" are hazy. When we look for life on Mars or elsewhere

in the universe, what are we looking for? Some definitions are quite specific with regard to the chemical properties of 'living' materials, while other definitions are more abstract:

> *What is the characteristic feature of life? When is a piece of matter said to be alive? When it goes on "doing something," moving, exchanging material with its environment, and so forth, and that for a much longer period than we would expect an inanimate piece of matter to "keep going" under similar circumstances. When a system that is not alive is isolated or placed in a uniform environment, all motion usually comes to a standstill very soon as a result of various kinds of friction; differences of electric or chemical potential are equalized, substances which tend to form a chemical compound do so, temperature becomes uniform by heat conduction. After that the whole system fades away into a dead, inert lump of matter. A permanent state is reached, in which no observable events occur. The physicist calls this the state of thermodynamical equilibrium, or of "maximum entropy." (Schrödinger, 1945)*

> *A physical system can be said to be living if it is able to transform external energy/matter into an internal process of self-maintenance and self-generation. (Varela, 1994)*

It is interesting to consider whether a virus like HIV1 or even computer virus is alive according to these definitions. Both HIV and computer viruses are self-replicating, but neither does very much self-maintenance. Whether we call these things "living" does not matter for most purposes: their properties are what matter. A sensible position on the issue is expressed in the following recent study:

> *Before the invention of molecular theory, people may (or may not) have believed that 'water' could be defined, but the best they could do in 'defining' it would be to discuss its sensible properties. In the absence of a compelling molecular theory, attempts at definition were doomed to interminable bickering over which of its sensible properties were essential to water's nature. We suggest that current attempts to define 'life' face exactly the same quandry. It is possible that in the future we will elaborate a theory of biology that allows us to attain a deep understanding of the nature of life and formulate a precise theoretical identity for life comparable to the statement 'water is H_2O.' In the absence of that theory, however, we are in a position analogous to that faced by someone hoping to understand water before the advent of molecular theory by 'defining' it in terms of the observable features used to recognize it. (Cleland and Chyba, 2002)*

The idea that "life" might be a particular, coherent, natural kind of thing, like water, rather than a rather arbitrary complex of many things, is an interesting one.

3.7.3 DNA and other molecular computing?

We have seen that the interpretation of the DNA code is not known, but what if we could use the code of DNA for our own purposes? Remarkably, this has actually been done. Adleman (1994)

encoded a classical problem in graph theory in DNA (an instance of the "Hamiltonian path" problem) and solved it with reactions triggered by enzymes, prompting optimistic speculation about how DNA might eventually allow the feasible computation of problems too hard for conventional computers (Gifford, 1994):

> *If we are able to construct a universal machine out of biological macromolecular components, then we could perform any computation by means of biological techniques. There are certainly powerful practical motivations for this approach, including the information-encoding density offered by macromolecules and the high energy efficiency of enzyme systems. At present there is no known way of creating a synthetic universal system based on macromolecules. Universal systems require the ability to store and retrieve information, and DNA is certainly up to the task if one could design appropriate molecular mechanisms to interpret and update the information in DNA. This ultimate goal remains elusive, but once solved, it will revolutionize the way we think about both computer science and biology.*

The original idea was to use chemical reactions to compute by substituting one sequence of nucleotides by another. The technological barriers in the way of using methods like these are still immense, but the idea has caught the attention of prominent researchers in the theory of formal languages and computation (Paun, Rozenberg, and Salomaa, 1998). And recently molecular computers have been built from pieces that comprise "logic gates" (AND, OR, NOT), and one recent demonstration of this work showed how a deoxiribozome-based computer could play tic-tac-toe (Stojanovic and Stefanovic, 2003). The prospect of feasible biomolecular computing of this kind is a distant but exciting one.

Exercises

This week we discussed the overall "neo-Darwinian" synthesis and some complications, and so it involves thinking about all the topics covered so far – good for review! The problems this week require thinking about all these things.

1. **vs Mendel and Hardy:** We mentioned some qualifications needed for "Mendel's laws" in section §3.2. What do these qualifications imply about the Hardy-Weinberg equilibrium?
2. **Self-organization:** As we observed in section §0.1, Darwin noticed commonalities among organisms, as evidence that we have common ancestors:

   > *What can be more curious than that the hand of a man, formed for grasping, that of a mole for digging, the leg of the horse, the paddle of the porpoise, and the wing of the bat should all be constructed on the same pattern, and should include the same bones, in the same relative positions.*

   And this fits with fossil evidence for "Cope's law," mentioned on page 90, and with the "clumpiness" of the phylogenetic tree on page 3 in lecture 0. More recently, as discussed in section §0.4, biologists have taken homologs of antenna genes out of mice, put them back into various positions in a developing fly larvae (e.g. on the "anal plates"), to have them trigger the growth of antennae in those unusual places.

   The idea that organisms are shaped (in part) by self-organization is the idea that some of their properties are determined more by basic requirements of their individual parts (organic molecules, genes, cells) than by the pressures to adapt.

   (i) Do Darwin's observations of similarities support this idea? (Briefly say why or why not)

   (ii) Does the discovery of antenna genes in mice support this idea? (why or why not)
3. **Evolution of death:** The 2002 Nobel prize for Physiology or Medicine went to Sydney Brenner, H. Robert Horvitz and John E. Sulston for their discoveries concerning "genetic regulation of organ development and programmed cell death." These passages are excerpted from the Nobel press release:

   > *All cells in our body are descendants from the fertilized egg cell. Cells differentiate and specialize to form various tissues and organs, for example muscle, blood, heart and the nervous system. The human body consists of several hundreds of cell types, and the cooperation between specialized cells makes the body function as an integrated unit. To maintain the appropriate number of cells in the tissues, a fine-tuned balance between cell division and cell death is required.*
   >
   > *Developmental biologists first described programmed cell death. They noted that cell death was necessary for embryonic development, for example when tadpoles undergo metamorphosis to become adult frogs. In the human foetus, the interdigital mesoderm initially formed between fingers and toes is removed by programmed cell death. The vast excess of neuronal cells present during the early stages of brain development is also eliminated by the same mechanism.*
   >
   > *Sydney Brenner realized, in the early 1960's, that fundamental questions regarding cell differentiation and organ development were hard to tackle in higher animals. Therefore,*

> *a genetically amenable and multicellular model organism simpler than mammals, was required. The ideal solution proved to be the nematode Caenorhabditis elegans. This worm, approximately 1 mm long, has a short generation time and is transparent, which made it possible to follow cell division directly under the microscope.*
>
> *As a result of these findings Sulston made the seminal discovery that specific cells in the cell lineage always die through programmed cell death and that this could be monitored in the living organism. He described the visible steps in the cellular death process and demonstrated the first mutations of genes participating in programmed cell death, including the nuc-1 gene. Sulston also showed that the protein encoded by the nuc-1 gene is required for degradation of the DNA of the dead cell.*
>
> *Robert Horvitz continued Brenner's and Sulston's work on the genetics and cell lineage of C. elegans. In a pioneering publication from 1986, he identified the first two bona fide "death genes", ced-3 and ced-4. He showed that functional ced-3 and ced-4 genes were a prerequisite for cell death to be executed.*
>
> *Later, Horvitz showed that another gene, ced-9, protects against cell death by interacting with ced-4 and ced-3. He also identified a number of genes that direct how the dead cell is eliminated. Horvitz showed that the human genome contains a ced-3-like gene. We now know that most genes that are involved in controlling cell death in C. elegans, have counterparts in humans.*

(i) Does the idea of cells programmed to die (rather than staying alive to produce more offspring) conflict with Darwin's proposals about natural selection?

(ii) In a "hierarchical theory of evolution" (described in §3.5), is cell death best described as selected for the gene, the cell, the organism, groups of organisms, or the species?

4. **Baldwin:** Why does Baldwin say in section V of his paper (on the web page) that natural selection is "entirely negative" while organic selection is a "positive agency?" Are these labels appropriate? (briefly say why)

5. **Baldwin:** In section IV of his paper (on the web), Baldwin says

   > *The intelligent use of phylogenetic variations for functional purposes in the way indicated, puts a premium on variations which can be so used, and thus sets phylogenetic progress in directions of constantly improved mental endowment.*

   Looking at the context of this remark in section IV, is this proposal the same as what we now call the the Baldwin effect – the acceleration of evolution ("phylogenetic progress") towards fitness peaks? (Briefly say how they are the same or different)

6. **Mendel.** We observed that Mendel's laws of dominance, independence and separation are not quite true. Reformulate each of these laws (with qualifications, or weaker claims) so that they are still significant and probably true.

7. **Organisms as the subject and object of evolution.** Lewontin (1983) says:

   > *Before Darwin, theories of historical change were all* **transformational***. That is, systems were seen as undergoing change in time because each element in the system undergoes an individual change during its life history. Lamarck's theory of evolution was transformational, for it regarded species as changing because each individual organism within the species underwent the same change…*

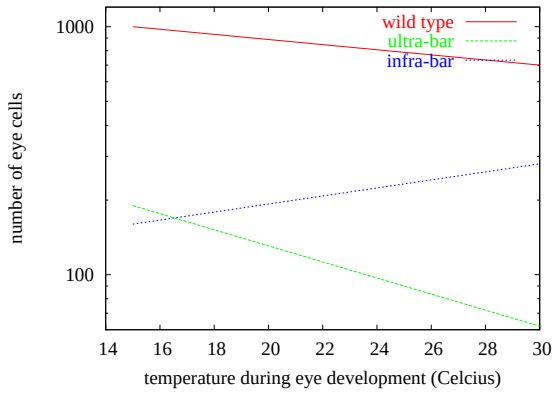
> *In contrast to these transformational theories of change, Darwin proposed a* **variational** *principle. Different individual members of the ensemble differ from each other in some properties, and the system evolves by a change in the proportions of the different types. There is a sorting out process in which some variant types persist, while others disappear so that the nature of the ensemble as a whole changes without any successive changes of individual members. Thus, variation of one kind, variation between objects in space, becomes transformed qualitatively into temporal variation. A dynamic process in time arises as the consequence of a static variation in space. There is no process other than the evolution of living organisms that has this variational form, at least as far as we know.*
>
> *In transformational theories, the individual elements are the* **subjects** *of the evolutionary process, for it is the change in the elements themselves that produces the evolution. These subjects change because of forces that are entirely internal to them,…Darwin's variational theory is a theory of the organism as the* **object***, not the subject of evolutionary forces. The variation between organisms arises as a consequence of internal forces, but these are autonomous and alienated from the organism as a whole. The organism is the object of these internal forces that operate independently of its functional needs or of its relations to the outer world. …The external chooses among many possible internal states, determining which shall survive.*

Is it true that "no process other than the evolution of living organisms" is "variational" in the sense described here? Explain. (hint: we talked about this!)

8. **vs Dawkins, part 1.** We mentioned one of Gould's criticisms of Dawkins in class (and in the notes), but Lewontin points to a more basic issue. Lewontin suggests that Darwin's idea that the organism is the object of evolutionary forces leads to views like Dawkins', but this is a mistake.

> *In* **The Selfish Gene***, Richard Dawkins speaks of organisms as 'robots' 'controlled body and mind' by the genes, as nothing but a gene's way of making another gene…But such a view…ignores…fundamental properties of living organisms…*
>
> *First, it is not true that the development of an individual organism is an unfolding or unrolling of an internal program…The organism is the consequence of an historical process that goes on from the moment of conception until the moment of death in which gene, environment, chance and the organism as a while participate at every moment.*

In defense of the first point, Lewontin mentions instances where a genetic variation produces one or another result, depending on the environment. For example, the number of light receptor cells in the compound eye of the fruit fly, Drosophilia, is usually about 1000 in the "wild type" fly genome, depending on the temperature, but this differs in two genetic variants called "ultra-bar" and "infra-bar:"



In this situation, we cannot answer the question: which of the ultra-bar and infra-bar variants determines the most eye cells. The reason is that the number of eye cells depends on the temperature.

(i) What are some other cases where this kind of thing happens – where a particular genetic endowment does not predict the results, because they depend on environment?

(ii) Is it true that such things are a problem for Dawkins' idea that ultimately, it is only the genes that are selected by the environment?

9. **vs Dawkins, part 2.** Continuing the passage above, a second problem for Dawkins is mentioned:

> *Second, it is not true that the life, death and reproduction of an organism are a consequence of the way in which the living being is acted upon by an autonomous external environment. Natural selection is not a consequence of how well the organism 'solves' a set of fixed 'problems' posed by the environment, but, on the contrary, the environment and the organism codetermine each other in an active way…*
>
> *The organism cannot be regarded as the passive object of autonomous internal and external forces. It is also the subject of its own evolution.*

The organism itself is part of the environment, and defines its own "niche," so that the environment is shaped by the organism and vice versa. The direction of influence does not go just one way, and the influence of an organism on an environment is not always positive:

> *Plant roots alter the physical structure and chemical composition of the soil in which they grow, withdrawing nutrients, but also conditioning the soil so that nutrients become more easily mobilized. Grazing animals actually increase the rate of production of forage, both by fertilizing the ground with their droppings and stimulating plant growth by cropping. …White pine trees in New England make such dense shade that their own seedlings cannot grow up under them, so hardwoods come in to take their place. It is the destruction of the habitat by a species that leads to ecological succession…The most powerful reconstitution of the environment that has been made by organisms is the gas composition of the atmosphere…It is living organisms themselves that have produced the oxygen by photosynthesis and have depleted carbon dioxide by fixing it in the form of carbonates in sedimentary rock…It is difficult to think of any physical force or universal physical low that represents a fixed problem to which all organisms must find a direct solution.*

Explain how these ideas conflict with Dawkins' idea that evolution applies only to genes, and then say how you think Dawkins might answer these points.

10. **Baldwin.** Suppose the Baldwin effect works in the way discussed in class (and in the notes). Is an evolutionary theory that incorporates this effect transformational (in the sense Lewontin describes in the first question, above)?

Selected Solutions

1. The Hardy-Weinberg equilibrium depends on the assumption of random distribution of the genes, so that they become "evenly mixed" and stay that way. The equilibrium does not say anything about **dominance**, but the even mixing does depend on properties of segregation and independence. Considering **segregation**, equilibrium is not destroyed if sometimes separation and recombination of the genes fails in some way, unless this happens in a "non-random" way. For example, if the recessive gene had some chemical property that made it more likely to have a separation or recombination failure of some kind, then this would destroy the equilibrium and could lead to a change in relative frequencies of the genes. And since **independence** does not hold because of "linkages" between genes, if there is any non-random influence on the linked genes, this also can destroy equilibrium. For example, if the recessive gene for some trait were linked to a trait that is maladaptive, this can have an effect.
2. (There are various acceptable ways to answer this one.)
   i. Darwin's observations about similarities in the bone structures of many animals does support the idea that self-organization, particularly from historical influences, is important. Clearly the particular bones in humans, moles, horses, porpoises and bats are suitably functional, but it would be absurd to think that the particular configurations they have taken are, in each case, due purely to design imposed from outside the organism (e.g. by selection). Rather, there must be something in the developing cells of each organism that determines these properties, something in each cell that comes from a common ancestor for these organisms. As Gould says, this "starts the ball rolling" in a specific direction.
   ii. It is one thing to notice that the common traits of organisms must be, in many cases, homologous, but the recent genetic studies of Hox genes and others shows that the genetic endowments in each of our cells apparently contain information relevant to many traits that are not expressed. This does provide further support for the original Darwinian observation about similarities, but conflicts with Darwin's idea that all design comes from natural selection, a force external to the organism.
3. (there are various acceptable ways to answer.)
   i. "Programmed death" apparently conflicts with Darwin's idea that selection will always favor the fittest organisms – at least, in this case, if you mean by "organism" the cell whose death is programmed. As far as each particular individual is concerned, it will always have increased fitness if it can live longer and reproduce more. So if cells (and also multicellular organisms) are "programmed" to die, this certainly does not increase their individual fitness.
   ii. "Programmed death" would not be selected at the level of the individual cell. It could be selected indirectly by selection for the organism containing the cells, or it could be selected for the "species" of cell. "Programmed death" could be selected at the species level since, in many natural settings, it definitely does make sense that killing off the older generation would reduce competition for food and resources with the new generation, and also would promote more genetic variation which could be healthy for

the species. But this death is good for the fitness of the species, not good for the fitness of the individual who is dying!

4. (again, various acceptable ways to answer.) Natural selection can only filter out variations, by killing or otherwise preventing some organisms from reproducing. It does not produce any adaptive change itself. So the label "entirely negative" does seem appropriate for natural selection. It kills off what doesn't work.

   Baldwin applies the term "organic selection" to "the organism's behavior in acquiring new modes or modifications of adaptive function." And when an organism can survive by adapting, this allows "all the time necessary to perfect the variations required by a complete instinct," and so "future development at each stage of a species' development must be in directions thus ratified by intelligence."

   In this way, organic selection guides phylogenetic development, instead of just killing off random changes that do not work. So it does seem appropriate to label organic selection a "positive agency." It guides instead of just filtering.

5. There are a couple of minor differences: Baldwin seems to think that "consciousness" and hence "mental endowment" is needed for adaptation, but the Baldwin effect is expected to apply in any instance of phenotypic plasticity – for example, to environmentally-based sex-change in fish, which is presumably not a conscious decision. Furthermore, he does not use the modern concept of "fitness." But Baldwin's main idea does seem to be the same as the modern one. He says that once an organism can survive by adapting, this allows, at the phylogenetic level, "all the time necessary to perfect the variations required by a complete instinct." In modern terms, an organism near a fitness peak can sometimes survive by adapting, and thus, merely surviving and reproducing, cause evolution to favor closeness to those peaks, "smoothing the landscape."

Review Questions

1. **Gene populations: Mendel and Hardy.**
   i. When Mendel crossed purebred (homozygote) wrinkled-pea plants with purebred smooth-pea plants, all the offspring produced smooth peas. Why did that happen?
   ii. Suppose that in a population, of the genes controlling dark color peas, $\frac{1}{3}$ are a recessive allele d for dark peas and $\frac{2}{3}$ are a dominant allele D for lighter peas. Assuming random mating, how many of the pea plants will have dark color peas?
2. **Gene populations: Mendel and Hardy.**
   i. Suppose we identify some particular genes of 5 plants in a greenhouse and find exactly these:

   AA Aa aa aa aa

   (Or equivalently, suppose we have 1000 of each of these 5 types, so that the small size of the population is not a concern.) If we mate all these plants together "randomly", what proportions of AA, Aa, and aa do we expect to find in the next generation?
3. **Mendel.** Lewontin says,

   > Mendel succeeded where others had failed partly because he worked with horticultural varieties in which major differences in phenotype resulted from alternative alleles for single genes. In Mendel's peas there was a single gene difference between tall and short plants, but in the usual natural populations of most plant species, there is no simple relation at all between height and the genes.

   If there were no simple relation at all between wrinkled/smoothness and the genes, what do you think would have happened in the first generation when Mendel bred smooth peas with wrinkled peas. (briefly explain why)
4. **Genetic transmission: equilibrium.** Hardy and Weinberg showed that in certain special conditions, the numbers of different kinds of genes in a population is perfectly constant, but we saw that in other conditions the relative frequencies of various genes can change very rapidly. Considering the genetic endowment of humans on the planet, it is certain that it is not at equilibrium: it is changing. List the main factors that are causing this change, beginning with the ones you think are probably most important.
5. **Genetic transmission: population.**
   i. Epidemiologists, who study diseases and epidemics in populations, are naturally concerned about the conditions under which "pathogens" (causes of disease) evolve to become "virulent" (harmful to their hosts). Clearly, in the most extreme case, a virus that requires a living host to replicate would not survive:

   (Ultimate virulence) Immediately upon entering the host, the host is killed (and so the virus is destroyed before it can even replicate).

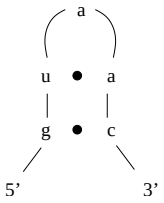
   A pathogen cannot be this virulent because it would become extinct immediately. How virulent can it be? That is, how would you define (Maximal virulence) – the worst a pathogen can be, such that it (the pathogen) can still survive.

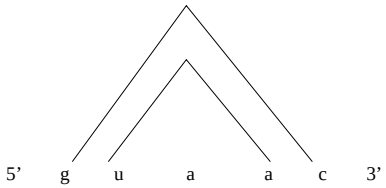
ii. Can you think of any pathogen that is "maximally virulent" in your sense? If so, what is it? If not, why wouldn't there be more pathogens like this?

6. **The language of DNA.** The language of DNA has "crossing dependencies." What are they?

RNA Parts:	5'	3'	a	u	g	c		
	Begin	End	Base	Base	Base	Base		
RNA-rule0:	x		$\mapsto$	x			**'start with any base'**	
	Base			Start				
RNA-rule1:	x	y		$\mapsto$	xy		**'extend to the right'**	
	Start	Base			Start			
RNA-rule2:	x	y		$\mapsto$	xy		**'extend to the left'**	
	Base	Start			Start			
RNA-rule3:	x	y	z		$\mapsto$	xyz	**'add Begin & End'**	
	Begin	Start	End			RNA		
RNA-loop:	x	y	z		$\mapsto$	xyz	**if x & z are attracting**	
	Base	Start	Base			Start		

7. **The language of RNA.** Use a tree diagram to show the steps in building an RNA molecule with the structural dependencies shown here, using the rules given above.





8. **Neo-Darwinism and Death once more.** A queen bee mates just once each with several males, storing the sperm in an organ called the spermatheca, for use throughout her life. The male honeybees, the "drones," die within an hour or so of mating.

   i. From a modern, neo-Darwinian perspective, what kinds of selective advantage could the bees that exist today have over a (hypothetical) bee where the drones survived to mate many times?

   ii. What differences between bees and mammals might explain why male mammals generally survive mating?

9. **The genetic clock.** Suppose that there is an organism containing just one strand of DNA with 1,000,000 nucleotides, and investigations reveal the following things:

   i. on average, each organism produces 2 offspring, and each offspring has, on average, 5 point mutations

ii. on average, a new generation is produced in 1 year (and the old generation dies)

Now suppose that we find two varieties of the organism which differ at 21 points. Approximately how long ago did the common ancestor live?

10. **RNA → Amino acids.** We saw that DNA determines RNA sequences, and the RNA sequences code the amino acid sequences in proteins, according to this table:

	u	c	a	g	
u	Phe	Ser	Tyr	Cys	u
	Phe	Ser	Tyr	Cys	c
	Leu	Ser	Stop	Stop	a
	Leu	Ser	Stop	Trp	g
c	Leu	Pro	His	Arg	u
	Leu	Pro	His	Arg	c
	Leu	Pro	Gln	Arg	a
	Leu	Pro	Gln	Arg	g
a	Ile	Thr	Asn	Ser	u
	Ile	Thr	Asn	Ser	c
	Ile	Thr	Lys	Arg	a
	Met	Thr	Lys	Arg	g
g	Val	Ala	Asp	Gly	u
	Val	Ala	Asp	Gly	c
	Val	Ala	Glu	Gly	a
	Val	Ala	Glu	Gly	a

i. Which amino acid is coded by the sequence: cgc

ii. Can a "point mutation" apply to this sequence without changing the amino acid the sequence codes? (if so, which mutation has this property?)

iii. Can a "point mutation" that applies to the first nucleotide change the amino acid the sequence codes? (if so, which mutation does it?)

iv. Can a "point mutation" that applies to the 2nd nucleotide change the amino acid the sequence codes? (if so, which mutation does it?)

v. Can a "point mutation" that applies to the 3rd nucleotide change the amino acid the sequence codes? (if so, which mutation does it?)

11. **The language of RNA.** The most obvious non-randomness in DNA and RNA is found in long repetitions. We saw that repetitions have "crossing dependencies" and so they cannot be defined by grammars that just extend a sequence to the right, or by a grammar that allows extensions to the right and left at once. But they can be defined by grammars that build expressions that have two parts. Use the following grammar to build the sequence, showing the derivation in a tree:

$$5'cgcacgc3'$$

RNA Parts:	5' Begin, 3' End, a Base, u Base, g Base, c Base	
RNA-rule0:	$\underset{\text{Base}}{x} \mapsto \underset{\text{Start}}{x}$	**'start with any base'**
RNA-rule1:	$\underset{\text{Base}}{x}\ \underset{\text{Base}}{x} \mapsto \underset{\text{Copy}}{x,x}$	**'start copies with any base'**
RNA-rule2:	$\underset{\text{Copy}}{x,x}\ \underset{\text{Base}}{y}\ \underset{\text{Base}}{y} \mapsto \underset{\text{Copy}}{xy,xy}$	**'extend copies with any base'**
RNA-rule3:	$\underset{\text{Copy}}{x,x}\ \underset{\text{Start}}{y} \mapsto \underset{\text{Start}}{xyx}$	**'splice copies together'**
RNA-rule4:	$\underset{\text{Begin}}{x}\ \underset{\text{Start}}{y}\ \underset{\text{End}}{z} \mapsto \underset{\text{RNA}}{xyz}$	**'add Begin & End'**

12. **Self-organization.** In the introduction to the class, we noticed that certain properties of biological units (molecules, organisms, groups of organisms) are apparently due to selection from alternative, less adaptive possibilities. But other properties are due to other kinds of organizing forces, some of which emerge from basic properties of the components of the biological units. And then in Lecture 3 notes (and in class), we listed some examples of this that have been discussed by biologists recently (Gould and others): examples where some possiblities are excluded not by selection but by other forces, and examples where certain possiblities arise because of properties of the basic parts. Which of these factors do you think has the largest influence on the nature of living things? (list at least one or two and briefly say why you think these have a significant effect)

Lecture 4

What is a language? First ideas

In section §3.7.2, we noticed that there are various different definitions of the term "life," and, at least at present, there is no science to resolve the question with anything analogous to the realization that "water is H_2O." A similar situation confronts us if we try to define the term "language." There is little consensus about what, exactly, should count as a "language." In mathematical and computational research, it is common to let any set of structured objects be called a "language." Others count a set of structured objects as a "language" only if each element of the set has some kind of "interpretation" or "meaning." Some prominent linguists, on the other hand, think that human language may be a special, natural kind of system, which might eventually allow a definition like the one we have for "water."

We will not try to decide on a "right" definition here: that is pointless! There is no "right" definition. Instead, we will introduce in this chapter a very liberal, mathematical definition based on information theory, and consider how it could apply in biological and even molecular settings. Then, in the next chapter, we will look at human languages and see that we find there very many special properties that "languages" in the broader sense of this chapter do not have.

The mathematical "theory of communication" is mainly concerned with the "average information" of a signaler, but it can provide also a notion of the content of a particular signal too, as described in this passage from Dretske (1983):

> *Communication theory only makes sense if it makes sense to talk about the probability of certain specific conditions given certain specific signals. …A signal r carries the information that s is F = the conditional probability of s's being F, given r (and k) is 1 (but, given k alone, less than 1).*

Dretske explains "The parenthetical *k*…is meant to stand for what the receiver already knows (if anything) about the possibilities that exist at the source."

We will explain this account of communication more carefully below, but we can notice right away that some biologists have clearly been influenced by this account of what communication is when they describe things that go on in the biological world:

Biological communication is the action on the part of one organism (or cell) that alters the probability pattern of behavior in another organism (or cell) in a fashion adaptive to either one or both of the participants. (Wilson, 2000)

Nearly all authors agree that communication involves the provision of information by a sender to a receiver, and the subsequent use of that information by the receiver in deciding how to respond. The vehicle that provides the information is called the signal…what other criteria are usually invoked to characterize animal communication? …The first is that the provision of information is not accidental but occurs because it benefits the sender…The second criterion is that the receiver must also benefit by having access to the provided information. (Bradbury and Vehrencamp, 1998)

We tend to think of biological signals as conveying or carrying information. In general, this characterization is accurate. (Hauser, 2000)

4.1 Communication as information transmission

The standard definition of **information** from Claude Shannon is extremely simple. Suppose we have a trick coin with heads on both sides. The obviously, $p(\text{heads}) = 1$ and the amount of information I give you by saying that the result is heads is 0. If the coin is fair, then $p(\text{heads}) = \frac{1}{2}$ and now saying that the result is heads tells you something. If instead of a 2-sided coin we have a 6-sided die, then the probability of rolling a 1 $p(\text{heads}) = \frac{1}{6}$ and telling you I rolled a 1 conveys even more information.

What we see in these examples is that there is an **inverse relation between the information conveyed and the probability**. The higher the probability of a result, the more predictable it is, the less information in that result.

The usual measure of information is base 2, the **bit**. A signal has x bits if on average you can find out what the signal is with x yes-no questions, or you can code it with x binary digits (using the naming rule from page 47, for a vocabulary of size 2).

To figure out how many digits you need, how many bits there are, Shannon proposed this: Letting $p(A)$ be the probability of A, the amount of information (or "self-information" or the "surprisal") of an event A, measured in "bits" is

$$i(A) = \log_2 \frac{1}{p(A)} = -\log_2 p(A).$$

In other words, the amount of information $i(A)$ in "bits" is the n such that $2^n = \frac{1}{p(A)}$. Because we have 1 over the probability, the more unlikely an event, the greater the information it has.

Considering the examples above again, if the coin has a head on both sides, $p(\text{heads}) = 1$ and since $2^0 = 1$

$$i(\text{heads}) = \log_2 \frac{1}{p(\text{heads})} = \log_2 \frac{1}{1} = \log_2 1 = 0 \text{ bits.}$$

This is what we wanted: in this case, saying the result is heads has no information.

If the coin is fair, though, since $\frac{1}{\frac{1}{2}} = 2$ and $\log_2 2 = 1$, that is, $2^1 = 2$, the information in a flip of a coin that turns up heads:

$$i(heads) = \log_2 \frac{1}{p(\text{heads})} = \log_2 \frac{1}{\frac{1}{2}} = \log_2 2 = 1 \text{ bit.}$$

Again, this is what we expect.

If instead of a coin we have a fair 6-sided die, since $\frac{1}{\frac{1}{6}} = 6$ and $\log_2 6 = 2.585$, that is, $2^{2.585} = 6$, the information in saying that a roll gave a 1 is:

$$i(\text{roll} = 1) = \log_2 \frac{1}{p(\text{roll} = 1)} = \log_2 \frac{1}{\frac{1}{6}} = \log_2 6 = 2.585 \text{ bits.}$$

Again, this is what we expect: more information than in the previous cases. Now let's put this idea to use in a definition of communication.

Communication happens when there is an event, a **signal** r that changes your estimate of the probability of something. For example, suppose that you think I am rolling a fair 4 sided die, or choosing one of the bases *c, g, t, a.* If each result is equally likely, and I tell you that the result is x, how much information is conveyed? As before, we can simply calculate

$$i(x) = \log_2 \frac{1}{p(x)} = \log_2 \frac{1}{\frac{1}{4}} = \log_2 4 = 2 \text{ bits.}$$

Notice that this only works if I decided to tell you the truth about the result! That is, there are really two events here: there is the happening of the result, and the happening of me telling you something. If I am perfectly honest, then the "conditional probability" that I say the result is x when the the result is x is 1. If I am not always honest, then in engineer's jargon, we say communication is "noisy" – that is, there is some chance that the signal does not correlate perfectly with the source.

Most of information theory is concerned with the information conveyed by an information source on average. For example, on average, how much information do we get per coin flipping event? Shannon's equation for this, the **entropy** H of a source s is this, where each of the possible events i at the source has probability p_i:

$$H = -\sum p_i \log_2 p_i,$$

That is, the entropy of the source in bits is the sum of the probability of each event times its surprisal. Considering a fair coin flipper, it is no surprise that the entropy is 1, since

$$\begin{aligned} H_{\text{fair coin}} &= -(p(\text{heads}) \log_2 p(\text{heads}) + p(\text{tails}) \log_2 p(\text{tails}) \\ &= -((\tfrac{1}{2} \log_2 \tfrac{1}{2}) + (\tfrac{1}{2} \log_2 \tfrac{1}{2})) \\ &= -((\tfrac{1}{2}(-1)) + (\tfrac{1}{2}(-1))) \\ &= -(-\tfrac{1}{2} - \tfrac{1}{2}) \\ &= -(-1) = 1 \text{ bit} \end{aligned}$$

Suppose the coin is biased, though, so that $p(\text{heads}) = \frac{9}{10}$. Then a heads has much less information than a tails, and an average coin flipping event has a lower entropy:

$$\begin{aligned} H_{\text{biased coin}} &= -(p(\text{heads}) \log_2 p(\text{heads}) + p(\text{tails}) \log_2 p(\text{tails}) \\ &= -((\tfrac{9}{10} \log_2 \tfrac{9}{10}) + (\tfrac{1}{10} \log_2 \tfrac{1}{10})) \\ &= 0.469 \text{ bits} \end{aligned}$$

I used a calculator for this last one. Using my calculator, the two calculations we just did look like this:

```
» -((1/2*log2(1/2)+(1/2)*log2(1/2)))
ans = 1
» -((9/10*log2(9/10)+(1/10)*log2(1/10)))
ans = 0.46900
```

The basic idea behind these results is common sense: the more predictable the outcomes, the less information the source has on average. Shannon was also concerned with communication through "noisy channels," and these are obviously relevant in studying biological communication too, but we will leave this complication aside for now.

4.2 Molecular communication

We have already mentioned on page 15 the possibility of cellular communication among cells in "axes formation," involving the exchange of substances at the molecular level. We mentioned in §0.4.4 that in a well-formed hydra, there is a chemical signalling the presence of a head, and this chemical keeps other cells from growing a new head.

Does this kind of talk about "signalling" or "communication" really make sense? Given the very general ideas about information sketched in the previous section, it does. On this

approach, perception of environmental conditions generally is a kind of information transmission. So, though it sounds odd to say: if having a head or not is equally probable, then the presence of the detectable chemical in the hydra carries 1 bit of information.

We also talked about DNA as a language. Does the sequence of bases in DNA carry information? It does, but it is hard to tell how much, since we do not know how probable each particular sequence is. If the whole DNA sequence were formed by a random choice from the four bases, then a sequence of length x is one out of 4^x possibilities, and so the sequence would have this quantity of information:

$$i(\text{sequence}) = \log_2 \frac{1}{\frac{1}{4^x}} = \log_2 4^x$$

So if $x = 1$ then there would be 2 bits, if $x = 3$ there would be 6 bits, and if $x = 6$ there would be 12 bits. So if there are $3\frac{1}{4}$ billion bases, as in the human genome, there would $3\frac{1}{4} \times 2 = 6\frac{1}{2}$ billion bits, if each base was randomly and independently chosen. But we have seen over and over that it is not the case that each base is randomly chosen. There is no way that copies like the one we saw in section §2.5.4 could happen by random choices of each base. So the fact is, it is obvious that a good deal of information is encoded in the genome, but we do not know how much.

4.3 Non-human animal communication

As in the coordination of activities at the molecular and cellular level, we expect to see animal communication when some kind of coordinated group activity is valuable or essential for achieving some survival or propagation-related goal.[1] Among multicellular organisms, we find communication systems with special properties, quite different than the signalling between cells and genes, and so it is no surprise that in the quotes from biologists on page 112, we see the biologists adding things to the definition of what they want to count as a language: it is signalling with some kind of benefit to the organism, or signalling that is "not accidental" in some other way.

4.3.1 honeybees, *Apis Mellifera*

Many species communicate in the cooperative effort to obtain food, but the dance of the honeybee is perhaps the most amazing non-human example of an informationally rich signalling system. The richness of this system was studied by zoologist Karl von Frisch in the 1940's, and it is still being actively studied.

Honeybees typically live in hives containing 1 queen bee, 200 or so male drones, and 20,000-100,000 female worker bees. A queen is a female bee that was selected in a previous nest for special care (extra food and a larger cell). A queen lives approximately 2 years and can lay

[1] In computer science and in many engineering settings, there have been theoretical studies of how much information is required between parallel processes in order to achieve some goal. Some of the results here are fundamental, and have a bearing on biological systems too (Breen et al., 1999; Klavins, 2002).

200,000 eggs per year, the drones live only 1-3 months to mate and then die, and the worker bees also live only a few months to work and then die. The workers are kept uninterested in reproduction by a pheromone secreted by the queen. Besides tending and defending the nest, the workers gather nectar and pollen which they bring back to the nest for use and storage. The bees compound eyes detect light in a different range from ours, insensitive to red but sensitive to ultraviolet, and sensors on their antennae detect fragrances. Many flowers depend on bees for cross-pollination and so they have evolved bull's-eye shapes, colors and fragrances to attract the bees.

A bee visits flowers in the morning and if one is found to be especially rich in nectar, she will look for other flowers of the same type, and will sometimes return to the nest and do a dance that indicates the distance, direction, and quality of the food. The food sources are sometimes quite distant, even miles away. This means that the bee must have some kind of cognitive map of the environment sufficient for keeping track of the relative position of the hive and the food sources. There is evidence that the other bees in the hive attend to visual, auditory, olfactory and tactile cues in the dance.

To figure out what the bee communicates in a dance, it is useful to consider what the bee knows about the location of the food source. There are two basic kinds of navigation strategies: **dead reckoning**, which is where you set a direction and travel at a certain rate for a certain time, and the other, **landmark navigation** involves going from one "mapped" landmark to the next on some kind of "cognitive map." Most animal navigation involves a mixture of the two.

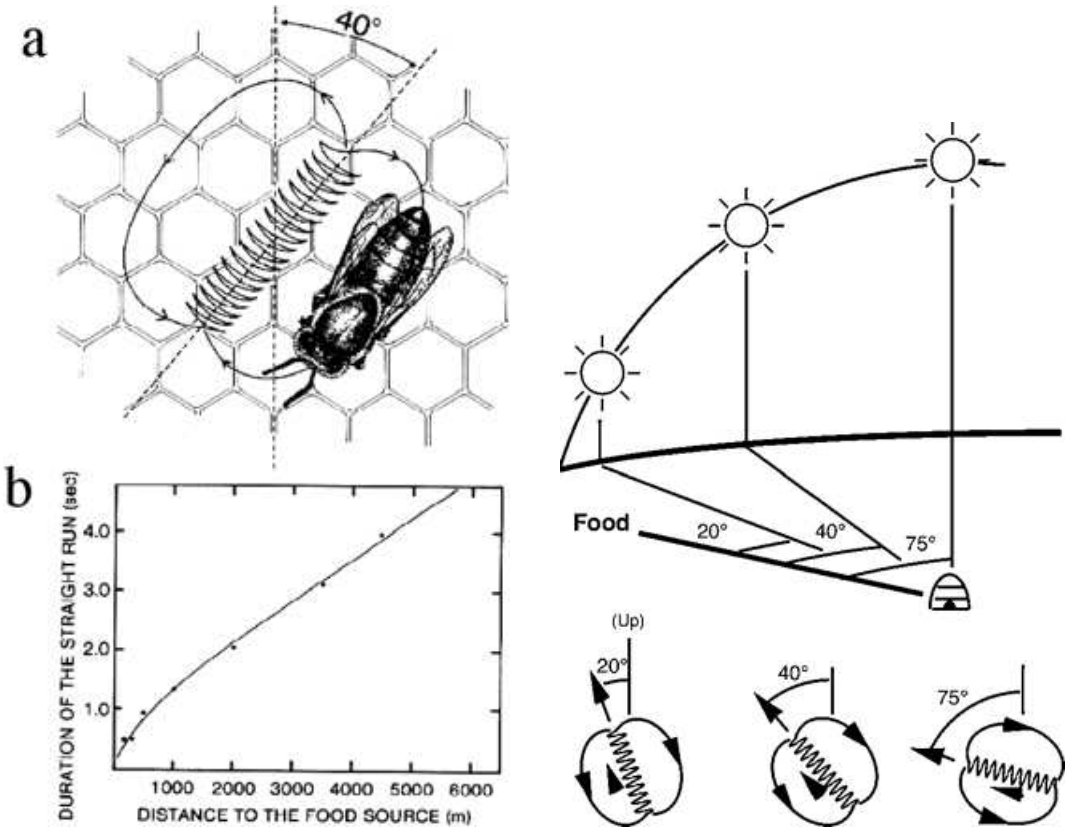
Experiments have demonstrated that when bees find a good food source, they remember the time of day and many properties of the source, and they are likely to return to the same place at the same time the next day.[2] This is adaptive, since flowers vary in the time of day at which they produce nectar. Some plants produce nectar only in the mornings, while others continue through the day. How do they identify the "locations" they remember? Do they have some kind of geometric, spatial map, or do they just remember some sequence of flying motions that got them to the position where they are? It turns out that they have a cognitive map, and this map is based at least partly on sensing motion – accelerations and durations. If you capture a bee and put it into a dark, airtight box, and move to a different position relative to the nest before releasing it, the bee can still fly almost directly to the nest (Gould, 1986). In another study, bees were captured and driven 20 kilometers away – about an hour drive – and then released. Most of the released bees made it back to the nest the same day (Janzen, 1971). This suggests that they use inertial dead reckoning at least in part.

It is also possible to show that bees also use landmarks, the position of the sun, and have a cognitive map of features around the nest. Recent studies have shown that a bee's estimate of distance from the nest is influence by what they see along the way, so that they overestimate distances flown through narrow tunnels. One of the most amazing demonstrations of their cognitive map comes from a discovery that when the dance indicated a food source on the opposite side of a lake, other bees were recruited to go to that location, but when the dance

[2]Some clues about the molecular basis of such timing abilities have recently been discovered. Neurons whose activity varies independently with a circadian rhythym have been identified (Pennartz et al., 2002), and Morré et al. (2002) have discovered proteins in plant and animal cells, "ECTO-NOX proteins," whose state oscillates regularly on a 24 minute cycle.

indicated a food source in the middle of a lake (because the food had been provided from a rowboat in the middle of the lake), the other bees in the hive paid no attention (Gould and Gould, 1988).

How does the bee's dance indicate a food source? Von Frisch noticed that the bees do a "round," circling dance for nearby food sources, and a "waggling," figure-8 dance for more distant sources. In a waggle dance done in a dark nest, direction is indicated by the orientation of the dance relative to gravity: an upward dance means straight towards the sun, downward means away from the sun, and angles in between interpreted as direction off that line. Distance is indicated by an elongated figure-8 for the more distant food sources, as shown in the picture here. Studies show that the bees dance is accurate to within 20° and direction is accurate to within 15%. The bee will not dance if the discovered food is superfluous in the hive, and the bee will not dance when their is no "audience:" it is a social activity.



from (Dyer, 2002)

Why do bees dance? The dancing behavior is not learned, but is entirely innate. And clearly the dance carries information, but we do not understand it well enough to quantify how much information it has. It appears that the dance is a ritualized kind of reenactment of the flight to the food source, and one naturally assumes that this reenactment might increase fitness since a hive should do better when good food sources are reported (Sherman and Visscher, 2002).

4.3.2 non-human primates

Compared to insects, it is no surprise that primates show quite different kinds of communication, related to a much wider range of activities: care-elicitation, alarm, food, and sex (competition, courtship). Monkeys and baboons have been studied quite extensively by Cheney and Seyfarth (formerly at UCLA, and now at the University of Pennsylvania), These animals are social, and show clear awareness of both social and family relationships (Cheney and Seyfarth, 1990; Cheney and Seyfarth, 2005).

Socializing grunts.

Baboons make relatively quiet grunting noises during their activities and these grunts seem to play various roles (Rendall et al., 1999). For example, there is one specific kind of grunt that apparently indicates a wish to reconcile after a fight. Baboons making these reconciliatory grunts were tolerated after a fight significantly more than baboons not making these grunts (Cheney and Seyfarth, 1997).

Contact barks.

Cheney, Seyfarth, and Palombit (1996) have shown that when a group of baboons is dispersed, they make loud "contact barks," especially when they are near the periphery of the group. Other baboons in the group, though, do not seem to answer these barks (Fischer et al., 2001), which leads Cheney and her collaborators to conclude that non-human primates cannot empathize with others; they cannot attribute mental states to other individuals.

Alarm.

Baboons and monkeys also make alarm calls when they spot predators or other dangers (Cheney and Seyfarth, 1996). They make a "sharp bark" in response to various predators, and females seem to make a different bark in response to crocodiles and snakes – a "crocodile bark."

These calls have been of particular interest because they raise the question of whether these calls are *referential* in the sense that they are recognized as indicating a particular kind of predator. Studies of baboons and monkeys have supported the conclusion that the calls are referential in this sense (Zuberbühler, Cheney, and Seyfarth, 1999). Similarly "referential" alarm calls have been found in other primates, including the mongoose (Manser, Seyfarth, and Cheney, 2002), ground squirrels and prairie dogs (Slobodchikoff et al., 1991).

Care-elicitation.

In a study of baboon contact barks, Rendall, Cheney, and Seyfarth (2000), noticed that adult females bark when they get separated from the group. In a study of vervet monkeys, Hauser (1989) also found that infants call when they want to be carried or to nurse. Mothers can

recognize the calls of their infants, and the calls sometimes trigger maternal retrieval and care-giving, but the mothers and infants do not seem to call back and forth in any kind of coordinated, "conversational" way.

Social convergence.

The bee dance is not learned. Are baboon and monkey vocalizations learned? Darwin noticed the similarities between human and non-human facial expressions, and recent study confirms that, the difference between threatening and friendly facial expressions seems to be, at least to a great extent, innate. Infant monkeys raised in isolation are still frightened by threatening faces but not by neutral and friendly ones (Sackett, 1970). So what about vocalizations? These are more controversial. A study of squirrel monkeys showed that being raised by a deaf mother or in isolation had little effect on subsequent vocal behavior (Winter et al., 1973). But a study of macaque monkeys showed that the particular auditory qualities of certain cooing sounds are shaped by their environment more than by heredity (Masataka and Fujita, 1989). The degree of "plasticity" in baboon and monkey vocalizations is limited.

Syntax?

In human languages and DNA, the particular sequence of basic elements makes a big difference in the message communicated: this has to do with "syntax," with the structural properties of the languages. Do we see evidence for this kind of syntax in any non-human primate? There is a recent argument that monkeys may use a simple two symbol system (Zuberbühler, 2002). The argument is that in Campbell's monkeys produce a low booming introduction before a certain alarm call, signifying an alert that does not pose a direct threat. Interestingly, another kind of monkey that inhabits the same locale seems to understand these calls too: showing little reaction to the Campbell's boom-introduced alarms, but a strong reaction to the non-boom-introduced alarms.

> …two recent studies suggest that monkeys and apes may effectively increase their vocal repertoire by combining existing calls and assigning these combinations to new contexts. Like many forest monkeys, Campbell's monkeys (Cercopithecus campbelli) give acoustically different alarm calls to leopards and eagles. In less dangerous contexts, they emit a low, resounding 'boom' call prior to the alarm calls. Sympatric diana monkeys (C. diana) respond strongly to the Campbell's monkey alarm calls. They also appear to be sensitive to the semantic changes caused by call combination, because they no longer respond to Campbell's monkeys alarm calls if they are preceded by a boom (Zuberbuhler 2002; see also Robinson 1984; Snowdon 1990). Similarly, chimpanzees frequently combine different call types when vocalizing, and in some cases also supplement calls by drumming their hands and feet against resonant tree buttresses (Mitani, 1993). In the Ivory Coast, male chimpanzees produce three acoustically different subtypes of barks: one when hunting, one when they encounter snakes, and a third, more generic bark type in a variety of different contexts. In two very limited circumstances, when traveling or encountering a neighboring group, the chimpanzees combine a bark with drumming (Crockford and Boesch, 2003). This signal combination has the potential to convey information that is qualitatively different from (and

> more specific than) the information conveyed by a single call type. Depending upon the definition one chooses, these call combinations may qualify as syntactical. Marler (1977), for example, distinguished between phonological syntax, in which call combinations carry a meaning that is more than just the sum of their parts, and lexical syntax, in which the component parts also play functional roles as subjects, verbs, modifiers, and so on. According to this distinction, the call combinations discussed above may be examples of phonological, but perhaps not lexical, syntax (but see Zuberbuhler 2002 for a slightly different view).
>
> (Cheney and Seyfarth, 2005)

We will get to see whether these 2-call combinations are similar to human phonology in the next section.

Summary

We introduced the extremely general, "information-theoretic" notion of communication as the perception of an event that carries information. Although all the animal communication systems discussed here involve events that carry information, it is no wonder that biologists want to add some conditions to what they want to count as "communication."

Comparing molecular communication at the cellular and sub-cellular level, bee dances and the grunts and barks of baboons and monkeys, we seem to have very different systems. We focused on studies of these animals in their natural habitats, and did not consider the recent attempts to teach sign languages to chimps like "Nim Chimpsky" in human and laboratory settings. (We mentioned these earlier, on page 9.) We may return to some details later, but it is obvious that the abilities of these animals are quite different from human linguistic abilities.

It is important to think about how puzzling this is. Chimps can solve problems, and know quite a lot about how things work. For example, a recent study (Hauser and Spaulding, 2006) showed that monkeys with very little exposure to humans realize that a knife can cut an apple but a glass of water can't. And that a knife can cut an apple in half but not put the halves back together. This was shown in a recent study, where an apple was put behind a window, a shade comes down and then a knife or a glass of water is shown being lowered behind the screen and then removed again, and finally the screen is raised, at which point the experimenters recorded how much time the monkeys spent looking at the scene. (See figures below.)

A similar methodology was used to show that monkeys understood that a glass of blue paint can stain a towel, but a knife cannot – even without any training about paints or knives.

The puzzle about the disconnect between produced speech and gestures on the one hand, and the ability to learn new things and solve problems on the other is well described by this passage from (Cheney and Seyfarth, 2005, italics added):

> The discontinuities between production and perception result in an oddly unbalanced form of communication: monkeys (and other animals) can learn many sound-meaning pairs but cannot produce new words, and they understand conceptual relations but cannot attach labels to them …Children's ability to compare another's perceptual state with their own forms the basis of a social referencing system that is integral to early word learning (Bloom

and Markson, 1998; Tomasello, 2003). Although there are precursors to these abilities in the social interactions and communication of monkeys and apes, they remain rudimentary (Cheney and Seyfarth, 1992; Anderson, Montant, and Schmitt, 1996; Tomasello and Call, 1997). Baboons recognize when calls are being directed at themselves and they seem to have some understanding of other individuals' intentions (Cheney and Seyfarth, 1997; Engh et al., 2006). In contrast to the communication of even very young children, however, monkey vocalizations appear designed to influence other individual's behavior, not their attention or knowledge. Although monkeys vary their calling rates depending upon the presence and composition of their audience, they do not act deliberately to inform ignorant individuals, nor do they attempt to correct or rectify false beliefs in others or instruct others in the correct usage or response to calls (Seyfarth and Cheney, 1986). *…In sum, the communication of non-human animals lacks three features that are abundantly present in the utterances of young children: a rudimentary ability to attribute mental states different from their own to others, the ability to generate new words, and lexical syntax.* We suggest that the absence of all three features is not accidental, and that the lack of one (a theory of mind) may explain the lack of the others (words and syntax). Because they cannot attribute mental states like ignorance to one another and are unaware of the causal relation between behavior and beliefs, monkeys and perhaps also apes do not actively seek to explain or elaborate upon their thoughts. As a result, they are largely incapable of inventing new words and of recognizing when thoughts should be articulated.

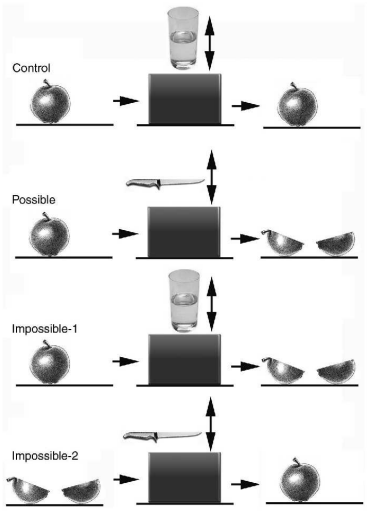


Fig. 1. A conceptual replication of Premack's apple study. Row 1 shows the possible control in which a glass of water is lowered and causes no change to the whole apple. Row 2 shows a possible transformation in which a knife is lowered and causes a transformation of the apple into two half apples. Row 3 shows an impossible transformation in which a glass is lowered and appears to cause a transformation of the apple into two half apples. Row 4 shows an impossible transformation in which two half apples appear to be transformed into a whole apple by a knife.

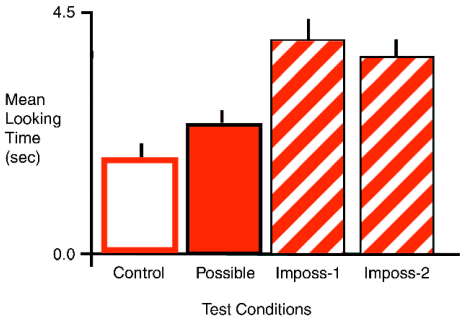


Fig. 2. Mean (+SD) looking time (in seconds) for each of the four conditions involving whole and half apples and either a knife or a glass as transforming agents. Two-tailed *P* value levels are indicated for each contrast based on *t* tests. Imposs-1, the impossible-1 test; Imposs-2, the impossible-2 test.

(Hauser & Spaulding 2006)

Lecture 5

What is a human language?

If I look pale, you can conclude that I have not been sunning myself at the beach very much for the past few weeks. So my looking pale is informative, it carries information, but we do not think of this as an instance of communication. Why not? One reason is that I do not look pale on purpose, but when I speak or make these marks on paper, I am making them with the intention of expressing something intelligible, something any English speaker could understand. But if we impose this requirement on communication, then nothing we have looked at before is clearly communication: the "language" of DNA is not produced because of anyone "intending" to express something; the bee's dance is automatic and it would be strange to think of a bee as having intentions about anything; and even in the studies of the baboons, we noticed that they do not converse. When one baboon calls out because it does not see the others, the others do not answer. Having an intention to communicate something specific to another organism is hard to demonstrate in any non-human. It may happen in baboons or chimps, but it is hard to make the case persuasive. But in humans, this is the rule. We don't even call an informative behavior or trait "communication" unless it is produced with the intention of communicating.

5.1 First observations

Human languages vary: in the biologists' jargon, there is a lot of "plasticity" in this behavioral trait. Speaking one does not enable you to speak another. We will say more about the differences in a moment, but let's first notice important properties that all languages have in common.

5.1.1 All languages: even a child can learn one

Children in a normal speech community, where "normal" can vary quite widely, regularly acquire competence in the language within a few years. And you don't have to be brilliant to do this; even Down Syndrome children can get the basics (Lenneberg, Nichols, and Rosenberger, 1964; Lackner, 1968). Language learning sometimes involves explicit instruction from a care-

taker, but need not do so. Most children learn their first few words before they are 1 year old, and by the time they are 6-8 months they have used 300 or so words: English children learn nouns like *milk, mother, father,…*, verbs like *eat, come, go, put,…*, a few prepositions like *up, down,…*, and some other special elements *yes, more, no, hi, bye-bye, oops,…* At around 18-24 months though, the rate of word acquisition seems to accelerate to 7-9 words a day, continuing at that rate until the child is about 6 years old (Carey, 1977).[1] At 18-24 months, the child usually starts making two word sentences like *want milk* and *big car*, and sometimes three words *no want this*, *the clown do* (Brown, 1973; Clark, 1993).

Another thing we see already from the examples above is that children do not acquire their proficiency in language by rote imitation of sequences they have heard. Children say things like *no want this* even when they have never heard anyone say that before. Even their later, more sophisticated speech could hardly be mistaken for an adult's:[2]

> Go me to the bathroom before you go to bed
> Yawny Baby – you can push her mouth open to drink her

Not only are children not producing sequences that they have heard, even when they are explicitly asked to imitate, they cannot do it. Ervin (1964) says:

> *Omissions bulked large in our cases of imitation. These tended to be concentrated on the unstressed segments of sentences, on articles, prepositions, auxiliaries, pronouns, and suffixes. For example, "I'll make a cup for her to drink" produced "cup drink"; "Mr. Miller will try," "Miller try"; "Put a strap under her chin," "Strap chin."*

Even when there is repeated, explicit correction, a child will have trouble complying: [3]

Child:	Want other one spoon, Daddy.
Father:	You mean, you want THE OTHER SPOON.
Child:	Yes, I want other one spoon, please, Daddy.
Father:	Can you say, "the other spoon"?
Child:	Other…one…spoon.
Father:	Say…"other".
Child:	Other.
Father:	"Spoon".
Child:	Spoon.
Father:	"Other…Spoon".
Child:	Other…spoon. Now give me other one spoon?

Here the child is not getting the point, at least not immediately, but most children get lots of special attention from the adults that care for them: speech directed to the focus of the child's

[1] Estimates of vocabulary size vary for several reasons. How much do you need to know about a word before you can be said to "know" it? Do you need to be able to use it "properly" in all contexts? Do you need to know exactly what it means, or all of its meanings? And even with answers to these questions, it is not clear how to test for this kind of knowledge.

[2] Examples from Melissa Bowerman, reported in Pinker (1994, pp275).

[3] This conversation from Martin Braine, reported in Pinker (1994, p281).

attention, practice and explicit instruction in conversational turn-taking, speech slowed by pauses that are inserted at structurally natural points. And it is no surprise that this happens across cultures: similar things have been found in studies of Kaluli speakers in Papua New Guinea (Ochs and Schieffelin, 1999), Sesotho speakers in South Africa(Demuth, 1986), and many others. It would be surprising if the language abilities of children did not benefit from this kind of special treatment, but there is evidence that children can learn a language even without such special training: children who are unable to speak can nevertheless learn to understand language (Stromswold, 1994), and merely hearing the sound of another language early in life can help you produce those sounds when you try to speak the language as an adult (Au et al., 2001). In any case, the child's abilities show that language is not a trove of remembered sentences, but something that they are creating according to their own principles.

5.1.2 All languages: unbounded complexes

Every human language has expressions of arbitrary size. That is, human languages do not have a 'longest sentence.' For any sentence you take, it is possible to make a longer one. In English, this can be done in many ways. For example, you can prefix almost any sentence with things like "Mary said" or "Fred said":

Grass is green.
Mary said grass is green.
Fred said Mary said grass is green.
Mary said Fred said Mary said grass is green.
…

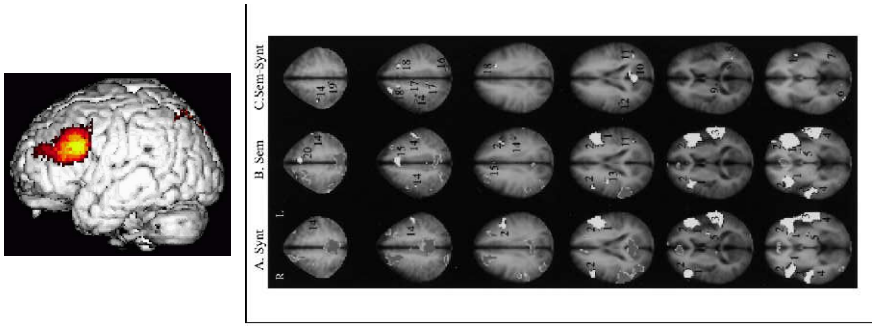
The set of sentences of any human language is infinite in this sense. There is no cutoff point in size.

5.1.3 All languages: fast, automatic analysis

Once you know a language, you cannot help hearing it as language. Even though (as will become clearer later) recognizing the words of a language is a complex task, competent speakers of human languages apparently do it effortlessly. Speakers can show behavioral responses to the meaning of a spoken word in context within 300-500 milliseconds of the onset of the word, which is sometimes before the word is completed (Chambers et al., 2002; Marslen-Wilson, 1975). But to respond to the meaning, the sounds have to be analyzed and classified, the word has to be recognized, and the word has to be related to the context in which it occurs. Another famous phenomenon is called the "Stroop effect" in simple tasks of comparing the colors of stimuli: in recognizing that red is different from blue, subjects cannot help being distracted when the red ink spells "blue," implying that recognition of the word is fast enough to interfere even when the task is explicitly non-linguistic (Stroop, 1935; MacLeod, 1991).

5.1.4 All languages: neural localization

There is also evidence that certain aspects of linguistic performance involve particular parts of the brain. Damage to a certain area of the left front surface part of the brain ("Broca's area") typically produces a complex of symptoms including certain difficulties with the production of "non-content" words like *the, a, of,*.... Activation in this and some other nearby areas can also be detected in electrical potentials on the scalp (event related potentials, ERP), by functional magnetic resonance imaging (fMRI), and positron emission tomography (PET):



left image from (Indefrey et al., 2001), right image from (Embick et al., 2001):

Although we know something about which parts of the brain are essential for linguistic abilities, we do not yet know very much about what computations are carried out, or how. Even something as seemingly simple as the ability to remember a word (or any other perceived event) has remained mysterious. The structure of neurons has been studied carefully, and we know something about how neurons fire and stimulate each other, but how does this activity conspire to encode information about the history of the organism, information that can persist and remain accurate for essentially the whole lifetime the organism? Only just recently are some basics of parts of the "neural code" coming to light, and much remains unclear at both the cellular and molecular level. One speculation is that in perception, there are chemical changes at connections between neurons (synapses) which facilitate or inhibit rates of activation (Rieke et al., 1997, for example).. And there is evidence that protein synthesis at these synapses during and shortly afterward is essential for long-term memory. Some of the genes and proteins apparently involved have been identified, proteins found in the human and the mouse, with homologs in *Drosophila* other animals. A molecule called adenosine 3',5'-cyclic monophosphate (cAMP) seems to play an important role, in cAMP responsive element binding protein (CREB), cAMP Response Element Monitor (CREM), and protein and Activating Transcription Factor (ATF) proteins – apparently important components in the neural plasticity behind long term memory and learning (Davis, 1996; Josselyn et al., 2003), but the coding mechanism supported by these mechanisms is not yet understood.

5.1.5 All languages: structural chunks

When you learn to write, one of the things taught is how to break language up into sentences. But sentence-like units, the sorts of units that can express "a complete thought," are implicit even in the languages of people who have not been taught to read or write. In DNA, triples of bases form codons, and longer stretches formed loops, copies and knots of various kinds. In spoken language, speech sounds form **syllables**, **morphemes**, **words**, **phrases**, **sentences**.

There is a similar "chunking" in visual perception. When we look at a scene like the Darwin's river bank (mentioned in the passage quoted on page 16), a swirl of moving colors hits the retina of our eyes, and this triggers certain reactions in the proteins there (mentioned on pages 12,90), but what we end up seeing is certain objects in a certain spatial arrangement (and often, the objects themselves have parts). And we can recognize whole objects even when they are partly occluded by others. We will see that a similar thing happens in language.

5.1.6 All languages: meaning

We mentioned already (page 7) Frege's idea about how we could possibly recognize the meanings of so many different sentences, most of which we have not heard before:

Semantic Compositionality: New sentences are understood by recognizing the meanings of their basic parts and how they are combined.

Since there is no bound on the size of meaningful expressions in any languages, all human languages must be compositional in this sense.

Human languages have many other properties in common, related to what expressions mean. Every human language has expressions ("names") that refer to particular people and things, and expressions ("verbs") that can combine with names to form an expression that is true or false. Every language provides a way to express *and, or* and *not*. There are many other common properties: surprising restrictions on the kinds of quantifiers human languages have, etc.[4]

5.2 Language structure: English

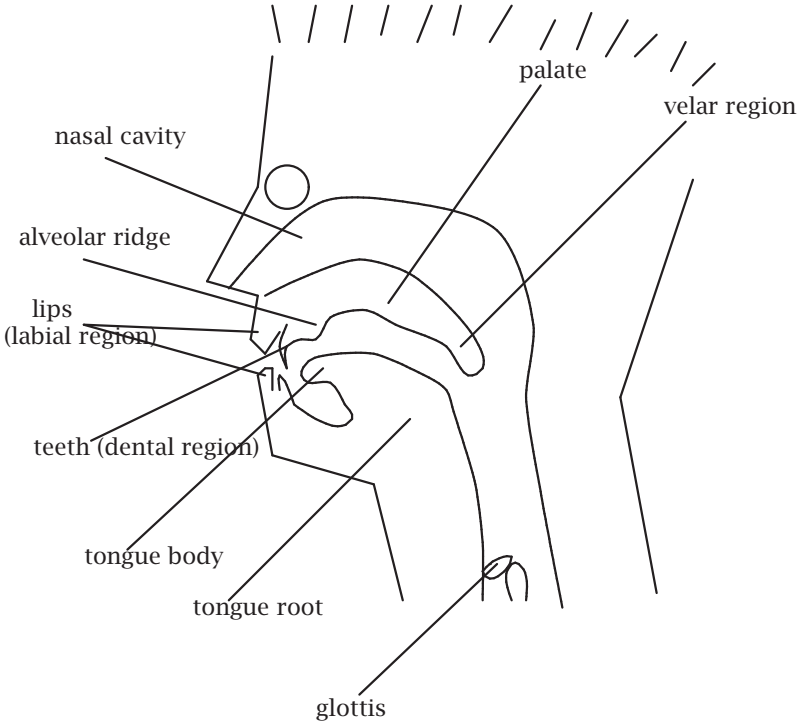
5.2.1 Basic gestures and gestural complexes

In spoken languages, the basic gestures, the basic units of speech sound are called **phonemes**. A phoneme sometimes has variants, as we will see, which are sometimes called **allophones**.

Identifying different phonemes with minimal pairs: Find pairs of different words that differ in a single sound: the differing sounds in these pairs are different phonemes, or variants of different phonemes.

Applying this method to standard American English, we obtain a list of 38 or so basic sounds (the list varies slightly depending on assumptions about which sounds should count as allophones). The sounds are produced by various parts of the mouth, nose and throat:

[4] Most of these will be beyond the scope of this class, but they are one of the standard topics in a class on semantics like Lx 125.



A speech sound that momentarily block the airflow through the mouth is called a **stop**.

			manner	voice	place
1.	[p]	s**p**it	stop	−voice	labial
1a.	[p^h]	**p**it	stop	−voice	labial
2.	[b]	**b**it	stop	+voice	labial
3.	[t]	s**t**uck	stop	−voice	alveolar
3a.	[t^h]	**t**ick	stop	−voice	alveolar
3b.	[ʔ]	bu**t** **'**n (button)	stop	−voice	glottal
4.	[k]	s**k**ip	stop	−voice	velar
4a.	[k^h]	**k**eep	stop	−voice	velar
5.	[d]	**d**ip	stop	+voice	alveolar
6.	[g]	**g**et	stop	+voice	velar
7.	[m]	**m**oat	nasal stop	+voice	labial
8.	[n]	**n**ote	nasal stop	+voice	alveolar
9.	[ŋ]	si**ng**	nasal stop	+voice	velar

The sounds [p] and [p^h] are counted as allophones, variants of the same sound, because switching from one of these sounds to the other never changes one word into a different word. The **fricatives** do not quite block airflow, but constrict air passage enough to generate an audible turbulence. The **affricates** are sound combinations: very brief stops followed by fricatives.

			manner	voice	place
10.	[f]	**f**it	fricative	−voice	labiodental
11.	[v]	**v**at	fricative	+voice	labiodental
12.	[θ]	**th**ick	fricative	−voice	interdental
13.	[ð]	**th**ough	fricative	+voice	interdental
14.	[s]	**s**ip	fricative	−voice	alveolar
15.	[z]	**z**ap	fricative	+voice	alveolar
16.	[ʃ]	**sh**ip	fricative	−voice	alveopalatal
17.	[ʒ]	a**z**ure	fricative	+voice	alveopalatal
18.	[h]	**h**at	fricative	−voice	glottal
19.	[ʧ]	**ch**ip	affricate	−voice	alveopalatal
20.	[ʤ]	**j**et	affricate	+voice	alveopalatal

The **liquids** [r l] and **glides** [j w] have less constriction than the fricatives.[5] Liquids can appear in a **syllabic** form, rather like unstressed [ər əl], indicated with a little mark: [r̩ l̩].

			manner	voice	place
21.	[l]	leaf	lateral approximant	+voice	alveolar
21a.	[l̩]	bottle	syllabic lateral approximant	+voice	alveolar
22.	[r]	reef	(central) approximant	+voice	retroflex
22a.	[r̩] or [ɚ]	bird	syllabic (central) approximant	+voice	retroflex
	[ɾ]	butter	flap	+voice	alveolar
23.	[j]	yet	(central) approximant	+voice	palatal
24.	[w]	weird	(central) approximant	+voice	labiovelar

The vowels are the most “sonorant” of all:

[5] As indicated, we use [r] for the American “r” sound. The standard IPA notation uses [r] for a trill “r”, and uses [ɹ] for the American “r”.

			tongue body height	tongue body backness	lip rounding	tongue root tense (+ATR) or lax (−ATR)
25.	[i]	b**ea**t	high	front	unrounded	+ATR
26.	[ɪ]	f**i**t	high	front	unrounded	−ATR
27.	[u]	b**oo**t	high	back	rounded	+ATR
28.	[ʊ]	b**oo**k	high	back	rounded	−ATR
29.	[ɛ]	l**e**t	mid	front	unrounded	−ATR
30.	[o]	r**oa**d	mid	back	rounded	+ATR
31.	[ɔ]	c**augh**t	mid	back	unrounded	+ATR
32.	[ʌ]	sh**u**t	low	back	unrounded	−ATR
33.	[e]	**a**te	mid	front	unrounded	+ATR
34.	[æ]	b**a**t	low	front	unrounded	−ATR
35.	[a]	p**o**t	low	back	unrounded	+ATR
a.	[ə]	ros**e**s	mid	back	unrounded	−ATR

A vowel which changes quality in a single syllable is a **diphthong**:

36.	[aɪ]	l**ie**s	+ATR
37.	[aʊ]	cr**ow**d	+ATR
38.	[oɪ]	b**oy**	+ATR

The language is not formed just by putting speech sounds in a sequence; there is non-randomness in the sequences of speech sounds. /kkkk/ is not a possible English word. One problem is: it cannot be "syllabified:" The idea that one of the natural units of speech is a **syllable** is familiar from traditional grammars and dictionary entries. Some speech sounds are louder, more "sonorous" than others, from the most sonorous vowels to the least sonorous stops:

The Sonority Hierarchy:

−sonorant			+sonorant			
stops	affricates	fricatives	nasals	liquids	glides	vowels (high,mid,low)

It is traditionally assumed that a **syllable** is formed from zero or more consonants, followed by a vowel, and ending with a shorter sequence of zero or more consonants (but we will see this is an approximation).[6]

> In any succession of sounds, some strike the ear more forcibly than others: differences of *sonority* play a great part in the transition effects of vowels and vowel-like sounds…In any succession of phonemes there will thus be an up-and-down of sonority…Evidently some of the phonemes are more sonorous than the phonemes (or the silence) which immediately precede or follow…Any such phoneme is a *crest of sonority* or a *syllabic*; the other phonemes are

[6] Some prominent approaches to phonology have tried to do without syllables altogether. Among those who accept syllables, it is a matter of controversy whether ASL has anything corresponding to a syllable structures – perhaps it could if vowels were equated with movements, and consonants with held positions…

> *non-syllabic*…An utterance is said to have as many *syllables* (or *natural syllables*) as it has syllabics. The ups and downs of *syllabification* play an important part in the phonetic structure of all languages. (Bloomfield, 1933, p120)

The consonants before the vowel, the vowel, and the consonants after the vowel are called the **onset**, the **nucleus** and the **coda**, respectively, with the nucleus as the only obligatory part.

Let's write:

C for consonants, and
V for vowels and syllabic consonants.

Then the basic parts of syllables are these **38 phonemes + syllabic (V) forms of 2 consonants**:

Basic parts:

p	b	t	k	d	g	m	n	ŋ	f	v	θ	ð
C	C	C	C	C	C	C	C	C	C	C	C	C

s	z	ʃ	ʒ	h	ʧ	ʤ	l	r	j	w	l̩	r̩
C	C	C	C	C	C	C	C	C	C	C	V	V

i	ɪ	u	ʊ	ɛ	o	ɔ	ʌ	e	æ	a	aɪ	aʊ	oɪ
V	V	V	V	V	V	V	V	V	V	V	V	V	V

Using these parts, there is just one way to build a nucleus, but there are three ways to build a the optional coda – depending on whether there are 1, 2, or 3 consonants:

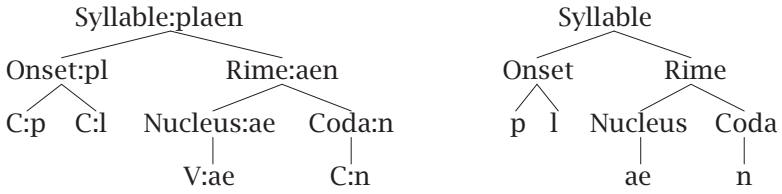
rule0:	$\underset{\text{V}}{x} \mapsto \underset{\text{Nucleus}}{x}$		**'build a nucleus'**
rule1a:	$\underset{\text{C}}{x} \mapsto \underset{\text{Coda}}{x}$		**'(opt) build coda with 1 C'**
rule1b:	$\underset{\text{C}}{x}\ \underset{\text{C}}{y} \mapsto \underset{\text{Coda}}{xy}$		**'(opt) build (allowed) coda with 2 Cs'**
rule1c:	$\underset{\text{C}}{x}\ \underset{\text{C}}{y}\ \underset{\text{C}}{z} \mapsto \underset{\text{Coda}}{xyz}$		**'(opt) build (allowed) coda with 3 Cs'**

Similarly, there are 2 ways to build a rime, 3 ways to build a coda, and 2 ways to build a syllable:

rule2a:	$\underset{\text{Nucleus}}{x} \mapsto \underset{\text{Rime}}{x}$	**'build rime without coda'**
rule2b:	$\underset{\text{Nucleus}}{x}\ \underset{\text{Coda}}{x} \mapsto \underset{\text{Rime}}{x}$	**'build rime with coda'**
rule3a:	$\underset{\text{C}}{x} \mapsto \underset{\text{Onset}}{x}$	**'(opt) build onset with 1 C'**
rule3b:	$\underset{\text{C}}{x}\ \underset{\text{C}}{y} \mapsto \underset{\text{Onset}}{xy}$	**'(opt) build (allowed) onset with 2 Cs'**
rule3c:	$\underset{\text{C}}{x}\ \underset{\text{C}}{y}\ \underset{\text{C}}{z} \mapsto \underset{\text{Onset}}{xyz}$	**'(opt) build (allowed) onset with 3 Cs'**
rule4a:	$\underset{\text{Rime}}{x} \mapsto \underset{\text{Syllable}}{x}$	**'build syllable without onset'**
rule4b:	$\underset{\text{Onset}}{x}\ \underset{\text{Rime}}{y} \mapsto \underset{\text{Syllable}}{xy}$	**'build syllable with onset'**

We will explain what we mean by allowed onset and allowed coda just below.

Using these rules, we can derive the word [plæn] – the word that we spell 'plan' – as shown on the left below. Linguists often use the abbreviated tree on the right:



Remember that the **root node** of the tree is on top – so the tree is upside-down, the way family trees often are. In this upside-down tree, the root has two parts, the onset and the rime. As in the DNA and RNA trees, we call these two parts **daughter nodes** of the root. The right daughter is the rime, which is in turn the **mother** of two more daughters: the nucleus and the coda. And in analogy with a real tree, the **nodes** that are furthest from the root, those along the bottom of the tree, are sometimes called **leaves**.

The possible onsets in English are restricted. (They are restricted in every language, but the exact restrictions vary.) In English:

(1) Any single consonant is a possible onset

(2) Only certain 2-consonant onsets are possible. Since there are 24 consonants listed above, (as our naming rule from page 47 tells us) there are 24^2=576 different pairs

of consonants. But the ones that occur in common English words are just those given by the 32 +'s in this table:

	w	j	r	l	m	n	p	t	k
p		+	+	+					
t	+	+	+						
k	+	+	+	+					
b		+	+	+					
d	+		+						
g	+	+	+	+					
f		+	+	+					
θ	+		+						
ʃ			+						
s	+			+	+	+	+	+	+

Maybe I missed a couple – this chart misses a few words with unusual sounds (borrowings from other languages, etc.). For example, *sphere* begins with the unusual onset [sf], which is not listed in this chart.

(3) The number of different 3-consonant sequences is 24^3=13,824. But in onsets, there are even fewer 3-consonant possibilities than there were 2-consonant possibilities!! I count just these 9:

	w	j	r	l	m	n
sp		+	+	+		
st		+	+			
sk	+	+	+	+		

See if you can think of any I missed.

(4) (Certain other onsets appear in words borrowed from other languages.)

Why are there so few possible onsets from the many possibilities? One idea is this famous one. The onsets and codas in English seem to respect this ordering according to the following principle:

Sonority principle: onsets usually rise in sonority towards the nucleus, and codas fall in sonority away from the nucleus.

This accounts for the impossibility of words with onsets like *rtag*, while allowing *trag*. And it accounts for the impossibility of words with codas like *gatr* while allowing words like *gart*. Similar sonority hierarchies play this kind of role in other human languages too.

Notice that the sonority principle seems to relate to the mechanics of pronouncing each of the sounds in a sequence: it would be hard to pronounce a word beginning with [kt] or [ðp] or [lgt]. So in a sense this pattern of increasing sonority is determined by properties of the sounds and the articulators themselves, not by some influence that is outside of the language. That is, it is a *self-organizing* influence.

A simple rule which properly divides most English words into syllables is this:[7]

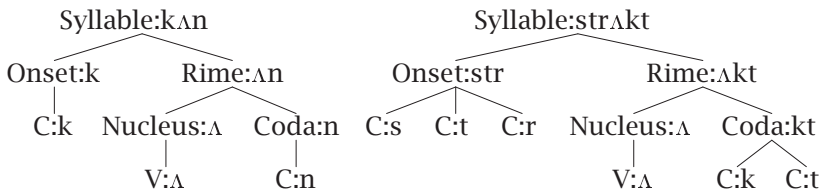
1. each +vocalic phone (vowels and syllabic liquids) is a nucleus.
2. Then, take the longest sequence of consonants preceding each nucleus which can form a possible onset to be the onset for the following nucleus.
3. Take all remaining consonants to be codas of the preceding nuclei.

For obvious reasons, this is sometimes called the "onsets before codas" rule; what it amounts to is: "maximize onsets." This principle also seems to hold across languages. Why would languages generally prefer consonants at syllable beginnings than at syllable ends? Is this a self-organizing influence too? The answer seems to be that yes, this fact may be due to the perceptual cues needed to recognize consonants, and the fact that final consonants are often unreleased (Ohala, 1990; Steriade, 1995).

For example, consider the word 'construct':

/kʌnstrʌkt/

This gets divided into two syllables this way:



[7] This rule works properly for many words (try *matron, atlas, enigma*), but it does not seem to provide quite the right account of words like *apple* or *gummy*. The first syllable of *apple* is stressed, and it sounds like it should include the consonant. Cases like these are called "ambisyllabic:" a consonant is ambisyllabic if it is part of a (permissible) onset but immediately follows a stressed lax (-ATR) vowel.

Morphemes and morpheme complexes

Frege's idea about compositionality (mentioned just above on page 127 and earlier on page 7) predicted that there would be a finite set of basic elements out of which sentences are built, so that we can interpret sentences based on the meanings of the parts and how they are put together. Surprisingly, the basic units of meaning are not phonemes. The phonemes /p/ or /b/ do not mean anything by themselves. The basic units of meaning are called **morphemes**. So the word *book* is a morpheme: no smaller part of it is meaningful by itself. And the word *case* is a morpheme. But the word *bookcase* has two morphemes in it. Also the word *books* has two morphemes: there is the noun morpheme *book*, and the plural morpheme *-s*. The plural morpheme is said to be a **bound morpheme**, because it can only occur attached to something else, while morphemes like *book* and *case* and *study* and *realize* are **free**. (And we see: some of the bound morphemes are one phoneme long! – not even a whole syllable.)

The morphemes come in categories: *book* and *case* and *bookcase* are both Nouns (abbreviated: N) because they can appear in similar positions in phrases:

(5) a. the book was expensive

b. the case was expensive

c. the bookcase was expensive

The word *bookcase* is an example of a **noun compound**: you can form a new noun by putting together two nouns. Sometimes a noun compound is spelled with spaces, and sometimes it is not

(6) a. the bookcase delivery was expensive

b. the bookcase delivery truck was expensive

c. the bookcase delivery truck repair was expensive

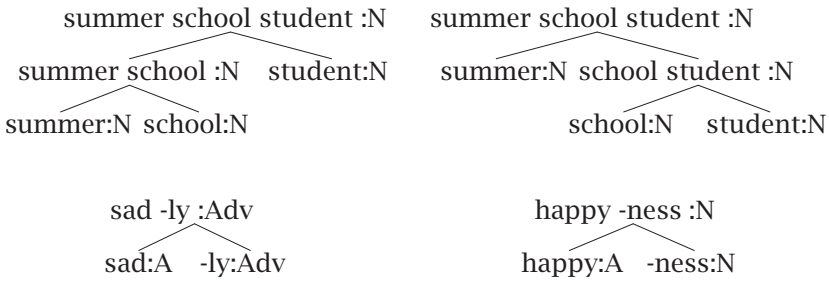
d. the bookcase delivery truck repair manual was expensive

Notice that noun compounding is **recursive** (in the sense already mentioned in the first lecture, on page 9): a noun can be extended to a larger one. We can formulate the rules for combining various kinds of morphemes with the same kind of notation that we used for RNA and DNA and proteins.

Free Morphemes:	(determiner)	D:	the, some, no, a, every, one, two,…
	(noun)	N:	student, penguin, cat, yard,
			kid, school, summer, winter, quarter,…
	(verb)	V:	laugh, cry, fall, sing, dance,…
	(transitive verb)	Vt:	like, praise, sing, teach, see,…
	(tense,modal)	T:	will,would,can,could,…
	(adjective)	A:	happy, sad, probable, rare,…
	(adverb)	Adv:	always, sometimes,…
	(preposition)	P:	in, on, with, about, near, by, from, to,…
Bound Morphemes:	(number)	Num:	-s
	(adverb)	Adv:	-ly
	(noun)	N:	-ness
	(noun)	N:	-er

Morphology:	x N	y N	↦	xy N	**noun compound rule**
	x N	-s Num	↦	x-s N	**plural rule**
	x A	-ness N	↦	x-ness N	**-ness rule**
	x A	-ly Adv	↦	x-ly Adv	**-ly rule**
	x Vt	-er N	↦	x-er N	**-er rule**

With these rules there are two ways to derive *summer school student*, but just one way to derive the adverb *sad-ly* and the noun *happy-ness* (which is actually spelled: *happiness*):



The two ways of deriving *summer school student* correspond to two different interpretations: the one on the left refers to a student who goes to summer school, but the one on the right is less natural: it refers to a school student from the summer, or something like that. Similarly for *leather football*, the natural derivation puts together *foot+ball* and then adds *leather*, but you could also put together *leather+foot* and then add *ball*. This would mean: a ball for leather feet (whatever that would be!).

English has many other suffixes, and it has prefixes, and both can occur in a single word:

other suffixes: peace-ful, forget-ful, kill-er, modern-ism, reptil-ian, orphan-age, defend-ant, annoy-ance, money-ed, neighbor-hood, class-ify, intens-ify, boy-ish, art-ist, restrict-ive, symbol-ize, ghost-ly, establish-ment, advis-ory, spac-eous, honest-y, assembl-y, robber-y, snow-y, natur-al

other prefixes: dis-engage, pre-test, un-qualified, in-accurate

combinations: modern-ist-ic, class-ifi-catory, nation-al-iz-ation, anti-dis-establish-ment-ar-ian-ism, anti-missile, anti-anti-missile

Alternative perspectives on morphology

The idea that morphemes are semantic atoms is largely satisfactory, but leads to unsatisfying accounts of some small things. First, there are idioms. Let's use the term 'idiom' to refer to something that looks like it is a complex of morphemes, but its meaning is not determined by the meanings of its parts. There are lots of familiar phrasal idioms like *your goose is cooked* or *they keep tabs on me* or *they swept it under the rug*. But with our definition of 'idiom', some compound words are idioms too. For example, someone who knows what *book* means and what *case* means could probably make sense of the term *bookcase*. But someone who knows what *sun* means and what *flower* means will not know what *sunflower* means, because it refers to a particular kind of flower. So *sunflower* is an idiom. So are *blueberry, deadline, monkey wrench, student body, red herring*. And with this definition of idiom, every idiom is a semantic atom – an expression whose meaning is not determined by the meanings of its parts. Nevertheless, we think of *goose be cooked* has having several morphemes. In what sense are those things morphemes, in the idiomatic context?

Another puzzle comes from words like *cranberry* and *boysenberry* and *huckleberry* – the units *cran-* and *boysen-* and *huckle-* are not usually regarded as meaningful. In *lukewarm* we know what the *warm* means but what is *luke-*? In *unkempt* and *uncouth*, we know what *un-* means, but what is *kempt* or *couth*? In *immaculate* and *impeccable*, we seem to see the the same *im-* that we see in *imprudent, impossible, immobile*, but what is *maculate* or *peccable*? These considerations suggest that some units that combine with morphemes may not be meaningful. So should we call them morphemes too?

A third puzzle for the idea that morphemes are semantic atoms comes from a puzzle about how morphemes are learned. A standard idea that fits well with the semantic atom conception is this one from a recent article:

> To learn that cat is the English-language word for the concept 'cat,' the child need only note that cats are the objects most systematically present in scenes wherein the sound /kæt/ is uttered (just as proposed by Augustine (398); Locke (1690); Pinker (1984); and many other commentators). (Snedeker and Gleitman, 2004)

What Augustine actually said in his *Confessions*, written around 398, was this:

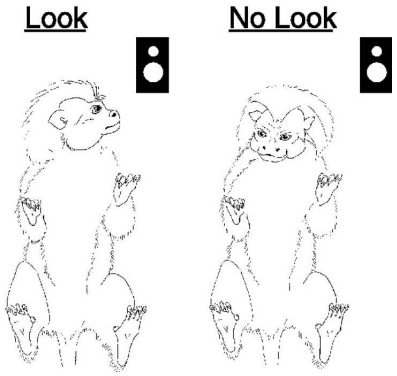
> When they (my elders) named some object, and accordingly moved towards something, I saw this and I grasped that the thing was called by the sound they uttered

> when they meant to point it out. Their intention was shown by their bodily movements, as it were the natural language of all peoples: the expression of the face, the play of the eyes, the movement of the other parts of the body, and the tone of the voice which expresses our state of mind in seeking, having, rejecting or avoiding something. Thus, as I heard words repeatedly used in their proper places in various sentences, I gradually learned to understand what objects they signified; and after I had trained my mouth to form these signs, I used them to express my own desires.

This sounds sensible, but notice that the learner faces three big difficulties on this approach: (i) the learner needs to know what the speaker means, in order to do this kind of correlation; (ii) since languages have lots of 'homophony', the learner needs to realize that different uses of expressions that sound exactly the same might mean completely different things (e.g. *there, their, they're*); and (iii) the learner needs to figure out which sequences of sounds are morphemes – where each word begins and ends.

There is a different conception of morphemes which does not say that they have to be semantic atoms, taking care of the problem with idioms and with *cranberries*, and which suggests a different way of determining where the edges of the words are, making things less difficult for the learner. The idea is basically the commonsense one that morphemes are commonly occurring units. A famous linguist, Zellig Harris, proposed this idea in the 1950's, suggesting that the morpheme boundaries are the places where it becomes relatively harder to predict what will come next. This idea has been developed in recent work by Goldsmith (2001), Brent (1999), and others.

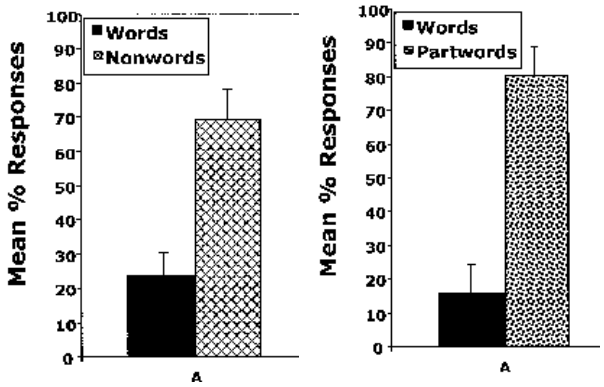
Recent studies show that not only human children (even at 7 months old), but also monkeys and other animals can notice chunks of this sort – sequences of sounds that usually go together. For example, one study (Hauser, Newport, and Aslin, 2001) played words to monkeys from a speaker, and noticed that when a new, unusual word is played the monkeys tend to look at the speaker, showing that the new word has caught their attention.



For 20 minutes one day, the monkeys heard the following 'training words' in random orders, with no pause at all between the words – the timing between the syllables of one word and the syllables of the next word were carefully controlled to provide no cues about word boundaries:

Training ords:	tupiro, golabu, bidaku, padoti
Test words:	tupiro, golabu
Test non-words:	dapiku, tilado
Test part-words:	tibida, kupado

Then the next day, by looking at the speakers, the monkeys showed that they were not suprised to hear 'test words' from the day before, but they more surprised to hear nonwords. But what is more interesting is that they were also surprised to here 'test part-words', which were constructed from the end of one word and the beginning of another. That is, the monkeys learned the word boundaries even without pauses between the words, just because there is more variation in what sounds appear next at the word boundaries (4 possibilities, randomly selected) than at syllable boundaries inside the words.



Rules for building sentences (TPs) and other phrases

We have seen that a sequence like:

the school teachers will sing happily

has 9 morphemes:

the school teach-er-s will sing happi-ly.

But it has 5 words, since *school teach-er-s* is a N(oun), and *happi-ly* is an Adv(erb).

the [school teach-er-s] will sing [happi-ly].

Above the level of words, we have larger units: **phrases** of various kinds. The subject of the sentence **the school teach-er-s** is a **Determiner Phrase (DP)**, and *sing happi-ly* is a **Verb Phrase (VP)**. Notice that English sentences always have to be tensed:

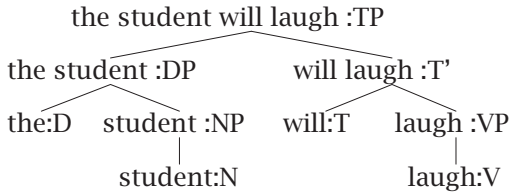
(future)	the [school teach-er] will sing happi-ly
(present)	the [school teach-er] sings happi-ly
(past)	the [school teach-er] sang happi-ly
(no good!)	* the [school teach-er] sing happi-ly

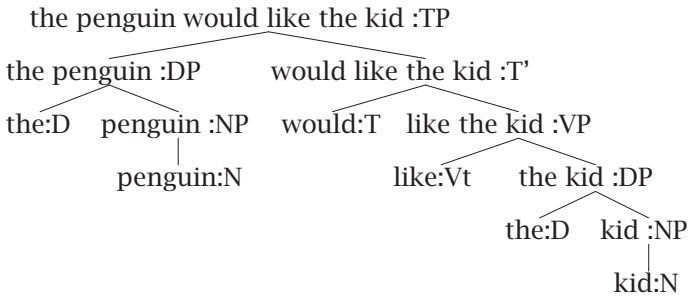
For this reason, we call a sentence a **Tense Phrase (TP)**, and we call *will sing happily* a **Tense-bar Phrase (T')**. The T' would usually be called "the predicate," but we call it T' since it has the tense in it, but is not the complete TP until the subject DP gets added.

Here is a simple set of 5 rules that lets us define the language with these structures in it:

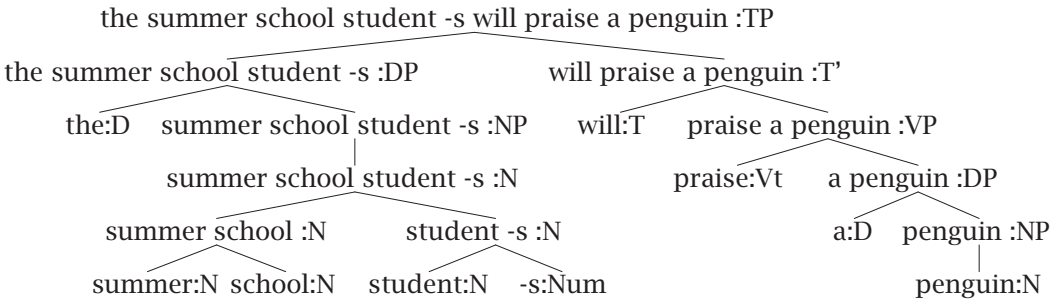
Syntax:	x X	↦	x XP	**X to XP, for X=N,A,Adv,V**
	x y Vt DP	↦	xy VP	**Vt takes DP object**
	x y D NP	↦	xy DP	**D takes NP object**
	x y T VP	↦	xy T'	**Tense with VP makes T′**
	x y DP T'	↦	xy TP	**T′ with subject DP makes TP**

Using this grammar exactly as we used the grammars for DNA, RNA and Proteins, we have derivations like this:





The derivations above use only the 5 rules for building phrases, but sometimes we also need to use the previous rules for building words, as in the following example:

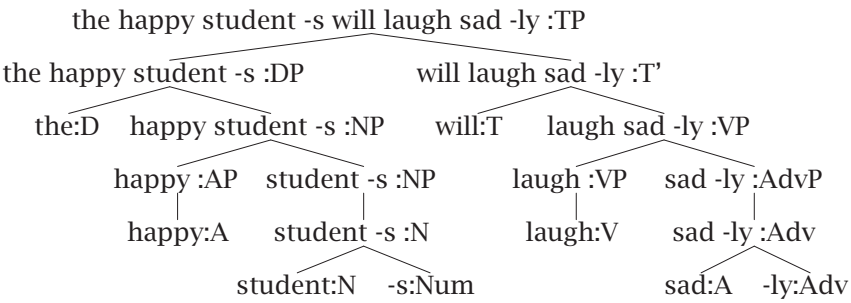


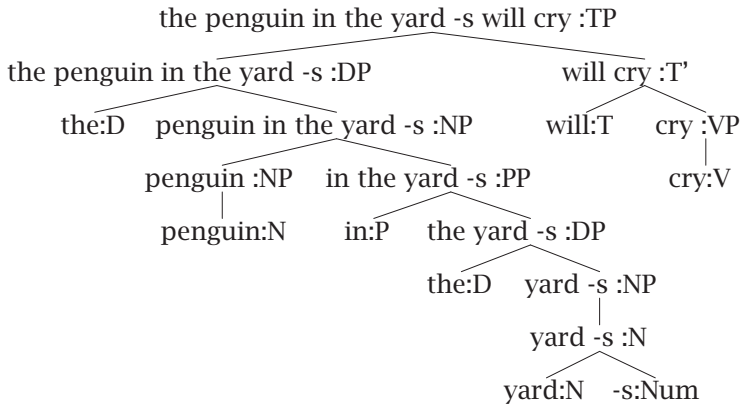
Rules for building and using <u>modifiers</u>

Let's add one more thing: modifiers. An **Adjective Phrase (AP)** can modify a noun phrase (NP). An **Adverb Phrase (AdvP)** can modify a verb phrase (VP). And a **Prepositional Phrase (PP)** can modify either a noun phrase (NP) or a verb phrase (VP). A PP is formed by putting a preposition (P) together with a determiner phrase (DP):

A first syntax of modifiers:	x AP	y NP	↦	xy NP	**AP modifies NP**
	x VP	y AdvP	↦	xy VP	**AdvP modifies VP**
	x NP	y PP	↦	xy NP	**PP modifies NP**
	x VP	y PP	↦	xy VP	**PP modifies VP**
	x P	y DP	↦	xy PP	**Prep takes DP object**

When an AP or PP modifies an NP, you still have an NP as the result – you just know more about it. And the same goes for VP modifiers: when an AP or PP modifies a VP, you still have a VP. Applying these rules, we can derive:





Exercises

Two notes:

- Use the phonetic notation and the rules introduced in this class. (Dictionaries and other classes may have used slightly different notations, but part of the exercise here is to use exactly the notation and rules we have introduced)
- If there is more than one structure, draw the most natural one (as discussed on page 136

This week we introduced phonemes, and the rules for making syllables out of them. And we introduced morphemes, the rules for making larger words out of morphemes, and rules for making phrases out of words. The problems this week test whether you understand how these rules work.

1. **Phonemes and syllables:**
   a. What does this American English say:

      ðə ætʌmz ar fonimz

   b. Draw the syllable structure for (all the syllables of) the last word
   c. Write the American English pronunciation of the following phrase in phonetic notation:

      she reads about syllables

2. **Phonemes and syllables:**
   a. What does this say:

      ju wɪl bi əsɪmɪletəd

   b. Draw the syllable structure for (all the syllables of) the last word
   c. Write the American English pronunciation of the following phrase in phonetic notation:

      he said go ahead, make my day

3. Using the rules and morphemes from the notes and handout, show the derivation of the following, in a tree:

the students will like the summer quarter

4. Using the rules and morphemes from the notes and handout, show the derivation of the following, in a tree:

   the singer would sing rarely in the winter

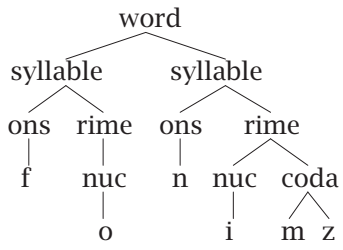
5. Using the rules and morphemes from the notes and handout, show the derivation of the following, in a tree:

   some happiness could teach every penguin

Solutions

1. **Phonemes and syllables: (some variation in pronunciation OK)**

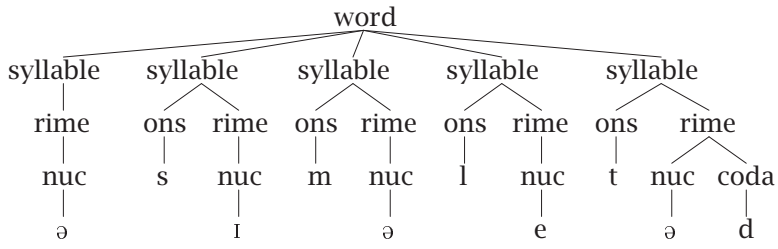
   a. The atoms are phonemes

   b.


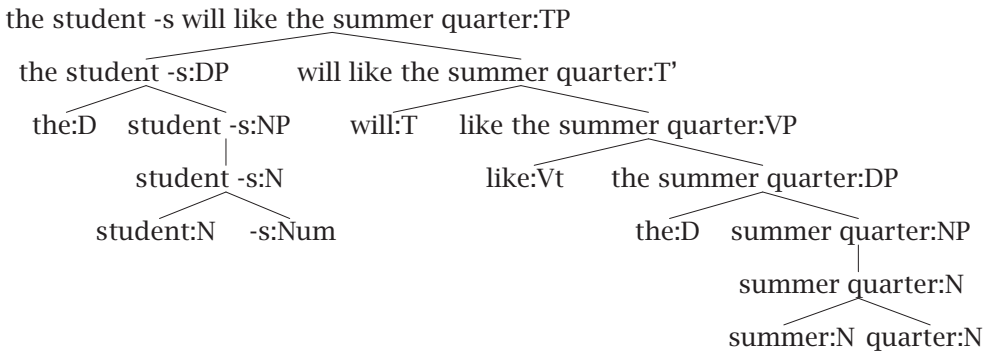
   c. [ʃi ridz əbaʊt sɪləbl̩z]

2. **Phonemes and syllables: (some variation in pronunciation OK)**

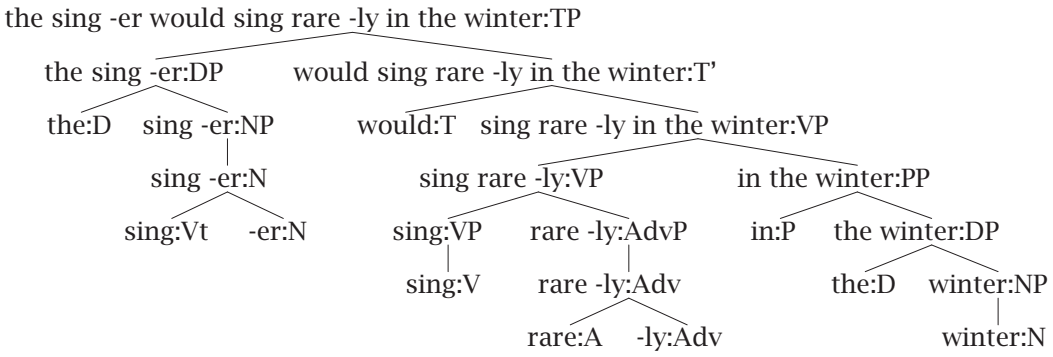
   a. you will be assimilated

   b.


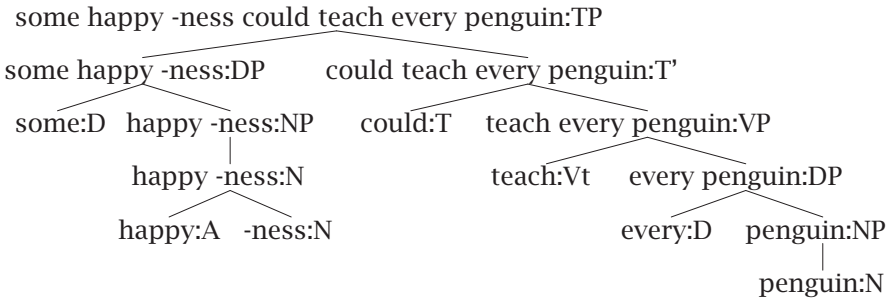
   c. [hi sɛd go əhɛd mek maɪ de]

3.


4.


the sing -er would sing rare -ly in the winter:TP
the sing -er:DP
would sing rare -ly in the winter:T'
the:D
sing -er:NP
would:T
sing rare -ly in the winter:VP
sing -er:N
sing rare -ly:VP
in the winter:PP
sing:Vt
-er:N
sing:VP
rare -ly:AdvP
in:P
the winter:DP
sing:V
rare -ly:Adv
the:D
winter:NP
rare:A
-ly:Adv
winter:N

5.


some happy -ness could teach every penguin:TP
some happy -ness:DP
could teach every penguin:T'
some:D
happy -ness:NP
could:T
teach every penguin:VP
happy -ness:N
teach:Vt
every penguin:DP
happy:A
-ness:N
every:D
penguin:NP
penguin:N

5.2.2 Language structure: a better model of English

Our rules for English can define a simple part of the language, and we can notice already some simple general properties:

1. The word-building rules all combine two things with some categories, X and Y, and in almost every case, they yield a Y. That is, the category of the result is usually determined by the category of the constituent on the right. This is sometimes called the **right hand head rule** for English words.

2. Above the level of words, the first phrase building rules are of two kinds: either they take an X to make an XP, or else they combine an X and a YP to make an XP. Linguists call the relation between the X and YP in these cases is called **selection**: Vt selects DP on the right; D selects NP on the right; T selects VP on the right; and then one case that goes the other direction, T' selects a DP subject on the left.

   The second set of phrase building rules were the **modifier rules**, and in each of those cases one phrase XP modifies a YP, and so the result is a modified YP.

While languages vary in many ways, it turns out that with respect to properties like these, the variation is much more limited.

We need to take one more step to see an aspect of language that some linguists regard as fundamental, a step that will preserve the basic features mentioned above. This important step can be motivated by noticing some things that the grammar above misses:

- Our grammar does not give us simple present tense sentences, and notice that the present tense marker *-s* is in a different place from the future tense marker *will*:

  we get: *the penguin will fall*
  but not: *the penguin fall-s*

- Our grammar does not generate any questions, even simple yes/no questions like this:

  we get: *the penguin will fall*
  but not: *will the penguin fall?*

  The verb seems to have moved from its usual position!

- Our grammar does not let us use auxiliary verbs like *have* and *be*

we get:	*the penguin will fall*
but not:	*the penguin have-s fall-en*
	the penguin will have fall-en
	the penguin will be fall-ing
	the penguin will have be-en fall-ing

This last problem caught the attention of the linguist (Chomsky, 1956). Our rules can relate subject (DP) and predicate (T'), and they can relate a determiner (D) and a noun phrase (NP), but only when these things are right next to each other. The new examples just above suggest two more surprising things:

i. *have…-en* and *be…-ing* are parts of the sentence too, even though they are not adjacent to each other. Furthermore, in examples like the last one, the *have…-en* dependency crosses the *be…-ing* dependency.

ii. to properly formulate a rule that relates simple sentences to the corresponding yes/no questions, it is natural to use a rule that builds not just strings, but strings **with some structure**.

For example, if we build a VP like *see the student*, then we can add the future tense *will* to the front, but the present tense *-s* would have to get added to the middle, after the verb. To avoid this problem, and to allow yes/no questions, we can split our strings as we did in the RNA and DNA languages in §2.5.4. When we put a V like *see* together with a DP like *the student*, we can keep the strings separate, producing the **pair of strings** (*see, the student*). Then we can still put a suffix after the verb if we need to.[8]

It is not hard to see how this would work, by slightly revising some of our first rules. In fact, we already used the same technique to define crossing dependencies in RNA.

Predicates with the auxiliary *be* are sometimes called "progressive," and predicates with the auxiliary verb *have* sometimes called "perfective," so we use the new categories **Prog** and **Perf** for these.

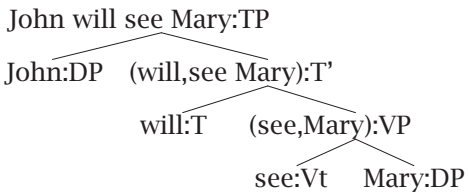
We modify our earlier rules to make a VP with 2 parts and a T' with 2 parts.[9] The rules (5,6) that move the suffixes into place and form Yes/No Questions are often called **movement rules**: they change the usual order of the words. We use the morphemes we had before (page 136), and we introduce few new ones. We use ϵ for **empty** parts of phrases.

[8] Chomsky, Joshi and many other linguists use rules that build and modify trees, but here pairs of strings suffice. Our rules here and in §2.5.4 are MCFG rules (Seki et al., 1991). Their relation to tree-transforming grammars is discussed in (Weir, 1988; Michaelis, 1998; Harkema, 2000; Stabler, 2001).

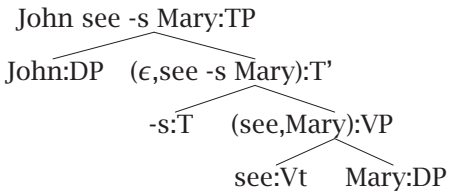
[9] If you study more syntax, you will see that we have split the VP and T' into their *head* and *complement* strings. In a more sophisticated theory, all categories are split into three parts: *specifier, head* and *complement*, plus any other components that are moving. The rule 5b is often called *affix hopping*, and the rules 4a,b, 5a,c, and 6b involve *head movement*.

New Morphemes:				(names)	DP:	John, Mary, Bill, Sue,…
				(progressive)	Prog:	(be,-ing)
				(perfective)	Perf:	(have,-en), (have,-ed)
				(tense)	T:	-s, -ed
Revised Syntax, with "Movement":						
	x X	↦	x XP	**X to XP, for X=N,A,Adv**		(1)
x D	y NP	↦	xy DP	**D takes NP object**		(2)
x Vt	y DP	↦	x,y VP	**Vt takes DP object (2 parts!)**		(3*a*)
	x V	↦	x,ϵ VP	**V to VP (no object, but still 2 parts)**		(3*b*)
x,y X	z,w VP	↦	x,zyw XP	**if X=Perf,Prog, XP has 2 parts**		(4*a*)
x,y Perf	z,w ProgP	↦	x,zyw PerfP	**PerfP has 2 parts**		(4*b*)
x T	y,z X	↦	x,yz T'	**if T is a <u>word</u> and X=VP,ProgP,PerfP**		(5*a*)
x T	y,z X	↦	ϵ,yxz T'	**if T is a <u>suffix</u> and X=VP**		(5*b*)
x T	y,z X	↦	yx,z T'	**if T is a <u>suffix</u> and X=ProgP,PerfP**		(5*c*)
x DP	y,z T'	↦	xyz TP	**Sentence: build TP as usual**		(6*a*)
x DP	y,z T'	↦	yxz TP	**Y/N Question: if y not empty**		(6*b*)

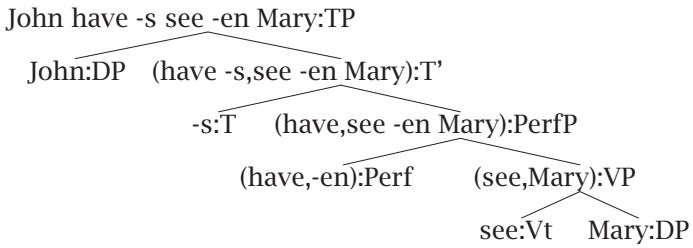
With these rules, we can derive a simple sentence like *John will see Mary*, much as before, but now the categories VP and T' are pairs of strings:



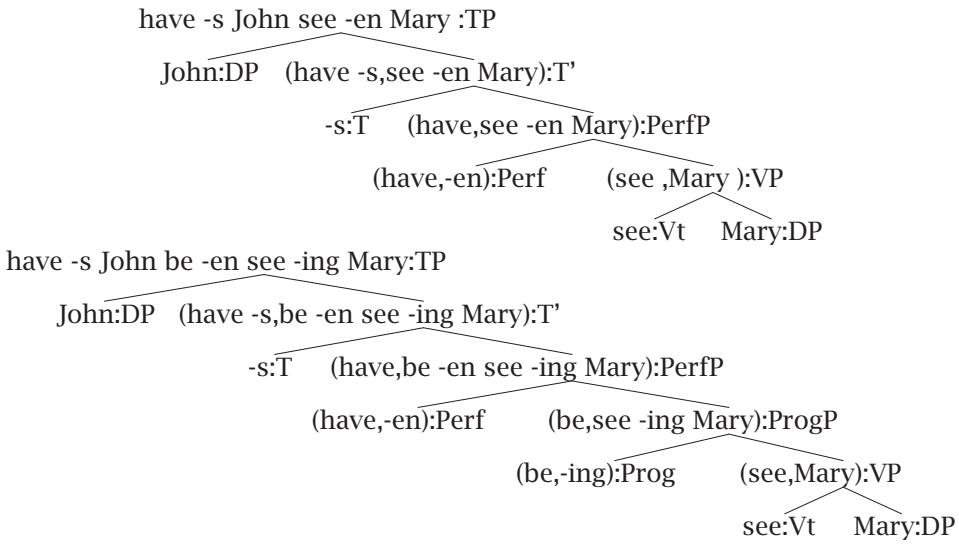
If we have present tense *-s* instead of the future *will*, the derivation looks similar, but notice how the tense affix moves onto the verb:



The affix *-en* is similarly attached to the appropriate verb in a derivation like this:



And we can form yes/no questions:

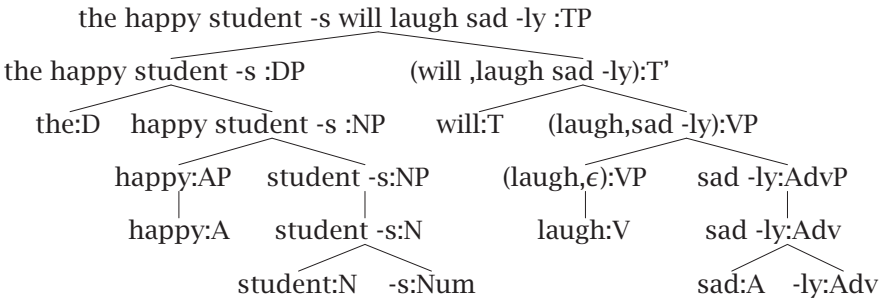


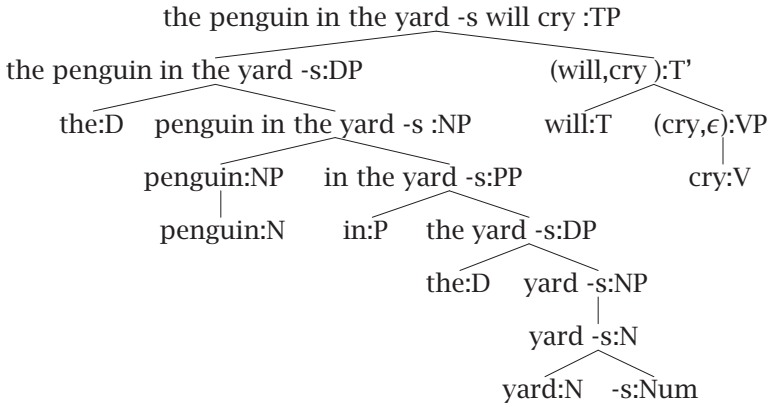
We can adjust the modifier rules so that VP has a pair of strings, and they will cover the sentences considered earlier:

Revised syntax of modifiers:

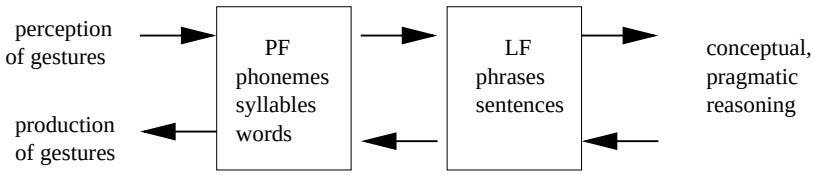
x AP	y NP	↦	xy NP	**AP modifies NP**
x,y VP	z AdvP	↦	x,yz VP	**AdvP modifies VP**
x NP	y PP	↦	xy VP	**PP modifies NP**
x,y VP	z PP	↦	x,yz VP	**PP modifies VP**
x P	y DP	↦	xy PP	**Prep takes DP object**

The trees on page 141, built with the first modifier rules, now look like this:





These derivations are **structure-dependent** in two senses: *first*, they involve recognizing structures (the subject and object DPs, the predicates (VP, T') and so on), and *second*, the rules themselves refer to parts of the structures already built (particular elements of the pairs of strings). It is natural to assume that human language recognition involves computing this structure from the perceived phonetic elements.



Calling the phonetic form PF, and the grammatically-defined structure LF (for "logical form"), various versions of this simple idea about language perception are expressed by some linguists, psychologists, philosophers:

> PF and LF constitute the 'interface' between language and other cognitive systems, yielding direct representations of sound, on the one hand, and meaning on the other as language and other systems interact, including perceptual and production systems, conceptual and pragmatic systems. (Chomsky, 1986, p68)

> The output of the sentence comprehension system…provides a domain for such further transformations as logical and inductive inferences, comparison with information in memory, comparison with information available from other perceptual channels, etc...[These] extra-linguistic transformations are defined directly over the grammatical form of the sentence, roughly, over its syntactic structural description (which, of course, includes a specification of its lexical items). (Fodor et al., 1980)

> …the picture of meaning to be developed here is inspired by Wittgenstein's idea that the meaning of a word is constituted from its use – from the regularities governing our deployment of the sentences in which it appears…understanding a sentence consists, by definition, in nothing over and above understanding its constituents and appreciating how they are combined with one another. Thus the meaning of the sentence does not have to be *worked out* on the basis of what is known about how it is constructed; for that knowledge by itself constitutes the sentence's meaning. If this is so, then compositionality is a trivial consequence of what we mean by "understanding" in connection with complex sentences. (Horwich, 1998, pp3,9)

In these passages, the idea is that reasoning about what has been said begins with the syntactic analyses of the perceived language.

Obviously, the rules for building gestural complexes (syllables, etc) and morpheme complexes (words, phrases, sentences) vary from one language to another, but different languages are similar in many ways too. The linguist Noam Chomsky (1971, pp26-28)proposes that one important similarity is the structure-dependence of the rules, like the ones in our revised syntax. He suggests that this is one of the surprising and distinctive features of human language,

and that it is assumed by human language learners not for simplicity or communicative efficiency but because of some genetically given bias:

> By studying the representation of sound and meaning in natural language, we can obtain some understanding of invariant properties that might reasonably be attributed to the organism itself as its contribution to the task of acquisition of knowledge, the schematism that it applies to data of sense in its effort to organize experience and construct cognitive systems. But some of the most interesting and surprising results concern rather the system of rules that relate sound and meaning in natural language. These rules fall into various categories and exhibit invariant properties that are by no means necessary for a system of thought or communication, a fact that once again has intriguing consequences for the study of human intelligence.
>
> Consider the sentence "The dog in the corner is hungry"...the subject ...is "the dog in the corner"; we form the question by moving the occurrence of "is" that follows it to the front of the sentence. Let us call this operation a "**structure-dependent operation**," meaning by this that the operation considers not only the sequence of elements that constitute the sentence but also their structure; in this case, the fact that the sequence "the dog in the corner" is a phrase, furthermore a noun phrase. [*nb: in these notes, it is a determiner phrase*]. For the case in question, we might also have proposed a "structure independent operation": namely, take the leftmost occurrence of "is" and move it to the front of the sentence. We can easily determine that the correct rule is the structure-dependent operation. Thus if we have the sentence "The dog that is in the corner is hungry," we do not apply the proposed structure-independent operation, forming the question "Is the dog that ____ in the corner is hungry?" Rather, we apply the structure-dependent operation, first locating the noun-phrase subject "the dog that is in the corner," then inverting the occurrence of "is" that follows it, forming: "Is the dog that is in the corner ____ hungry?"
>
> Though the example is trivial, the results is nonetheless surprising, from a certain point of view. Notice that the structure-dependent operation has no advantage from the point of view of communicative efficiency or "simplicity." If we were, let us say, designing a language for formal manipulations by a computer, we would certainly prefer structure-independent operations.
>
> ...Notice further...though children make certain errors in the course of language learning, I am sure that none make the error of forming the question "Is the dog that in the corner is hungry?" despite the slim evidence of experience and the simplicity of the structure-independent rule.

Since these questions bear on the question of which aspects of human language abilities are genetically determined, we will consider these suggestions again later.

5.3 Language structure: Quechua

It is interesting to compare a European and now also American language like English with other languages that are not closely related, to get an idea of how different languages can be. The language of the Incas in South America was Quechua, and many dialects of Quechua continue to be spoken, mainly in Peru, Bolivia, and Ecuador. We introduce the basic sounds and syllable structure (for one dialect of this language), then the morphemes and phrase structure, as we did for English.

5.3.1 Gestural complexes: phonology

			manner	voice	place
1.	[p]	s**p**ot	stop	−voice	labial
2.	[p^h]	**p**op	stop	−voice	aspirated labial
3.	[p']	–	stop	−voice	glottalized labial
4.	[t]	s**t**uck	stop	−voice	alveolar
5.	[t^h]	**t**ick	stop	−voice	aspirated alveolar
6.	[t']	–	stop	−voice	glottalized alveolar
7.	[q]	–	stop	−voice	uvular
8.	[q^h]	–	stop	−voice	aspirated uvular
9.	[q']	–	stop	−voice	glottalized uvular
10.	[k]	s**k**ip	stop	−voice	velar
11.	[k^h]	s**k**ip	stop	−voice	aspirated velar
12.	[k']	s**k**ip	stop	−voice	glottalized velar
13.	[ʧ]	**ch**ip	affricate	−voice	alveopalatal
14.	[ʧ^h]	–	affricate	−voice	aspirated alveopalatal
15.	[ʧ']	–	affricate	−voice	glottalized alveopalatal
16.	[m]	**m**oat	nasal stop	+voice	labial
17.	[n]	**n**ote	nasal stop	+voice	alveolar
18.	[ñ]	–	nasal stop	+voice	palatalized alveolar
19.	[s]	**s**ip	fricative	−voice	alveolar
20.	[ʃ]	**sh**ip	fricative	−voice	alveopalatal
21.	[x]	–	fricative	−voice	velar
18.	[h]	**h**at	fricative	−voice	glottal
22.	[ɾ]	butter	flap	+voice	alveolar
22.	[r]	reef	(central) approximant	+voice	retroflex
21.	[l]	leaf	lateral approximant	+voice	alveolar
21.	[ʎ]	–	lateral approximant	+voice	palatal
24.	[j]	yet	(central) approximant	+voice	palatal
25.	[w]	weird	(central) approximant	+voice	labiovelar

			tongue body height	tongue body backness	lip rounding	tongue root tense (+ATR) or lax (−ATR)
26.	[i]	b**ea**t	high	front	unrounded	+ATR
27.	[u]	b**oo**t	high	back	rounded	+ATR
28.	[o]	r**oa**d	mid	back	rounded	+ATR
29.	[e]	**a**te	mid	front	unrounded	+ATR
30.	[a]	p**o**t	low	back	unrounded	+ATR

This list is similar to the list of English phonemes in some respects, and different in others. In Quechua, there is no difference between /p/ and /b/, but there is a difference between /p/, /p^h/ and /p'/. We show that these are phonemes with minimal pairs like these:

	word	meaning	word	meaning	word	meaning
i.	tanta	collection, combination	t^{h}anta	old, worn out	t'anta	bread
ii.	p'ataʎ	to bite	p^{h}ataʎ	to explode, to blow up	paʎ	he
iii.	ʧ'aʧ'u	treacherous, tricky	ʧ^{h}aʧu	ragged, tattered	ʧu	make
iv.	k'ank'a	rooster	k^{h}anka	slimy, clammy	kasaʎ	to be
v.	q'ata	turbid, muddy	q^{h}ata	mountainside	noqa	I

Syllables allow at most one consonant in onset and coda positions (at least to a good first approximation). So while English allows 3-consonant onsets like [spl] in syllables like [splæʃ] ('splash'), this does not happen in Quechua.

5.3.2 Morpheme complexes: syntax

Compared with English, Quechua word structure is very rich. A whole sentence can be expressed by a word or two. For example (Herrero and Sánchez de Lozada, 1978; Stabler, 1994):[10]

(7) wañu-chi-chi-lla-sa-nku-ña-puni.
die-make-make-DEL-PROG-3PL-DUR-EMP
'they are still just having people killed as always'

Languages that string together long sequences of morphemes into single words, complexes that are pronounced according to word stress rules, are called **agglutinating** or **agglutinative**. Languages that are primarily agglutinating include Quechua, Turkish, Mongolic, Manchutungusic, Finnish, Japanese, Korean, Hungarian, Malayalam, Telugu, Zulu,...These are usually distinguished from **polysynthetic** or **fusional** languages that put together complex words but with extensive sound changes in the morphemes depending on context, as in Mohawk, Mayali, Nahuatl, Southern Tiwa, Chuckchee,... And these languages types are both distinguished from **isolating** or **analytic** languages which tend not put many morphemes together: English, Chinese, Vietnamese, Samoan,....

Quechua suffixes must be in a certain order, as in English:

[10] In the word-for-word translations, IND for indicative, PROG for progressive, FUT for future, NEG for negative, S for singular, PL for plural, DAT for dative, ACC for accusative, LOC for locative, DEL for delimitative, DUR for durative, and EMP for emphatic elements.

(8) nation-al-iz-ation, *nation-iz-al-ation, *nation-ation-iz-al

And as in English, most prefixes and suffixes cannot be repeated, but some can:

(9) a. assemble, dis-assemble, *dis-dis-assemble
b. Darwin, Darwin-ian, *Darwin-ian-ian
c. art, art-ist, *art-ist-ist
d. missile, anti-missile, anti-anti-missile

In (7) just above, we see one of the few Quechua suffixes that can be repeated: *-chi*, meaning 'make'. This suffix can attach to quite a wide range of verbs:

(10) Riku-ni
see-1S
'I see it'

(11) Riku-chi-ni
see-make-1S
'I show it' or 'I make him see it'

(12) Riku-chi-chi-ni
see-make-make-1S
'I have it shown'

The person and number of the verbs is similar, except that there is an inclusive *we* – used when the hearer is included, and an exclusive *we* – used when the hearer is excluded. This distinction gets lost in the translation to English:

noqa kasa-ni	I be-1s	'I am'	noqanchej kasa-nchej	we(incl) be-1p	'we are'
			noqayku kasa-yku	we(excl) be-1p	'we are'
qan kasa-ni	you be-2s	'you are'	qankuna kasa-nkichej	you(pl) be-2p	'you guys are'
pay kasa-n	you be-3s	'he/she/it is'	paykuna kasa-nku	they be-3p	'they are'

English marks "case" only in pronouns (e.g. nominative *she* vs. accusative *her*), but Quechua marks it explicitly: *-ta* for accusative case (object of verb or preposition), *-man* for indirect objects or locatives, and *-wan* for instruments (like English *with*):

(13) Mariya t'anta-ta miku-n
Maria bread-ACC eat-3s
'Maria eats bread'

This is the most common word order – Subject Object Verb – but in a simple sentence like this all 6 orders of these words are perfectly fine and roughly synonymous.

Notice that we can get all 6 orders using simple grammars, if we "split" the verb phrase into 2 parts so that we can reorder the verb and object. – This is the same thing we did to get crossing dependencies in DNA and auxiliary verbs in English. We also add a very simple

rule here to add the 3rd person marker *-n* to the verb, which is the only suffix when tense is simple present. (This was suggested by the question in class.) And we let the case marker *-ta* get added by the VP rule itself:

Morphemes:	(names,bare nouns)		DP:	Juan, Marya, pay, t'anta,…	
	(transitive verb)		Vt:	mikhu,riku,kasu,…	
	(3rd person suffix)		T:	-n	
Simple fragment of Quechua Syntax, with "Movement":					
x Vt	y DP	↦	x,y-ta VP	**Vt takes DP object**	(1)
-n T	y,z VP	↦	y-n,z T'	**add only 3rd person, for present tense**	(2)
x DP	y,z T'	↦	w TP	**where w is x,y,z in <u>any</u> order**	(3)

The last rule allows the subject x to be combined with the verb y and object z in any of the 6 possible orders. The actual grammars used by Quechua speakers is of course much more complex than this one, but it is natural to get the possible reorderings with rules like this that let you "move" each constituent into one of the possible positions.[11]

Quechua has some other very general differences from English. Instead of prepositions, Quechua has postpositions – these are like prepositions but they <u>follow</u> their objects instead of preceding them, like *wan*, 'with', in the following example:

(14) mikhu-n t'anta-ta mantekilla-wan ima
eat-3s bread butter-with too

And another strange thing: adverbs seem to get marked for case:

(15) Allin-ta llank'a-nki
good/well-ACC work-2
'you work well'

(16) Allin-ta riku-ni
good/well-ACC see-1
'I see well' or 'I see the good one'

[11] There are many studies of how to get the appropriate orders of constituents in languages that allow much more variation than English. Rambow (1994) shows that certain "scrambling" constructions in German cannot be appropriately defined with context free (or even "tree adjoining") grammars, and advocates a kind of grammar that is similar in power to the ones we are using here.

5.4 Language typology

Darwin predicted an analogy between the phylogeny of animals and the phylogeny of languages in *Origin of Species*

> On the view which I hold, the natural system is genealogical in its arrangement, like a pedigree; but the degrees of modification which the different groups have undergone, have to be expressed by ranking them under different so-called genera, sub-families, families, sections, orders, and classes.
>
> It may be worth while to illustrate this view of classification, by taking the case of languages. If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, had to be included, such an arrangement would, I think, be the only possible one. Yet it might be that some very ancient language had altered little, and had given rise to few new languages, whilst others (owing to the spreading and subsequent isolation and states of civilization of the several races, descended from a common race) had altered much, and had given rise to many new languages and dialects. The various degrees of difference in the languages from the same stock, would have to be expressed by groups subordinate to groups; but the proper or even only possible arrangement would still be genealogical; and this would be strictly natural, as it would connect together all languages, extinct and modern, by the closest affinities, and would give the filiation and origin of each tongue. (Darwin, 1859, §13)

He even observes how "Rudimentary organs may be compared with the letters in a word, still retained in the spelling, but become useless in the pronunciation, but which serve as a clue in seeking for its derivation." (§13) Let's review some basics first.

How many languages are there? It is hard to tell when to call two dialects different languages, as it is hard to tell when two varieties of organism are different species. With species, the usual rough criterion is the possibility for mating, and with languages, the usual rough criterion is mutual intelligibility. With this rough criterion, most linguists estimate that there are more than 4000 languages. Some linguists think there are more than 6000.

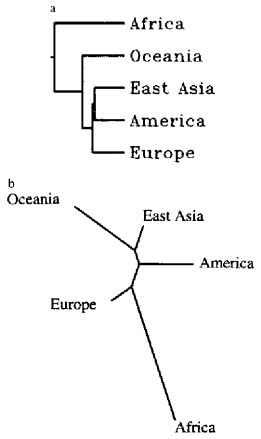
How many phonemes does each language have? The two cases we have looked at are typical: English has approximately 38, and Quechua has approximately 30. Hawaiian has just 13 phonemes, and the Austronesian language Rotokas has only 11 (Clark, 1990), while at the other extreme, the African Khoisan language Ju'Hoansi is claimed to have 89 consonants, 34 vowels and 7 phonemic tone patterns (Miller-Ockhuizen, 2001).

How many morphemes does each language have, each language of each particular individual? How should we tell whether someone really "knows" a word? This is not clear, but the usual rough estimate is that children master some 10,000 morphemes by the age of 6 (as was mentioned on page 124). How many more are learned depends on whether the individual becomes literate and reads, and many other factors.

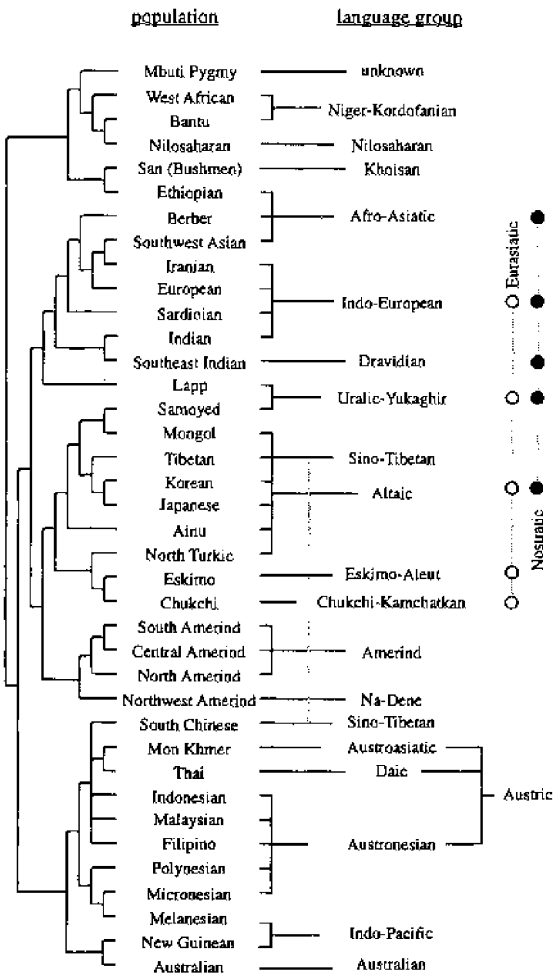
Languages are changing constantly. Your English is different from your parents' English, and furthermore, there are population-wide trends. Consider these examples in the recent history of English, from (Joseph, 2000):

1. Nū wē sculon herian heofon-rīces Weard (Caedmon,ca.660)
   now we shall to-praise heaven's guardian
2. Whan that Aprille with its shoures soote… (Chaucer, ca.1400)
   when April with its sweet showers…
3. Tush, never tell me! I take it much unkindly that thou, Iago, who hast had my purse as if the strings were thine, shouldst know of this. (Shakespeare, 1604)

Here we see not only changes in pronunciation, but also words that have disappeared (*sculon, herian, Weard,…*) and changes in clause structure (you can no longer say *I came <u>when that</u> April with its sweet showers wooed me*). Does anything remain the same? We consider some properties that remain in a moment, but first, notice that with so many things changing, it is no surprise that it has been difficult to detect in languages remnants of their remote history (contrasting with DNA, in which certain parts have remained remarkably stable all the way back to bacteria). It is easy to see that there are clusters of similar, apparently related languages, which most people are aware of: Romance, Germanic, and even larger groupings like IndoEuropean. Furthermore, as Darwin expected, these clusters correlate with genetic similarity among their speakers (Cavalli-Sforza, 1997) Based on extensive surveys of genetic similarities, Cavalli-Sforza plots the relations among the major continental groups like this:



Recent studies of DNA and models of genetic change support the conclusion that the common ancestor of all living humans (at the left of the higher tree, and in the middle of the lower one) lived in Africa 100,000-200,000 years ago. (Jorde, Bamshad, and Rogers, 1998). It is not possible to attain such depth of comparison among languages, but still there is a correlation (no surprise!):



Obviously, this does not imply that our genetic differences code for language differences! We can refute this simple idea with the familiar observation that a child of any ethnicity born into any human linguistic community can learn the language of the community perfectly well. What the correlation shows is that as ethnic groups split off and become relatively isolated for a while, not only do they develop distinctive genetic traits, but also distinctive languages and cultures.

5.5 Universals: first ideas

Some properties that we find in all human languages are easy to list now:

1. Every human language is infinite (has infinitely many declarative sentences – no longest one)
2. Every spoken human language is interpreted compositionally, in the sense that the meanings of many new utterances are calculated from the meanings of their parts and their manner of combination (Frege's proposal)
3. Every spoken human language distinguishes vowels and consonants,
   and among the consonants, distinguishes stops/fricatives/affricates
   from the more sonorant glides/liquids/nasals.

   In all signed human languages, there are similar basic gestures, and fundamental distinctions between handshapes, locations and movements. (Sandler and Lillo-Martin, 2001).
4. Every human language has transitive and intransitive sentences, but the major constituents (subject, object, verb) occur in different orders in neutral sentences:

   SOV (Quechua,Turkish,Japanese,Navajo,Burmese,Somali,Warlpiri,American Sign Language)

   SVO (English,Czech,Mandarin,Thai,Vietnamese,Indonesian)

   VSO (Welsh,Irish,Tahitian,Chinook,Squamish)

   very rare:

   VOS (Malagasy,Tagalog,Tongan)

   OVS (Hixkaryana)

   OSV ?

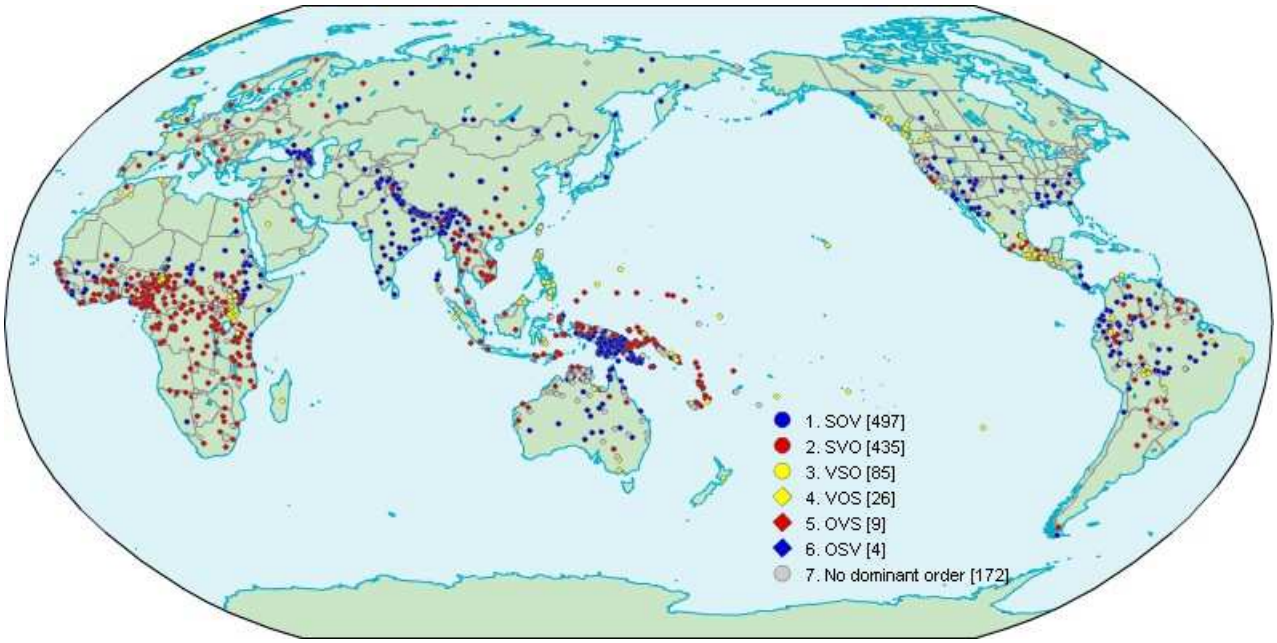
Property 4 shows that there is rather limited variation in the order in which elements are **selected** and rearranged by **movements**. The linguist Greenberzg (1963) included these strong tendencies in his catalog of 45 more specific universals:

G1. In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object. Subjects tend to be on the left.

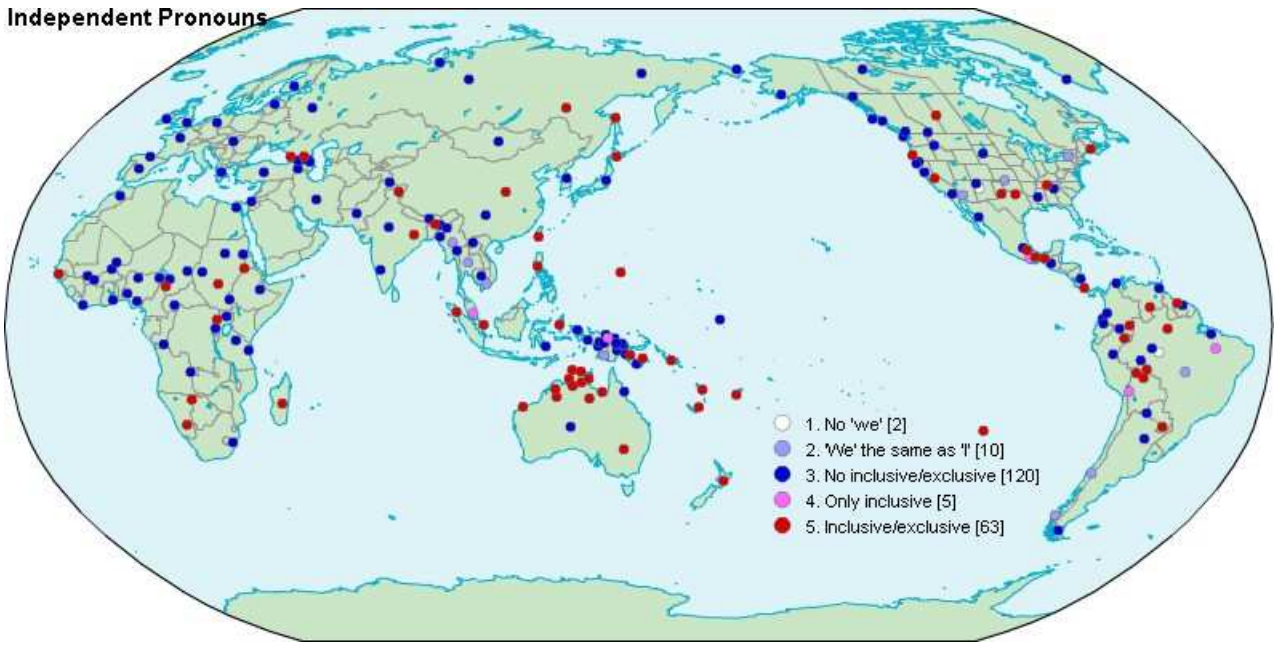
G3. With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional. (e.g. Quechua) So if a verb selects its object on the left, prepositions almost always do too.

These are confirmed and elaborated by recent studies (Hawkins, 1994, and others)

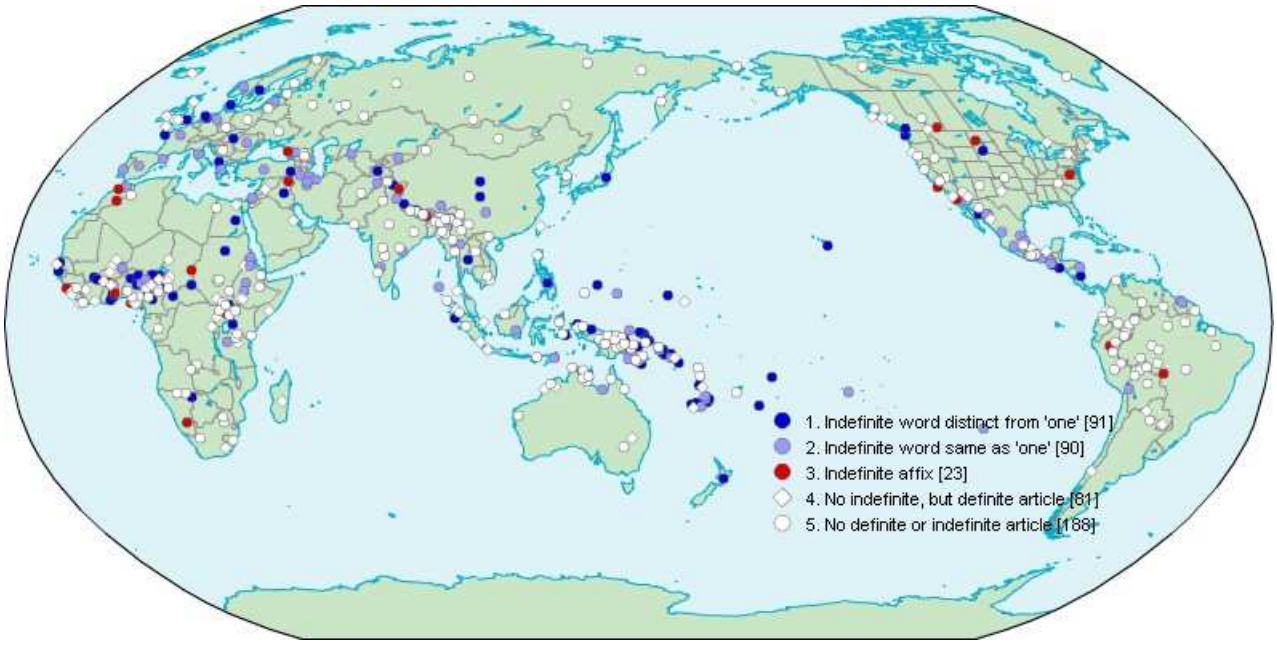
The order of major constituents (SVO,SOV,...) is one basic property that varies in the world's languages, and recent studies show that the geographical distribution of this variation is far from random. The following maps are from Haspelmath et al. (2005):



Another variation that we saw in the English-Quechua contrast is the marking of inclusive/exclusive forms of pronouns:



And unlike English, Quechua has no indefinite pronouns:



In the first and third of these maps, the status of English stands out as exceptional, with Quechua having the more usual properties. And in all these cases, we see nonrandom geographical distributions.

Exercises

Note:

- Use the phonetic notation and the rules introduced in this class. (Dictionaries and other classes may have used slightly different notations, but part of the exercise here is to use exactly the notation and rules we have introduced)

This week we introduced a kind of structure into the rules for morpheme complexes that we had already seen in DNA duplication, but here we saw Chomsky proposes it for English auxiliaries, and it is useful for reordering constituents in Quechua too. We then made some first observations about language differences, language typology, and language universals.

1. Consider a human language with 32 phonemes. Since $2^5 = 32$, any one of 32 things can be specified by 5 bits. Explain why it is incorrect to assume that a 10-phoneme utterance in this language carries 50 bits of information. (Hint: it is exactly analogous to the reason, discussed in lecture 4, that a 10-nucleotide sequence of DNA does not, in general, specify 20 bits of information.)

2. Using the more sophisticated rules for English morpheme complexes introduced this week, present a derivation tree for the sentence (break it into morphemes, and show the derivation tree, as done in class and notes):

   John has been teaching the student

3. Using the more sophisticated rules for English morpheme complexes introduced this week, present a derivation tree for the sentence (break it into morphemes, and show the derivation tree, as done in class and notes):

   The penguin praised Bill rarely

4. Show the structure of all the syllables in the following Quechua sentence. (It means: "They made me drink it." The standard spelling shown here is phonemic except that "ch" corresponds to the phoneme [ʧ], and "j" corresponds to the phoneme [x].)

   Ujyachiwarqanku.

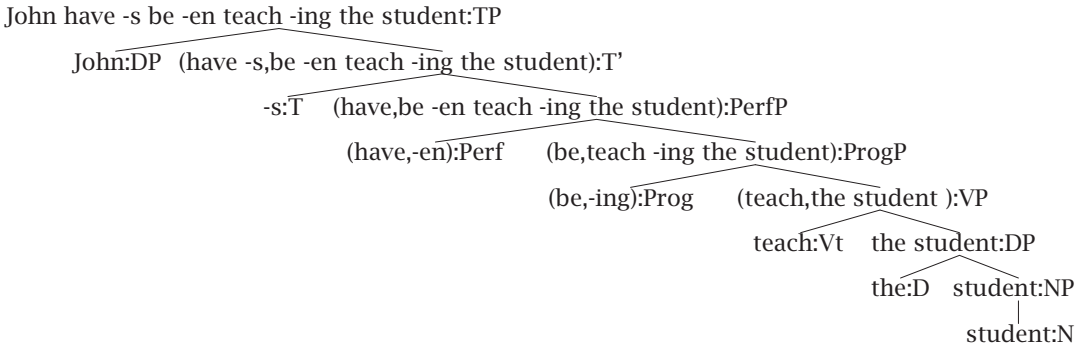
5. Using the rules for the "Simple fragment of Quechua Syntax, with 'Movement'", present a derivation tree for the following Quechua sentence, which means *he sees Maria.* (Break it into morphemes, and show the derivation tree, as done in class and notes):

   Pay Marya-ta riku-n

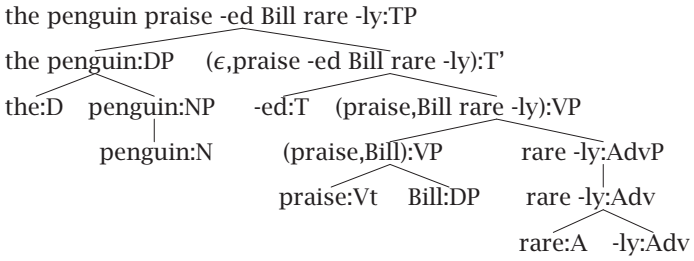
Solutions

1. If, in each position, every phoneme were equally likely, then each phoneme would be one choice out of 32. By Shannon's definition from page 112, that's 5 bits of information. But no human language allows you to put any phoneme in any position. Phoneme sequences are very unlike random sequences! For example, in English, if you start with [p], the next sound cannot be just any English phoneme. For example, the next phoneme cannot be [k] or [p] or [t] or [z] or…. So the choices at each point are much less than all the phonemes, and so the information conveyed by a sequence of 10 phonemes, in a human-like language with 32 possible phonemes, is much less than 50 bits.

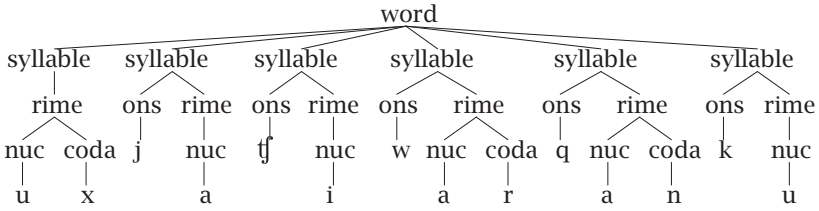
2.



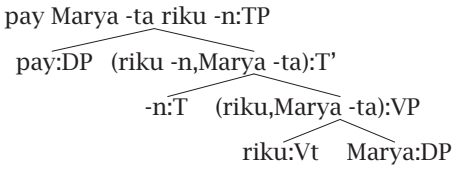
3.



4. We said Quechua syllables can have at most 1 consonant in onset and in coda, and that languages generally prefer to avoid codas, so the structure must be this:



5.



Lecture 6

Origins of human linguistic ability: Selection, exaptation, self-organization

Human languages vary; they exhibit "phenotypic plasticity." But we have also seen that they have many properties in common. In this chapter we review and elaborate some proposals about the origins of human language abilities, the abilities we have to learn, produce and recognize linguistic structures.

In this chapter we will consider some of the origins of these abilities: not the origins of any particular language, but of the human "faculty of language," the ability to learn and use any particular human language.

6.1 Innateness

Evolution is a source of order in human language only if it exhibits the basic properties identified by Darwin (discussed in §0): heritability, variation, and selection. Are (at least certain aspects of) human language ability inherited? The proposal that they are is sometimes called the "innateness" hypothesis: the human language faculty is, at least in part, genetically determined. The fact that they are is suggested by their regular acquisition by children (from limited and widely various evidence), and it is also suggested by the neural and vocal tract specializations for language. We have further direct evidence from heritability and molecular genetic studies.

One way to study genetic factors in language ability is to compare fraternal and identical twins raised in different environments. Since identical twins have the same genetic material while fraternal twins share only about half their genetic material, genetically determined traits should correlate more highly in identical twins. (This provides one of the more sophisticated strategies for answering Exercise 3 on page 39.) There have been a number of studies of development dyslexia – an inability in learning to read despite a normal environment – and specific language impairment (SLI) – a language deficit that is not accompanied by general cognitive or

emotional problems. A number of these studies decisively establish the heritability of both dyslexia and SLI. See, for example, Stromswold (1998), for a review.

A family called "KE" of 30 people over 4 generations was discovered, with a language disorder that affects approximately half the family. The affected family members have difficulty controlling their mouth and tongue, but they also have problems recognizing phonemes and phrases. Careful comparison of the genomes of the normal and impaired family members, together with the discovery of an unrelated person with a similar impairment, has led to the identification of one of the genes apparently involved in the development of language abilities (Hurst et al., 1990; Gopnik, 1990; Vargha-Khadem et al., 1995; Lai et al., 2001; Kaminen et al., 2003). The gene, FOXP2, is encoded in a span of about 270,000 nucleotides in chromosome 7. Parts of this span encode an 84 amino acid sequence, shown in part in the following figure from Lai et al. (2001) (the one-letter codes for amino acids were given in §0 on page 12 of these notes):

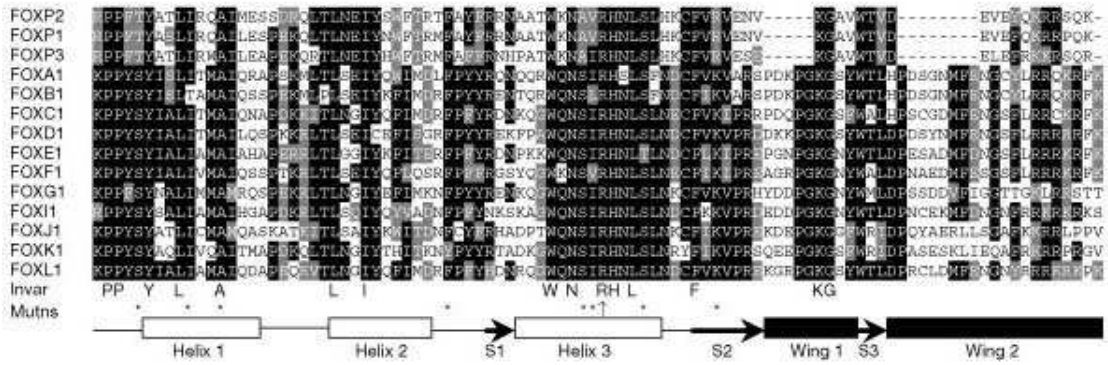


Figure 4 Forkhead domains of the three known FOXP proteins aligned with representative proteins from several branches of the FOX family. All sequences are from *Homo sapiens*. Residues that are invariant in this selection of forkhead proteins are given beneath the alignment. Asterisks show sites of the substitution mutations in FOXC1, FOXE1 and FOXP3 that have been previously implicated in human disease states. The upwards arrow indicates the site of the R553H substitution identified in FOXP2 in affected members of the KE pedigree. The proposed structure of the forkhead domain as established by X-ray crystallography is shown, containing three α-helices, three β-strands (S103) and two 'wings'.

In the impaired individuals, a single G(uanine) nucleotide is replaced by A(denine), resulting in a change from the amino acid H (Histidine) to R (arginine).

These discoveries have been discussed and debated in the popular press. For example, Chomsky (1991) points out that this discovery is entirely compatible with his claims that some grammatical universals are innately specified:

> Philip Lieberman writes [Letters, NYR, October 10] that he objects only to "biologically implausible formulations" of Universal Grammar "that do not take account of genetic variability." In response, Lord Zuckerman observes correctly that it is a "truism" that a "genetically based 'universal grammar' " will be subject to variability. It remains only to add that the truism has always been regarded as exactly that.
>
> Lieberman states that "until the past year, virtually all theoretical linguists working in the Chomskian tradition claimed that Universal Grammar was identical in all humans," thus denying the truism. By "the past year," he apparently has in mind Myrna Gopnick's results, which he cites, on syntactic deficits. The claim that Lieberman attributes to "virtually all theoretical linguists?" is new to me; to my knowledge, Gop-

> nick's results, far from causing deep consternation, were welcomed as interesting evidence for what had been assumed. Perhaps Lieberman has been misled by the standard assumption that for some task at hand – say, the study of some aspect of language structure, use, or acquisition – we can safely abstract from possible variation. To quote almost at random, "invariability across the species" is a "simplifying assumption" that "seems to provide a close approximation to the facts" and is, "so far as we know, fair enough" for particular inquiries (Chomsky, 1975). Note that one who takes the trouble to understand what is always assumed might argue that this approximation is not good enough, and that problems might be solved by moving beyond it. A serious proposal to that effect would, again, be welcomed, another truism.

And Gopnik (1992) emphasizes that what has been discovered is not the one and only language gene, but rather one of the genes possibly involved in the development of language abilities:

> Zuckerman, responding to Chomsky, raises as an ongoing question an issue that I believe can be clearly settled by looking at inherited language impairment over several generations. In *The New York Review* he claims that whether "man's syntactical abilities [are] due to one set of interacting genes or more than one [is] anyone's guess." While this question may have been "anyone's guess" in the past, there is now converging evidence from several studies that provide a clear answer: though certain cases of developmental language impairment are associated with a single autosomally dominant gene, these impairments affect only part of language – the ability to construct general agreement rules for such grammatical features as tense and singular/plural – and leave all other aspects of language, such as word order and the acquisition of lexical items, unaffected. These facts answer Zuckerman's question: Language must be the result of several sets of interacting genes that code for different aspects of language rather than a single set of interacting genes. In fact it is misleading to think of "language" or "grammar" as unitary phenomena. Inherited language impairment shows that different parts of grammar are controlled by different underlying genes.

6.2 An argument for emergence by selection

Since there is clear evidence for genetic control of at least some aspects of human linguistic abilities, it makes sense to ask whether these abilities emerged by natural selection. In a series of publications, Pinker has argued that it did (Pinker and Bloom, 1990; Pinker, 2000; Pinker, 2001). The main argument is summarized in the following passages

> Evolutionary theory offers clear criteria for when a trait should be attributed to natural selection: complex design for some function, and the absence of alternative processes capable of explaining such complexity. Human language meets this criterion: grammar is a complex mechanism tailored to the transmission of propositional structures through a serial interface...(Pinker and Bloom, 1990)

By "propositional structure," they presumable mean a structure that can express a proposition, something that can be true or false. Is Pinker right about what 'evolutionary theory' tells us? Can we just look at a trait, and if it shows a 'complex design', conclude on the basis of 'evolutionary theory' that it should be attributed to natural selection?

Of course not. Two prominent biologists have responded to this proposal: Lewontin – known especially for his work in population genetics, the "Lewontin-Kojima model" etc – and paleontologist Stephen Jay Gould. (See also Orr 1995, Berwick 1997.)

Gould, Lewontin and others have pointed out that when we consider traits whose history is preserved in the fossil record – bone structure, etc. – we find "recruitment" or "exaptation" more often than simple selection. That is, many traits of organisms emerge because other traits have been selected. For example, Gould and Vrba (1982) point out that feathers may have emerged because they were good for catching insects, and wings may have emerged because they are good for casting a shadow on water so that prey in the water can be seen. The first use of wings and feathers for flight may have come later, and then, of course, it proved very adaptive. But feathers were probably not initially selected because of their potential contribution to their flying ability. They coin the term "exaptation" for features not specifically selected for or features previously designed for another function, which have been coopted for their current use. They often call such features "spandrels," which is the name of the space between an arch and the corner of a rectangular structure supported by arches. Many churches and other buildings are designed with arch supports, and the spandrels result; they are not designed specifically to have spandrels. Lewontin (1998) says,

> The phenomenon of recruitment in the origin of new functions is widespread in evolution. Birds and bats recruited bones from the front limbs to make wings…The three bones that form the inner ear of mammals were recruited from the skull and jaw suspension of their reptilian ancestors. The panda's thumb is really a wrist bone recruited for stripping leaves from bamboo.

But the idea that language is a spandrel, that it might have emerged by exaptation, is criticized by Pinker and others:

> The key point that blunts the Gould and Lewontin critique of adaptationism is that natural selection is the only scientific explanation of adaptive complexity. "Adaptive complexity" describes any system composed of many interacting parts where the details of the parts' structure and arrangement suggest design to fulfill some function. The vertebrate eye is the classic example…It is absurdly improbable that some general law of growth and form could give rise to a functioning vertebrate eye as a by-product of some other trend such as an increase in size of some other part.

Similar points have been made by (Williams, 1966; Dawkins, 1986). Lewontin responds again:

> Unfortunately, we are not told…how to measure the complexity of linguistic ability as compared with, say, the shape of our faces nor what (unmeasured) degree of complexity is required for for natural selection to be the only explanation.

6.3 Another argument for emergence by selection

In class discussion, we mentioned briefly that now that we can sequence particular genes and identify variants, statistical studies of the kinds of genetic variation occuring near the variations can, at least in principle, provide clues about whether some variant has been selected or not. If FOXP2 is a language gene, perhaps we can see evidence of this kind that it has been selected.

6.3.1 How is FOXP2 expressed? What abilities depend on it?

As mentioned at the beginning of the chapter, there is a gene FOXP2 on the 7th human chromosome. The genes is approximately 270,000 nucleotides, coding a protein with 715 amino acids, and a single nucleotide variation alters one of the amino acids that affects the protein (Lai et al., 2001). What are the phenotypic consequences of this change, exactly? The question of whether the consequence actually affects any distinctively linguistic ability has been controversial. A recent study by Watkins, Dronkers, and Vargha-Khadem (2002) tries to resolve some of the controversy by administering a wide range of tests to many of the family members, testing linguistic abilities and cognitive abilities and coordination. They did find linguistic deficits: the affected family members had an impaired ability to repeat words and to properly produce past tenses, but they also had trouble repeating non-words, and they had difficulty with both regular and irregular past tenses. Furthermore, the impaired family members showed deficits on "almost every test." Based on a statistical analysis of the test results, the study concludes:

> We suggest that, in the affected family members, the verbal and non-verbal deficits arise from a common impairment in the ability to sequence movement or in procedural learning. Alternatively, the articulation deficit, which itself might give rise to a host of other language deficits, is separate from a more general verbal and non-verbal developmental delay.

6.3.2 Another mutation of FOXP2

Another mutation has been found that truncates a different FOXP2 protein, R328X. This one also has linguistic effects, as described in these passages from a recent report:

> Our screening of probands also identified three novel exonic allelic variants in the coding region, each of which is predicted to yield a change in FOXP2 protein sequence…Crucially, one of these coding changes was a heterozygous CrT transition in exon 7, yielding a stop codon at position 328 of the FOXP2 protein (R328X)…
>
> The R328X mutation is highly likely to have functional significance, since it leads to dramatic truncation of the predicted product, yielding a FOXP2 protein lacking critical functional domains…
>
> The development of the children carrying the R328X mutation was assessed using the Griffiths (1970) Mental Development Scales…

> Assessment of the proband when he was 4 years old indicated developmental delays in the domains of speech and language, and social skills. He communicated mainly using single words and was unable to repeat multisyllabic words. Eye-hand coordination was satisfactory, but he had difficulty with activities in the Practical Reasoning domain…During informal assessment of articulation, he had difficulty in producing consonants at the beginning of words and became frustrated and significantly less intelligible during word repetition.
>
> His younger sister has a history of motor and oropharyngeal dyspraxia, otitis media, and oesophageal reflux. On assessment using the Griffith Scales, at age 1 year 8 mo, she showed her poorest performance in the Hearing and Speech domain. She did not speak any words and could not identify objects, and her vocalization was poor. However, she was interested in puzzle-type toys and was able to put different shapes into form boards; her general motor skills at this age appeared normal…
>
> The mother, who also carried the R328X mutation, reported a history of speech delay in childhood. At present, she has severe problems with communication. She volunteered to bring a relative to the consultations, because she could not understand the nuances of what was said and was afraid of misinterpretation. She had poor speech clarity and very simple grammatical constructions. Her speech had less varied cadence than most people's, but her vocabulary was satisfactory. Her receptive difficulties were compounded by performance anxiety. (MacDermot et al., 2005)

6.3.3 Statistical evidence of selection of the FOXP2 gene?

The statistical tests for evidence of selection look for differences between the kinds of variation found around the selected site and the kinds of variation that would be expected if the variation at the site was completely neutral. This is difficult, because the rate of neutral variation is not constant: it varies with species, with sex, with demographic and other factors that are not well understood Kreitman (2002) reviews 12 of these statistical tests for selection, and is very cautious about drawing conclusions, especially in populations that have been expanding, since the "signatures of positive selection and expanding population are similar." Furthermore, he says

> With the availability of so many ad hoc statistical tests to detect selection, it is not unlikely that one or another of the tests will support a departure from neutrality…In practice, researchers do not report all of the tests they have carried out on the data, but rather focus on the statistically significant ones.

One widely cited study does provide both positive and negative results. Hamblin, Thompson, and Rienzo (2002) studied vivax malaria resistant blood group alleles. The mortality rate of this kind of malaria is a little less than 5%, but it is not a surprise that the resistant blood group is more frequent in African populations than in Italian or Chinese populations. Still, only some of the popular statistical tests for selection produced positive results in this case.

Enard et al. (2002) applied some of these same tests to the FOXP2 data, and carefully compared the sequence to homologous sequences in other species. They found that the human FOXP2 differs from the mouse at 3 points, and differs from chimpanzees, gorillas, and rhesus monkeys at 2 points. Two of the tests for selection did show significant departures for neutrality (the tests H, and Tajima's D), but they note that "population growth can also lead to negative D values throughout the genome. However, the value of D at FOXP2 is unusually low compared with other loci." Considering all the evidence, though, they conclude

> human FOXP2 contains changes in amino-acid coding and a pattern of nucleotide polymorphism which strongly suggest that this gene has been the target of selection during recent evolution.

Given the state of our understanding of these statistics and of the neutral, null model, the results should be interpreted cautiously (cf. e.g. Sabeti et al. 2006). But the prospects for this kind of study in the future look exciting.

6.4 An argument for emergence by exaptation

Hauser, Chomsky, and Fitch (2002) suggest that if a martian came to earth and meticulously observed Earth's living creatures,

> ...it might note that the faculty mediating human communication appears remarkably different from that of other living creatures; it might further note that the human faculty of language appears to be organized like the genetic code – hierarchical, generative, recursive, and virtually limitless with respect to its scope of expression. With these pieces in hand, this martian might begin to wonder how the genetic code changed in such a way as to generate a vast number of mutually incomprehensible communication systems across species, while maintaining clarity of communication within a given species.

To focus the question more precisely, Hauser, Chomsky and Fitch tentatively distinguish two different things whose origins we might ask about:

FLN: the faculty of language in the narrow sense: the abstract computational system whose key component "generates internal representations and maps them into the sensory-motor interface by the phonological system, and into the conceptual-intentional interface by the (formal) semantic system"

FLB: the faculty of language in the broad sense: the broader system that encompasses the FLN together with aspects of the associated sensory-motor and conceptual-intentional systems: categorical perception, concept formation, the programming and coordination of motor output, etc.

Then they consider 3 different hypotheses, adopting the 3rd:

~~*Hypothesis 1:*~~ *FLB (including FLN) is strictly homologous to nonhuman animal communication.* "FLB is composed of the same functional components that underlie communication in other species."

~~*Hypothesis 2:*~~ *FLB is a highly complex adaptation for language.* Like the vertebrate eye, the "FLB, as a whole, is highly complex, serves the function of communication with admirable effectiveness, and has an ineliminable genetic component. Because natural selection is the only known biological mechanism capable of generating such functional complexes, proponents of this view conclude that natural selection has played a role in shaping many aspects of FLB, including FLN, and, further, that many of these are without parallel in nonhuman animals."

Hypothesis 3: FLN emerged recently, and is unique to our species, while other parts of FLB are primarily based on mechanisms shared with nonhuman animals.

Hauser, Chomsky, and Fitch (2002) argue for hypothesis 3. Since this hypothesis gives quite different stories about FLN and FLB, the question of what FLN includes becomes paramount. Here they adopt a rather surprising view:

Hypothesis 3a: "FLN comprises only the core computational mechanisms of recursion" "as they appear in narrow syntax and the mappings to the interfaces" and furthermore, "we see little reason to believe …that FLN can be anatomized into many independent but interacting traits."

Hypothesis 3b: "certain specific aspects of human language" like FLN may be spandrels, "by-products of preexisting constraints rather than end products of a history of natural selection." This becomes a reasonable position once 3a is adopted, since then the FLN is quite simple, not a complex of independent, interacting parts like the eyes of mammals, and so the argument from design (hypothesis 2), at least for FLN, is "nullified."

These hypotheses make a very restricted claim, a claim about just one aspect of language: the FLN. Hauser, Chomsky and Fitch are not so clear about what the FLN includes, wanting to avoid commitments that are inessential to their argument, but they say (p1571), "All approaches agree that a core property of FLN is recursion, attributed to narrow syntax in the conception just outlined. FLN takes a finite set of elements and yields a potentially infinite array of discrete expressions…At a minimum, then FLN includes the capacity for recursion." Furthermore, they suggest:

Hypothesis 3c: "…the core recursive aspect of FLN currently appears to lack any analog in animal communication and possibly other domains as well."

Although this remark emphasizes the uniqueness of recursion, they say that investigations of this hypothesis should consider domains like number, social relationships, and navigation.

6.5 The ability to produce, recognize, and represent recursive structure

Recursion has been mentioned many times in these notes, but since it is central in the hypotheses of Hauser, Chomsky and Fitch, let's now look at it again. It is common to say that a definition of a notion is recursive if it it uses the notion itself (as in the definition of Fibonacci numbers on page 14), and that a structure is recursive if it has complex parts that can contain other complexes of the same kind (as in the sentences of English, mentioned on page 9, or in the noun compounds of English, mentioned on page 135). Recursion (in these senses) is everywhere!

The numbers can be defined recursively. For example, we can generate the numbers with rules that say 0 is a number, and that the result of adding 1 to any number is a number. We could write this with rules like we used before:

$$\textbf{Basic element:} \quad \begin{array}{c} 0 \\ \text{Number} \end{array}$$

$$\textbf{Generative rule:} \quad \begin{array}{c} x \\ \text{Number} \end{array} \quad \mapsto \quad \begin{array}{c} x+1 \\ \text{Number} \end{array}$$

The generative rule in this definition of Number uses the definition of Number, but recursion like this is very common. It does not distinguish speaking a sentence from doing arithmetic, eating a carrot or taking a walk; crudely, we could define the structure of eating a carrot or any other meal this way:

$$\textbf{Basic element:} \quad \begin{array}{c} \text{first bite} \\ \text{Meal} \end{array}$$

$$\textbf{Generative rule:} \quad \begin{array}{c} x \\ \text{Meal} \end{array} \quad \mapsto \quad \begin{array}{c} x+\text{another bite} \\ \text{Meal} \end{array}$$

Of course, we don't have to define eating activities this way, but for many activities that extend previous results, it is very natural to do so. In computer science, recursion is used for very many things.[1] Let's consider animal cognition in a little more detail to see if the theories avoid recursion. We do not need to look at numerosity in particular, but since that is what Hauser, Chomsky and Fitch suggest, let's try it first.

Recursion in animal cognition: numbers and representations of numbers

A number of studies of animal conceptions of numerosity have revealed more than might have been expected initially (Gallistel, Gelman, and Cordes, 2003). For example, in a task where rats

[1] In computer science, the use of recursion is distinguished from the conditional iteration ("while-loops," etc), but they are expressively equivalent: any function you can compute with one can be computed with the other.

need to push a bar some particular number of times before a food pellet appears in an alcove (where the food in the alcove is not visible from the bar), it was found that they can count fairly reliably to 20 and higher.

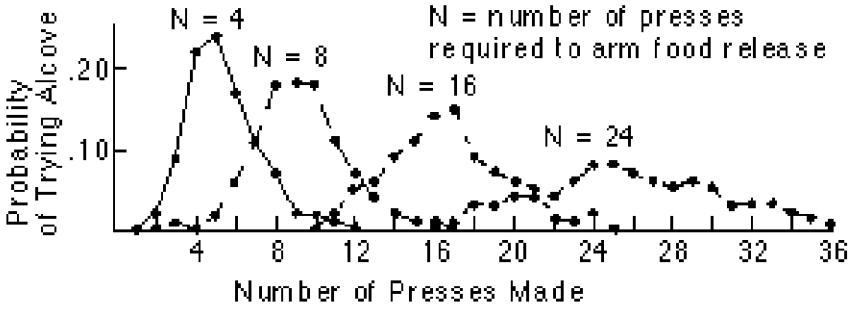


Figure 2. *The probability of breaking off to try the feeding alcove as a function of the number of presses made on the arming lever and the number required to arm the food-release beam at the entrance to the feeding alcove. Subjects were rats. Redrawn from* (Platt & Johnson, 1971) *by permission of the authors and publishers.*

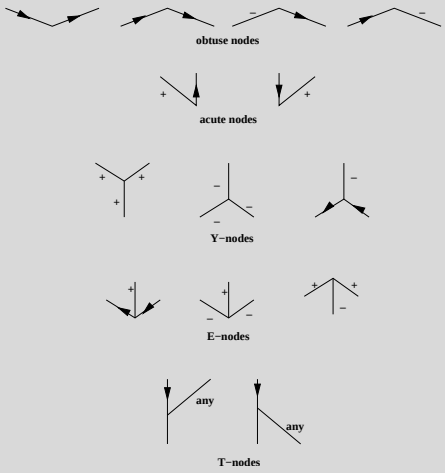
Based on this and many other studies, (Gallistel, Gelman, and Cordes, 2003) suggest that "a system for arithmetic reasoning with real numbers evolved before language evolved," but that the question of whether there is non-verbal reasoning with discrete numbers, like the integers, is more difficult to assess in both human and non-human animals. If animals have a rich representation of real numbers, why don't we find more animal communication about such things, especially in the baboons and chimpanzees which are genetically similar to us? Hauser, Chomsky, and Fitch (2002, p1575) observe: "A wide variety of studies indicate that nonhuman mammals and birds have rich conceptual representations. Surprisingly, however, there is a mismatch between the conceptual capacities of animals and the communicative content of their vocal and visual signals."

Recursion in object recognition

A better case for recursion in animal cognition can, perhaps, be made in perceptual domains. Many animals can recognize certain kinds of visually presented objects and relations. Can they recognize objects inside of, or in front of other objects? Yes. This looks like a recursive ability. How people and other animals do this is not well understood, but it is hard to imagine an account of the capability which would not be recursive.

Simple recursive methods have been found that can recognize objects whose edges are completely inside another's as well as objects that are in front of part of another edge, as in the drawing on the left, in contrast to the "nonsense" drawing on the right.

The problem of how to recognize solid objects from drawings like this has been well studied (Clowes, 1971; Huffman, 1971; Waltz, 1975). Object recognition is especially simple in examples like the ones above where every surface is normal to the X, Y or Z axis (Kirousis and Papadimitriou, 1988). Each edge can be classified as convex (+), concave (-), or a contour (►) that just marks the edge of an object against the background. Convex edges (like ab in the figure above on the left) are coming toward the viewer, concave edges are going away from the viewer and contours (like ac in the figure above) mark the edge of an object. With this classification of edges, there are only 14 possible kinds of edge intersections:



The object recognition consists in finding a categorization of each edge such that every intersection is one of these 14 kinds and each surface has a consistent orientation – something which is easy for the cubes in the figure above on the left, but impossible for the figure on the right. The computations required for this kind of object recognition are intriguingly similar to those required to recognize sentences, with "nonsense" objects failing to have an analysis just like "nonsense" sentences. Object recognition is naturally implemented with matrix multiplication (Kirousis, 1990) and given the similarity in the tasks, it is perhaps not surprising that the fastest known methods for recognizing languages with structured expressions (as in our RNA, English and Quechua grammars) are also matrix multiplications (Nakanishi, Takada, and Seki, 1997) – typically carried out with recursive algorithms (Cormen, Leiserson, and Rivest, 1991, §31.2). In both domains, the algorithms are looking for finitely many different kinds of elaborations of finitely many objects.

Critical summary

One of the points made by Hauser, Chomsky and Fitch is uncontroversial: many of the cognitive and perceptual abilities that are being exercised when we use our language (memory, acoustic perception, coordination,...) are things that could have been selected for other, non-linguistic purposes. So the question is: once we have some sophistication in all these other abilities, what needs to be added in order to obtain human-like linguistic abilities? Hauser, Chomsky and Fitch suggest: recursion (or something just slightly more than this). But there are three

related kinds of objections to this idea. *First,* recursion too is already implicated in perception and other non-linguistic faculties. We have recursion in our notions of number, and also in our conception of objects in space. Do these provide a source for the kind of recursion found in language? The discussion so far suggests at least that recursion itself is not a distinction of human languages. We plausibly find it in many other cognitive domains. It is true that we do not see other animals communicating about recursive structures (like numbers of things), but this does not indicate that they do not compute recursively in other tasks. *Second,* maybe the essential feature of human language is not just its recursion, but something more specialized to language. A *third* objection is that recursion is so basic, it may be a mistake to think it is innately determined at all; maybe it is a property that is introduced (rather easily, for humans) into particular languages and then transmitted from generation to generation by learning. (This last idea is discussed further in the next lecture – see exercise 1 on page 201.)

The hypothesis that human language is not distinguished just by its recursiveness, but by the fact that it is recursively defined over structured representations (of the sort in our rules for RNA sequences, English and Quechua) is a more interesting claim, but even this looks like it is unlikely to be distinctively linguistic.

6.6 Structure-dependence and language complexity

While Hauser, Chomsky and Fitch emphasize recursion in human languages, earlier studies questioned the plausibility of evolutionary accounts of other aspects of language. Chomsky himself drew attention to properties common to all languages that do not arise in response to the requirements of communication or other functional considerations:

> A traditional view holds that language is a "mirror of the mind." This is true, in some interesting sense, insofar as properties of language are "species-specific" – not explicable on some general grounds of functional utility or simplicity that would apply to arbitrary systems that serve the purposes of language. Where properties of language can be explained on such "functional" grounds, the provide no revealing insight into the nature of mind. Precisely because the explanations proposed here are "formal explanations," precisely because the proposed principles are not essential or even natural properties of any imaginable language, they provide a revealing mirror of the mind (if correct).
>
> …In contrast, consider the fact that sentences are not likely to exceed a certain length. There is no difficulty in suggesting a "functional" explanation for this fact; for exactly this reason, it is of no interest for the study of mind…Or consider the observation known as "Zipf's law": namely, if the words of a long text are ranked in order of frequency, we discover that frequency is expressible as a function of rank in accordance with a fixed "law" (with a few parameters) …a fact that can be explained on quite general grounds…Or consider a third case. It has been observed that hearers have great difficulty in interpreting sentences in which a relative clause is completely embedded in another relative clause: for example, the sentence "The

> book that the main read is interesting" is readily interpretable, but the sentence "The book that the man the girl married read is interesting" is much less so. This observation is easily explained…[and so] the result is of little interest. (Chomsky, 1971, pp44-5)

It is interesting to consider whether any of the grammatical universals are "mirrors of the mind" in the sense that they may have evolved to subserve other mental functions besides just language.

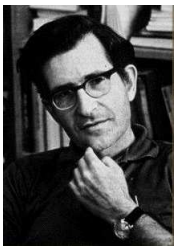
As noted in §0 and §5, all languages have **subject-predicate structures:** they have parts that can refer to particular things and parts that express properties that things have. These notions (esp. "refer" and "express properties") are not perfectly understood, but if something like this turns out to be true, it is plausible that it may be due to some basic facts about how we think about things in our environment, rather than to special requirements of human language. But we observed in §5 that Chomsky draws attention to a more subtle property of language that he called **structure dependence**. We saw this in the rules for auxiliary verbs in English, and in the rules for reordering elements in Quechua. Furthermore, he says it is something that is not needed (at least, not very much) for computer languages. Chomsky says:

> Consider the sentence "The dog in the corner is hungry"…the subject …is "the dog in the corner"; we form the question by moving the occurrence of "is" that follows it to the front of the sentence. Let us call this operation a "structure-dependent operation," meaning by this that the operation considers not only the sequence of elements that constitute the sentence but also their structure…
>
> Though the example is trivial, the results is nonetheless surprising, from a certain point of view. Notice that the structure-dependent operation has no advantage from the point of view of communicative efficiency or "simplicity." If we were, let us say, designing a language for formal manipulations by a computer, we would certainly prefer structure-independent operations.
>
> Notice further…though children make certain errors in the course of language learning, I am sure that none make the error of forming the question "Is the dog that in the corner is hungry?" despite the slim evidence of experience and the simplicity of the structure-independent rule. (Chomsky, 1971, pp26-28)

This does not make the notion of structure-dependence perfectly clear, nor does it make a clear suggestion about the origins of this property, but a natural guess is that the idea is something like this: linguistic expressions with this kind of 'structure' are not needed to simplify the system or facilitate communication, so maybe the structure reflects how we think, 'mirroring the mind' in this sense. It is difficult to assess whether this claim is this really true until we pin the terms down more carefully. It is now known that the introduction of structured expressions (of the sort we have in our rules for RNA sequences, English and Quechua) makes the grammars more expressive, and that more expressive grammars often allow more concise definitions of languages. Let's pause to consider this.

As mentioned at the beginning of section §4, the information-theoretic definition of "language" is different from the linguists' conception of that term, and the linguists' conception is

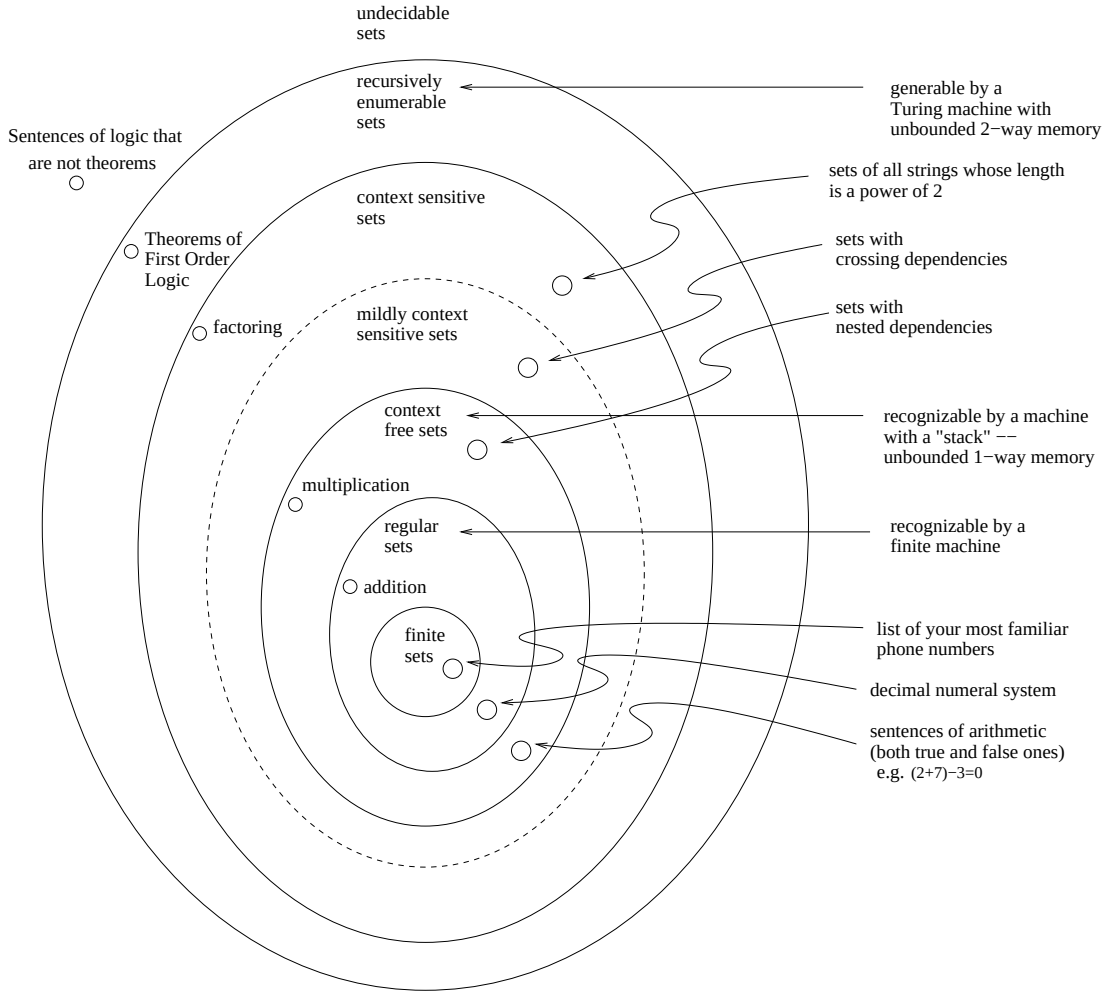
different from the commonsense one. For one thing, when we ask what language you speak, we are not interested in the one that teachers and "intellectuals" say you should speak, but in the language you actually do speak. (If you speak multiple languages, then we are interested in how you use each of them, and whether you use both of them – e.g. in "code-switching" utterances that have words from more than one language.) Fortunately, the different language users in a community have 'similar' languages, each one some slight variant of "English" or "Chinese" or whatever, so it is possible, at least as an approximation, to ask about the "properties of English," or of "Chinese," where by that we mean the common properties of the language of most speakers of those languages.


Chomsky

Like the sequences of nucleotides in DNA, the common sequences of morphemes in human languages are very far from random. They have natural units of various sizes. As we did for DNA, we propose definitions of human languages that capture this non-randomness, the "chunks" of structure. In §2.5.4, we saw that DNA has nested and crossed dependencies. (Chomsky, 1956) noticed this, and it turns out to be very important. It turns out that the kinds of computing systems that can recognize languages that simply extend to the right is **different from** systems that can recognize languages with nested dependencies, and these are **different from** systems that can recognize languages with crossing dependencies. These different kinds of patterns form part of the **Chomsky hierarchy**, shown below. Many problems in math and computer science have been located in this hierarchy too.[2]

Notice the position of languages with (unbounded) crossing dependencies in the hierarchy drawn below. We can represent crossing dependencies with grammars that define structured expressions (of the sort in our rules for RNA sequences, English and Quechua), but this kind of expressiveness might well be useful for other cognitive faculties as well. (I don't think Chomsky would like the suggestion that this kind of complexity is what he meant by "structure dependence" in the grammar; it is not quite clear. But this interpretation makes sense of his suggestion that he has a property in mind that we see in human languages but not in standard computer programming languages.)

[2]The Chomsky hierarchy is covered in detail in any standard introduction to the theory of computation, like Computer Science 181.



It is easy to show that that different kinds of computing systems – with different sorts of access to memory – are required to recognize each of these kinds of patterns.

6.7 Self-organization of language abilities?

This chapter has considered the emergence, not of particular human languages, but of the human language abilities that make those languages possible. Our understanding of these abilities (and especially their neural realization – cf. §5.1.4) is still at an early stage, and so I do not know how to regard any significant aspect of them as finding their shape by self-organization. But as we will see next week, a case can be made for regarding certain aspects of particular languages as self-organized – as emerging from basic properties of their parts and global constraints. As we will see there, certain properties related to complexity and effort do not need to be imposed from outside the organism, by selection, but can be seen as properties that any human activity (or any language-like human activity) would have, analogous to mechanical limits on biology. (This is something that Pinker & Bloom seem not to have considered.)

Exercises

1. Pinker and Bloom (1990) argue that a paperweight could have been created for any number of things, while a television is so complex it is exceedingly unlikely that it would be created for anything other than receiving and displaying television signals. And they list the properties of language that they think exhibit the kind of "adaptive complexity" we see in a television but not a paperweight, on pages 11-13 of the version linked on the webpage. This is sometimes called the **design argument**: language seems to have a complex, adaptive design.

   a. Which of the properties listed by Pinker and Bloom on pages 11-13 of their paper on the web page best support their argument? Briefly explain why.

   b. Which of the properties listed by Pinker and Bloom look most like they could have come from something other than selection? Which of them could have come from prior use outside of language (exaptation), or which could be due to basic requirements on how the parts work (self-organization)?

2. Explain why the diversity of human languages might seem to undermine the design argument. And briefly explain whether you think the response to this worry in Pinker and Bloom (1990) is persuasive.

3. In the passage quoted on page 177, Chomsky suggests that the structure-dependence of language does not improve efficiency of communication and does not simplify the grammar or language perception and production. This might be taken to suggest that language did not emerge by selection: it has these important features that do not have any adaptive function, features could easily have been otherwise. But Pinker and Bloom argue that this would be a mistake: a property that has no function or that could have been different, does not show that the faculty did not evolve by selection. Why do they say this? Are they right?

4. Consider the following parody of Hauser, Chomsky, and Fitch (2002):

   > Let's define EB, the "(vertebrate) eye in the broad sense" as the whole collection of systems that comprise the visual system: the whole eye itself, plus the extraocular muscles and coordination abilities, the optic nerve, visual cortex,.... And let EN be the "eye in the narrow sense," by which we mean the essential ability to detect light. EB obviously includes EN, but EN itself is not complex: it is really just the ability of rhodopsin and related proteins to change state when they are hit by light. Now consider these claims: (NH1) EN is simple, and so the argument from design is nullified, (NH2) basic parts of EB other than EN plausibly evolved independently of vision (though of course, once deployed in vision they may have been further modified). We conclude that the eye in the narrow sense, EN, did not emerge by selection, and that selection acted on EB only after EN emerged, simply adjusting existing systems to the new capability.

   Is this position as plausible as the position of Hauser, Chomsky and Fitch with respect to FLN and FLB? Explain why or why not.

5. In a recent textbook called *Evolution and Human Behavior* (2000), John Cartwright says:

> There has always been a strong anti-adaptationist tradition in linguistics. Noam Chomsky, one of the world's leading linguists, and Stephen Jay Gould, a prolific and widely read evolutionary theorist, have both repeatedly argued that language is probably not the result of natural selection. Gould's position seems to stem from a general concern about the encroachment of adaptive explanations into the territory of human behaviour. …he has used the term 'Panglossianism' to deride those who see the products of natural selection in every biological feature. Gould seems to have a view of the brain as a general purpose computer that, being flexible, can readily and quickly acquire language from culture without needing any hard wiring. Gould's output and influence have been great but one cannot help but feel that his scepticism towards an evolutionary basis for language stems in part from a political agenda that may be well intentioned but unreasonably resistant to any claims for a biological underpinning of human nature.
>
> Chomsky takes the view that language could have appeared as an emergent property from an increase in brain size without being the product of selective forces. He argues that when 10^{10} neurones are put in close proximity inside a space smaller than a football, language may emerge as a result of new physical properties. Chomsky's position is all the more surprising since he has battled long and hard to show that a language facility is something we are born with and not something that the unstructured brain simply acquires by cultural transmission.

OK, here are the questions:

i. What are the main <u>scientific</u> considerations that support Chomsky's and Gould's view that human language abilities (broadly construed) may have emerged for reasons that have nothing to do with language?

ii. What are the main <u>scientific</u> considerations on the other side, considerations suggesting that human language abilities (broadly construed) may have emerged because they were selected for their communicative value?

iii. Do you agree that the weight is so strongly in favor of the selectionist perspective that to make sense of Gould's position we need to assume that it derives from some political agenda? Briefly explain.

6. In another recent book called *Not by Genes Alone* (2005), anthropologists Peter Richerson and Robert Boyd write:

> When the environment confronts generation after generation of individuals with the same range of adaptive problems, selection will favor special-purpose cognitive modules that focus on particular environmental clues and then map these cues onto a menu of adaptive behaviors. Evidence from developmental cognitive psychology provides support for this picture of learning – small children seem to come equipped with a variety of preconceptions about how the physical, biological, and social world works, and these preconceptions shape how they use experience to learn about their environments. Evolutionary psychologists think the same kind of modular psychology shapes social learning. They argue that

culture is not "transmitted" – children make *inferences* by observing the behavior of others, and the kind of inferences that they make are strongly constrained by their evolved psychology. Linguist Noam Chomsky's argumentthat human languages are shapted by an innate universal grammar is the best-known version of this argument, but evolutionary psychologists think virtually all cultural domains are similarly structured.

i. Are Richerson and Boyd right about selection favoring "special-purpose modules" in the situation they describe? Why does this happen? (remember our discussion of the Baldwin effect and "genetic assimiliation")

ii. The idea that "special-purpose modules" are selected for problems like language learning seems, at least on the face of it, inconsistent with Chomsky's recent proposal that language (language in the "narrow sense") looks like it may have emerged not by selection but rather in a simple, sudden and uniquely human step, perhaps as a kind of exaptation or recruitment of an ability from another domain? Where is this inconsistency coming from, and what is the right view about the matter? (defend your view!)

Selected Solutions

There are various defensible responses to these questions, but it is important to at least mention the main points.

1. a. The following properties of human languages are listed by Pinker and Bloom to indicate their 'adaptive complexity':

   i. Grammars are built around N,V,A,P with characteristic roles, meanings, and subcategories

   ii. Phrases are built from some head X combined with specific kinds of phrases and affixes

   iii. Rules of linear order, which often signal what the subject, object (etc) are

   iv. Case affixes on N and A can sometimes signal what the subject, object (etc) are (see p.155)

   v. Verb affixes signal tense, aspect

   vi. Auxiliaries (either affixes or in VP-peripheral position) signal truth value, modality, force

   vii. Languages frequently have pronouns and related elements

   viii. Mechanisms of "complementation and control" provide embedded sentences and their interpretation

   ix. Wh-words question particular parts of sentences

   They conclude: "Language seems to be a fine example of 'that perfection of structure and coadaptation which justly excites our admiration' (Darwin 1859)." But what I notice in this list is that it is not so easy to see "complexity of design" and "perfection of structure" in the listed features, as it is in the structure of the eye, for example.

   Maybe the best support for their argument comes from the verb affixes and auxiliaries, since each depends on the other to function properly in human language.

   Or maybe the best support their argument comes from the way some combination of linear order and case-affix marking combine to determine what the subject and object of each sentence is, since here we often have two different kinds of things working together.

   b. The categories N,V,A,P might be just learned, and have the properties they do because we refer to things (N), talk about their relations (V) to one another, and modify our descriptions in various ways (A,P) in all languages, and this may well have a non-linguistic, conceptual basis.

2. The diversity of human languages might seem to undermine the design argument, because it could seem that languages are entirely learned, and a general learning mechanism would suffice.

   But Pinker & Bloom respond by pointing out that "there is no psychologically plausible multipurpose learning program that can acquire language as as special case, because the

kinds of generalizations that must be made to acquire a grammar are at cross-purposes with those that are useful in acquiring other systems of knowledge from examples."

This is a persuasive point (and it is supported by the results of Gold and others showing that no learning strategy can learn just anything, discussed in class and mentioned later on page 187).

3. Pinker & Bloom point out that a trait or organ can be selected because it has a valuable property, even if it has other properties that have no adaptive value, and also that "the fact that one can conceive of a biological system being different than it is says nothing about whether it is an adaptation." These points are clearly right! The structure-dependence of language, even if it had no adaptive value (which is very debatable!), would not show that language abilities were not selected, so long as they have other properties that do have selective value.

4. The parody misses the important fact that the eye in the broad sense, the EB, includes many adaptations that would have no value at all if it were not for the light-sensing proteins in the retina. This is why the argument from design applies to the whole complex of the EB: at least many of these things must have evolved together.

   In language though, the parts of the LFB not included in the LFN are things that would be valuable even without LFN: memory, coordination, perception,...

Lecture 7

Origins of particular languages and structures: Selection, exaptation, self-organization

Darwin proposed that biological evolution has these basic ingredients:

variation: in organisms, genetic variation is introduced by mutations.
Furthermore, we see that geographic isolation can lead to genetic divergence, and so then special things can happen when organisms from different ecosystems are brought into contact.

reproduction: in organisms, this is the mechanism for transmission of certain traits

selection: only a lucky few organisms survive to reproduce

Darwin imagined that selection was the formative influence in life, as we see in the famous conclusion to his *Origin of Species* that we quoted in the first lecture notes, on page 17. But we observed that there are other kinds of explanations for the properties of organisms:

exaptation: a trait can emerge as a consequence of selection for another trait, or a trait can be selected for one reason and come to fulfill a quite different function later

self-organization: some traits do not need to be imposed by an external force like selection (killing off less fit individuals before they reproduce), but rather they can emerge because of basic properties of the organism and the environment itself. This kind of principle helps us make sense of certain limits, of traits that many organisms share, and of special traits that emerge repeatedly in convergent evolution.

Furthermore, we saw that selection (and self-organization) act (simultaneously) at different levels:

hierarchical theory: selection can act at many levels simultaneously: it acts on genes, cells, multicellular organisms, demes (groups of related organisms), species,...

After a quick survey of these basic ideas in the first few weeks of the class, we noticed that they raise the question of what should count as 'life,' and what other kinds of things could evolve.

It is surprisingly natural to regard language and other cultural artifacts as entities that are evolving, even though they may not be 'living.' In the development of cultural entities, we have close analogs of all the ingredients outlined above:

variation: languages vary as new words and structures are introduced. Furthermore, we see that geographic isolation can lead to linguistic divergence, and so then special things can happen when languages come into contact. Also languages can change when one generation "misunderstands" or "reanalyzes" constructions of the previous generation – these are like mutations, or "transmission errors."

reproduction: in organisms, languages are transmitted by learning

selection: not all words and structures survive: roughly, only the most useful ones will persist

Notice that the method of reproduction, of transmission, differs from biological reproduction in that it is **Lamarckian**, and it seems likely that, at least with respect to many traits, inheritance is blending rather than particulate. That is, acquired traits can be transmitted, and the response to seeing two different ways of doing something is not always one or the other, but sometimes a kind of "blend" of the two.[1] Consequently, language changes can be very rapid and can introduce novel structure. "Popular" new words and constructions can spread like wildfire! Furthermore, there can be other influences on language besides selection.

exaptation: a trait can emerge as a consequence of selection for another trait, or a trait can be selected for one reason and come to fulfill a quite different function later.

self-organization: some traits do not need to be imposed by an external force like selection (losing the less useful words and constructions) but rather they can emerge because of basic properties of the language itself. This kind of principle could help us make sense of certain traits that languages share, and special traits that emerge repeatedly in convergent evolution.

hierarchical theory: selection can act on particular languages in particular speakers, on a whole community of speakers (so the whole community of English speakers could be regarded as analogous to an organism), or even to groups of languages.

Let's first consider language transmission (learning) a little more carefully, since it plays such an important role in this picture, and then consider various properties of human language that might fit into this picture.

[1] Darwin actually worried about whether particulate inheritance would remove variation, but Fisher showed that variation would persist; in fact, it efficiently preserves variation. So now the opposite question comes up: with blending Lamarckian inheritance, will enough variation persist for selection to have a shaping influence? Yes, at least in some situations. See, e.g. Boyd and Richerson (1985, pp71ff).

7.1 Language transmission: learning

Looking at the origins of language abilities, we observed that all human languages share quite a large number of distinctive properties, so we are in a strange position when we turn to consider how languages are learned. It is strange, because when we think about learning there is often an implicit assumption that, at least in principle, anything can be learned. It might be more or less difficult to learn one thing or another, but commonsense does not usually begin with the assumption that the learner comes with some ideas at the start, and that the learner is only capable of learning certain kinds of things. Philosophers call this perspective a "rationalist" or "nativist" one, as opposed to an "empiricist" one. Chomsky puts the point again this way:

> Even knowing very little of substance about linguistic universals, we can be quite sure that the possible variety of languages is sharply limited. Gross observations suffice to establish some qualitative conclusions. Thus, it is clear that the language each person acquires is a rich and complex construction hopelessly underdetermined by the fragmentary evidence available. This is why scientific inquiry into the nature of language is so difficult and so limited in its results...it is frustrated by the limitations of available evidence and faced by far too many possible explanatory theories, mutually inconsistent but adequate to the data...Nevertheless, individuals in a speech community have developed essentially the same language. This fact can be explained only on the assumption that these individuals employ highly restrictive principles that guide the construction of grammar. Furthermore, humans are, obviously, not designed to learn one human language rather than another; the system of principles must be a species property. Powerful constraints must be operative restricting the variety of languages. (Chomsky, 1975, pp10-11)

Chomsky's nativist view gets a surprising kind of support from mathematical studies of learning. The rough idea is easy to describe. We can think of the language learner as a function from evidence to hypotheses about the world. In the case of language learning, the evidence is some sequence of utterances (possibly with context), and the learner's hypotheses are grammars. (Human language learners typically get correction and instruction, not just examples to learn from, but as we mentioned on page 125, it seems that correction and instruction is not necessary.) Idealizing, we could imagine that the learner can remember everything, and that the sentences heard would include everything in the language if the learner could listen forever. In this very idealized setting, it is easy to describe a learner for a finite language:

non-generalizing learner: At each point, the learner guesses that the language is exactly what has been heard so far.

This learner is not very interesting, but will succeed if the language is finite. That is, this learner can learn any of the finite languages. Obviously, though, if we added an infinite language to the collection, then this learner could not learn it. What is more surprising, though is that no learner can learn a class of languages that includes all the finite ones plus some infinite ones. In a simple mathematical setting, this result was proven by Gold (1967), and became a

foundation for the mathematical study of learning.[2] And with probabilistic methods too, we find again that only learning problems with certain special structural properties can be solved with feasible resources.

At this point, someone might object that although the patterns we see in language are not length-bounded, it is an idealization to think of the language as actually infinite, and does not imply that the learner actually needs to learn anything that is really infinite. But this is a confusion. In the first place, the claim is just that learners naturally notice patterns and that these patterns are not length-bounded. In the second place, the same kind of point would apply even to patterns that were length-bounded: if you generalize (= if you notice patterns in data), this is going to lead you to assume that certain kinds of things would not occur accidentally, and this has the obvious impact on what you can learn.

The basic idea here is sometimes called a "poverty of the stimulus" argument. When a language is infinite (when the patterns in it are unbounded), you don't need to see them all to recognize them, and so we have to explain how it is that we all extend our grammars to sentences we have never heard before in essentially similar ways. This might be explicable if we have an "innate idea" about what kind of thing we all expect language to be, but is hard to see how it could be explained if any extension of our experience could count as part of the language. The study of how children actually learn their language shows that what is happening is very complex (cf. §5.1.1), but it is clear that they do, in fact, regularly generalize in certain natural ways. That is not surprising, but it can lead the philosophers to troubling conclusions about what we could possibly know about the universe.

7.1.1 Locke and others against innate ideas

The "empiricist" wants to stick with the sensible-sounding idea that knowledge can only come from the evidence presented to our senses. The British philosopher John Locke famously defended this perspective, saying in *An Essay Concerning Human Understanding* (1690) that the mind is like a "blank slate" or "blank page" upon which experience writes:

> I know it is a received doctrine, that men have native ideas, and original characters, stamped upon their minds in their very first being. This opinion I have at large examined already; and, I suppose what I have said in the foregoing Book will be much more easily admitted, when I have shown whence the understanding may get all the ideas it has; and by what ways and degrees they may come into the mind; – for which I shall appeal to every one's own observation and experience.
>
> All ideas come from sensation or reflection. Let us then suppose the mind to be, as we say, white paper, void of all characters, without any ideas: – How comes it to be furnished? Whence comes it by that vast store which the busy and boundless fancy of man has painted on it with an almost endless variety? Whence has it all the materials of reason and knowledge? To this I answer, in one word, from Experience. In that all

[2] This mathematical subject is booming recently, with conferences and several recent, good texts devoted to it and to its application to language learning (Kearns and Vazirani, 1994; Jain et al., 1999; Hastie, Tibshirani, and Friedman, 2001; Duda, Hart, and Stork, 2001).

> our knowledge is founded; and from that it ultimately derives itself. Our observation employed either, about external sensible objects, or about the internal operations of our minds perceived and reflected on by ourselves, is that which supplies our understandings with all the materials of thinking. These two are the fountains of knowledge, from whence all the ideas we have, or can naturally have, do spring.

He warns us against those who are "extending their Enquiries beyond their Capacities, and letting their Thoughts wander into those depths where they can find no sure Footing; 'tis no Wonder, that they raise Questions and multiply Disputes, which never coming to any clear Resolution, are proper to only continue and increase their Doubts, and to confirm them at last in a perfect Skepticism."

Certainly we can share Locke's desire for clarity, but (speaking of clarity) notice that we cannot really tell whether his view that "all ideas come from sensation or reflection" even conflicts with modern nativism until we see what he means by the reflection and "internal operations of our minds" that act on "the materials of thinking." The nativist view described above does not say that you can acquire a language with no sensory input at all, but only that certain generalizations from that experience and not others are natural. Looking more carefully at the mechanisms Locke provides for associations of ideas, and at the kind of structure he imagines the senses impose upon experience, it is no surprise that his proposals are not up to the task of explaining language acquisition as we now understand it. Now, the mechanisms proposed for learning in linguistic and other domains are not only more complex, but they bring a bias towards particular kinds of conclusions that Locke would have worried about, but which do not seem to be escapable. Nevertheless efforts to escape all bias in learning, or failing that, to set what bias there is on a firmly "rational" foundation, continue! (Not to mention the efforts to deny there is bias even when it is perfectly plain.)

In sum, language is transmitted by learning, but we do not assume that just anything can be learned. Rather, language learning seems to fill out a structure whose basic outlines are invariant across languages. (In the same way, we assume genetic traits are transmitted and selected, but we do not need to assume that everything could emerge that way: many things may be biologically impossible.) Within all this structure, there is lots of variation, and some aspects of this variation can be passed from one generation to the next by learning, which is a Lamarckian evolutionary mechanism, since acquired properties can be and are transmitted in this way.

7.2 Language variation

What introduces language variation? The sources of variation are many, and their interactions complex. They include

- individual creativity of various kinds (new names and acronyms, new pronunciations of existing forms, extensions of meaning,...).

- "imperfect" transmission. Like a mutation, a linguistic structure can be misinterpreted by later generations and find its role in the language significantly changed as a result.

  For example, the English word *orange* seems to have come from the Sanskrit *naranga*, Arabic *naranja* or Spanish *naranja*, but then it seems likely that *a norange* was reanalyzed as *an orange*. (Pinker, 1994, p245)

  Slightly more elaborate and consequential reanalyses can happen in a similar way too. For example, it appears that English modals like *will, would, may, might,...* were originally normal verbs triggering an *-en* ending on the following embedded verb, but the ending was lost and then *will, would, may, might,...* were reanalyzed as the special forms we know today (Lightfoot, 1999; Roberts and Roussou, 2002; Roberts, 1993)

- **language isolation** leads to diverse forms, and then **language contact** can yield novel results.

Why does the biological endowment for language allow so much plasiticity? Why is language variation so extensive? Why wouldn't our genetic endowment determine more aspects of language than it does. This question is about our basic language abilities, and so it really belongs in the previous chapter, but it comes up now because we are considering how cultural evolution could shape our languages. The proposals about this are very speculative and controversial:

- thinking back to plasticity's cost/benefit tradeoffs discussed in §3.6, it is natural to propose: a language flexible enough to offer new expressive capabilities can be advantageous (Nowak, Komarova, and Niyogi, 2002).

  The addition of new expressive capability could explain the spread and persistence of certain new words. Inhabitants of the deserts are likely to have words for varieties of cactus, while inhabitants of the arctic are more likely to have words for seals.

  Notice that this kind of pressure would never explain why any whole language community (like English) would come to dominate others (like all the indigenous American languages). Any human language is easily extended to express whatever can be expressed in any other; they do not differ in expressive potential.

- Dyson (1979) suggests that linguistic diversity divides groups and isolates them, facilitating more rapid evolution

  Rejecting this idea, Pinker (1994, pp240f) and Baker (2001, pp210f) point out that traits are not selected because of their later benefits. Apparently their idea is that it is implausible that human biological evolution has already been accelerated enough by language to make this explanatory (Pinker, 1994; Baker, 2001, for example).

  But Dyson's suggestion applies readily to cultural evolution: isolation will produce divergence and more rapid cultural evolution. And cultural evolution will have biological consequences. For example, paleontologists have speculated about the rather sudden extinction of other hominids right around the time when *Homo sapiens* began to show some technological sophistication:

  > Although the source of *H. sapiens* as a physical entity is obscure, most evidence points to an African origin perhaps between 150,000 and 200,000 years ago...About

> 40,000 years ago, the Neandertals of the Levant yielded to a presumably culturally rich *H. sapiens*, just as their European counterparts had…
>
> The earliest *H. sapiens* sites [in Europe] date from only about 40,000 years ago, and just 10,000 or so years later the formerly ubiquitous Neandertals were gone. Significantly, the *H. sapiens* who invaded Europe brought with them abundant evidence of a fully formed and unprecedented sensibility…The pattern of intermittent technological innovation was gone, replaced by constant refinement. Clearly, these people were *us*.
>
> …anatomically modern humans behaved archaically for a long time before adopting modern behaviors. That discrepancy may be the result of the late appearance of some key hardwired innovation not reflected in the skeleton, which is all that fossilizes. But this seems unlikely, because it would have necessitated a wholesale Old World-wide replacement of hominid populations in a very short time, something for which there is no evidence.
>
> It is much more likely that the modern human capacity was born at – or close to – the origin of *H. sapiens*, as an ability that lay fallow until it was activated by a cultural stimulus of some kind. If sufficiently advantageous, this behavioral novelty could then have spread rapidly by cultural contact among populations that already had the potential to acquire it. No population replacement would have been necessary to spread the capability worldwide.
>
> It is impossible to be sure what this innovation might have been, but the best current bet is that it was the invention of language. (Tattersall, 2003)

Language and cultural evolution could certain play a role in this kind of event.

- Weakening Dyson's idea a little: perhaps language differences just promote group solidarity. But Pinker (1994, p241) and Baker (2001, pp212f) suggest that the diversity we see far exceeds what would be required to distinguish groups. In sum, we are in need of more clearly defined proposals to sort out all this controversy!

Given the rate of language change, only recent advances in travel and communication make a (near-)universal inter-lingua conceivable. In earlier times, language communities could change in distinctive ways, for much longer periods, without contact with other languages.

7.3 Selection, exaptation, or self-organization?

We already observed on page 158 that variation in human languages correlates with variation in the human genome. Languages like English, French and German have similarities (esp. related lexical items!), which is no surprise given their well-known historical connections, and so they are grouped into the "Indo-European" family of languages. We have a number of language groups like this, and some more speculative super-classifications of these language families. But we did not consider there whether the development of one language from another could itself be regarded as an evolutionary change. This perspective is not always explicit, but of course it is the standard idea about cultural development and diversification. So now let's

briefly consider some properties of language and of language change, watching to see what kinds of explanations should be offered from an explicitly evolutionary perspective.

7.3.1 Discrete syntax

The very first thing that should be mentioned among the universals of human languages is a property that they share with the DNA language: they are discrete symbol systems. That is, human languages are "digital" not "analog," in the sense that a word's identity is categorical rather than varying along a continuum. The communication of bees and many chemical signaling systems are analog, with the meanings of a symbol varying along a continuum that corresponds to a continuum in what is meant. But human language does not work that way.

With the information revolution of the 1900's came a recognition of the advantages of digital information transmission, even when the information being transmitted is really analog, as we see in music recording technology for example. There are two main reasons for this. (1) Digital signals can be more resistant to noise. a little bit of noise will not suffice to change one word into another. And (2), errors can be recognized in digital signals if there is a little bit of redundancy,

In connection with Shannon's measure of "information" in Lecture 4, we already mentioned that human languages are, in fact, redundant, but here are some other ways to look at it. First, it is possible to replace a phoneme by a cough or some other noise in such a way that the listener does not even notice that the phoneme is missing – this is sometimes called the "phoneme restoration effect" (Warren and Warren, 1970; Warren and Sherman, 1974). And second, there are many ways to degrade linguistic input while leaving it still intelligible. For example, deleting all the vowels in usually still leaves a readable written text:

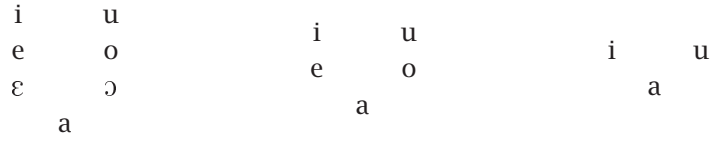
> Thanks to the redundancy of language, yxx cxn xndxrstxnd whxt x xm wrxtxng xvxn xf x rxplxcx xll thx vxwxls wxth xn "x" (t gts lttl hrdr f y dn't vn kn whr th vwls r). In the comprehension of speech, the redundancy conferred by phonological rules can compensate for some of the ambiguity in the sound wave. For example, a listener can know that "thisrip" must be *this rip* and not *the srip* because the English consonant cluster *sr* is illegal. (Pinker, 1994, p181)

This kind of redundancy is desirable when the communication channel is (at least sometimes) "noisy," and we can understand this property as emerging and persisting in language (at least in part) for that reason. So this is an example of a property that could emerge by **selection** for communicative efficiency, but since it is found in all human languages, it is probably more naturally attributed to basic, genetically conditioned mechanisms of categorization.

7.3.2 Dispersion in sound systems

Another property related to the noise tolerance of language is a kind of dispersion. When the sound system of a language has just 3 vowels, it never has just [o ɔ ʊ]. Rather, languages tend to choose vowels that are as perceptually and articulatorily distinct as possible like [i a u]. We

find that the trio [i a u] and enrichments of it like the following, occur in the world's languages (Lindblom, 1998; Flemming, 2002)



So it is natural to regard the sounds always as "dispersed" in perceptual and articulatory space. What explains these facts? The dispersion of sounds is naturally regarded as a kind of self-organization (Lindblom, MacNeilage, and Studdert-Kennedy, 1984). That is, the properties of the sounds themselves determine their suitability in one or another sound system, relative to the global requirements of perceptual and articulatory distinctiveness.

Interestingly, the consonants do not seem to be dispersed in a similar way…or at least not obviously so (Lindblom and Maddieson, 1988). Understanding why consonants are distributed in the world's languages as they are is a topic of ongoing research.

7.3.3 Vocabulary introduction, meaning extension, related changes

Of course, new vocabulary is introduced into languages all the time: new people, new technologies, and new discoveries all precipitate new vocabulary, and extensions of old vocabulary to new things. This is probably where language are most visibly and most rapidly changing. Computers used to be people who did calculations, but now they are machines on our desks. And there is the famous example of the words of French origin *pork, beef, veal* coming to replace the more "common" use of the Old English forms for *pig or swine, cow, calf* as the names for what we eat. Another kind of extension of meaning that is fairly common is the popular use of words for extremes to apply to things that actually are not so extreme. My shoes are "awesome;" the class was "fantastic." Once these words become mainstream, they no longer have the same extreme feeling, and so we need new ones.

Lightfoot (1982, p153) compares this to a possible origin of major structural changes:

> It is a fact of biological necessity that languages always have devices to draw attention to parts of sentences, and people may speak more expressively by adopting a novel or unusual construction, perhaps a new word order…Dislocation sentences fall under this rubric: *Mingus, I heard him*, and *He played cool, Miles*. These forms, still regarded as novel in English and as having distinct stylistic force…However, such expressive forms characteristically become bleached and lose their novelty as they become commonly used…The special stylistic effect slowly becomes bleached out and the constructions lose their particular force, become incorporated into the normal grammatical processes and thereby require speakers to draw on their creative powers to find another pattern to carry the desired stylistic effect.

English changed from SOV to SVO, and this kind of change is fairly common. Many factors undoubtedly play a role in such a major change, and only some of them are now understood.[3]

7.3.4 Origins of English reflexive pronouns

There seem to be cases where a vocabulary item or structure originates with one role, and later speakers of the language start to use it with another role. We already observed that this happens when a word's meaning is extended to apply to new cases, but it can also happen in more surprising ways: For example, in a study of the origins of English reflexive pronouns (*himself/herself/itself/...*), Keenan (1997) observes that in Old English *self* was an independent word that followed definite DPs to indicate contrast or emphasis:

- Ne sohte ic na hine, ac he sylf com to me
  not sought I not him, but he self came to me
  'I did not seek him, but rather he came to me'

Later, by about 1050 or so, *self* disappeared as an independent word, but instead of eliminating *him self*, it starts getting used as a definite DP by itself:

- He becom þa to anre birig,..., & þa circlican þeawas him sylf þaer getaehte
  He came then to a town,..., and the churchly ways himself there taught

This looks like it could be regarded as **exaptation**: the old *him sylf* that was used for one purpose starts getting used as *himself* in a different role.

7.3.5 Compositionality, recursion and learnability

The language structures described in the Fregean style, by atoms plus combinations, are discrete – a property mentioned above – but also sometimes compositional and recursive. In human languages, the meanings of expressions are usually built up from the meanings of their parts. And we find plentiful recursion: almost every category of expression is one that can contain other instances of the same category. Where do these aspects of language come from?

One way of thinking about Frege's insight (mentioned on in §7) is that compositionality could have a learnability explanation. If the expressions of a language have common parts interpreted in common ways, then experience with those parts will extend to sentences that have never been heard before. The value of this property for language learning is obvious, and it has been demonstrated in simple computer simulations (Kirby, 1999a) and mathematical studies (Komarova and Nowak, 2001) Recursion could then follow from compositional language structure, when one proposition is embedded in another, when a named individual is part of a named group, or when a named action is part of a named complex of actions. The reasons for human languages being compositional again has this complex status: it is a universal property and so it is natural to assume that it may be specifically provided for (genetically) by the way

[3] A class on historical linguistics, like Linguistics 110, explores these matters carefully.

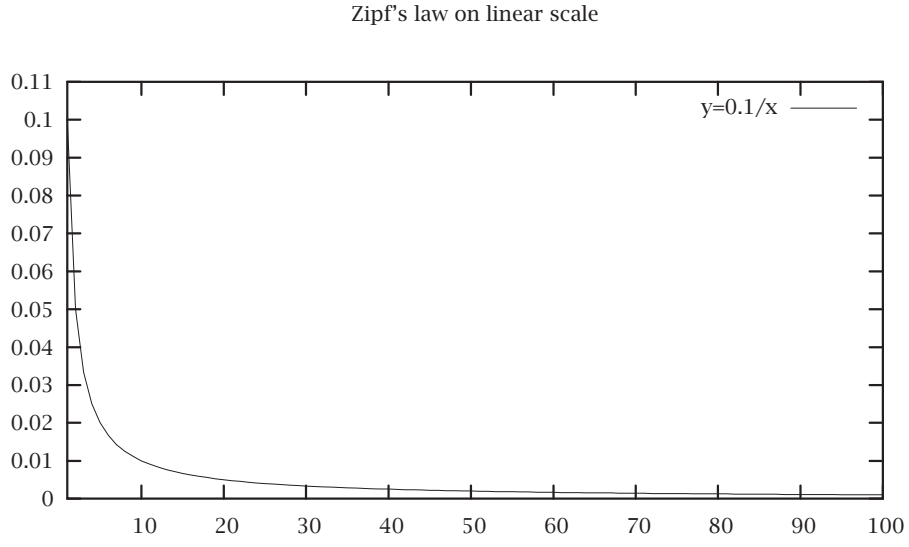
we recognize patterns, but also it is a property of language that would persist in a culture once it emerged because it has higher 'fitness' in the sense of being more easily transmitted.

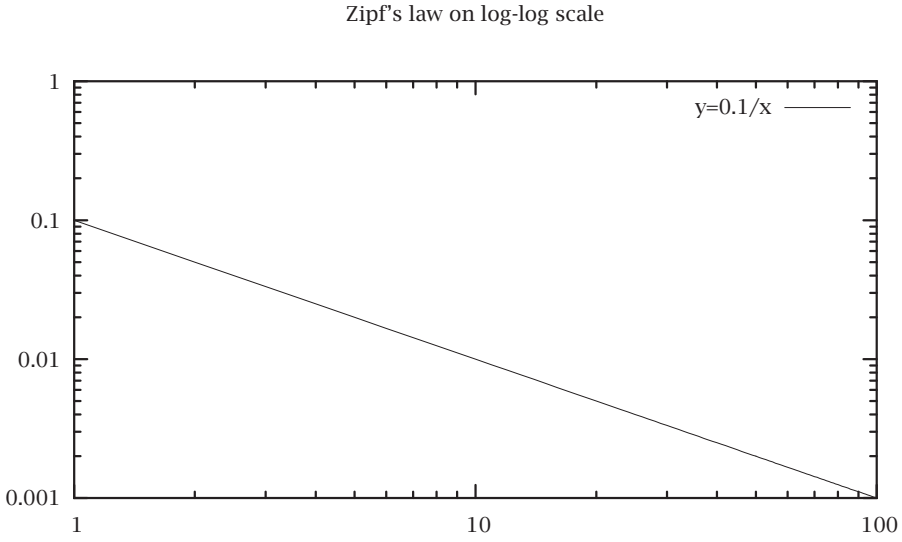
7.3.6 "Effort" and statistical properties

In early studies of texts, Zipf (1949) noticed that the distribution of word frequencies is not normal, and that there is a relation between word frequency and word length. In particular, in most natural texts, when words are ranked by frequency, from most frequent to least frequent, the product of rank r and frequency μ is constant; that in natural texts the function f from ranks r to frequencies is a function

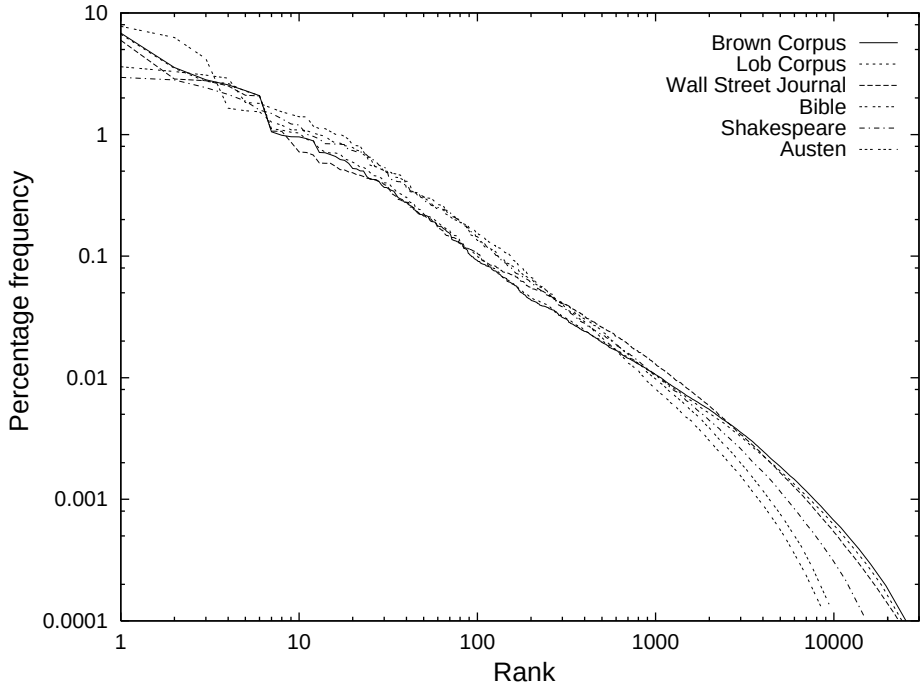
$$\mu(r) = \frac{k}{r}.$$

Plotted on a regular scale we get the usual inverse exponential curve, which becomes a downward sloping line on a log-log scale:




Zipf's law on log-log scale
y=0.1/x
1
0.1
0.01
0.001
1
10
100

We get this kind of relationship in most texts and collections of texts (Teahan, 1998):



Zipf proposed that the shortness of frequent words comes from a "principle of least effort:" frequently used vocabulary tends to be shortened. This idea may seem intuitively right – we know cases of explicit shortenings of words – but the statistical evidence in favor of the view that this kind of shortening explains Zipf's curve is extremely weak, because Zipf-like distributions emerge even with pure random word generators, as long as the word termination character is among the randomly distributed elements.[4] Consequently, there is no reason to assume that the process of abbreviation is a significant factor unless the distribution of words of various sizes departs significantly from what might be expected anyway. Since Zipf's regularities can emerge entirely from local tendencies to end words at some point (other things being equal), and to use certain words more than others, no more elaborate hypothesis is needed. Although many statistical properties of language remain unexplained, it seems unlikely that Zipf's explanation is a major factor.[5]

7.3.7 Other complexity bounds

Besides the tendency for words and sentences to be short, there are other, more surprising restrictions on languages. One famous one is the following. Notice that you can modify the noun [man] with a phrase like [who Bill likes] as in (1a), and you can question various parts of that statement as in (1b) and (1c):

[4] Mandelbrot (1961), Miller and Chomsky (1963, pp456-463).
[5] Cf. Li (1992), Niyogi and Berwick (1995), Perline (1996), Teahan (1998), Baayen (2001).

(1) a. The man [who Bill likes] shot the gangster
b. Did the man [who Bill likes] shoot the gangster?
c. Who did the man [who Bill likes] shoot?

But you cannot question *Bill*, like this:

(2) a. * Who did the man [who likes] shoot the gangster?

Similarly, you can modify [the teacher] with [who inspired Bill], but again you cannot question *Bill*

(3) a. I do know the teacher [who inspired Bill]
b. * Who do I know the teacher [who inspired]?

This fact is sometimes described this way (Ross, 1967):

The complex NP constraint: No wh-phrase can move from inside a complex NP
(where a "complex NP" is an NP with a clause, a sentence-like phrase, inside it)

We find this restriction (or something very close to it) in Japanese and many other languages:

- Otto ga kabutte ita koto o watakusi ga sinzita boosi wa akai
  Otto wearing was think I believed hat red
  'The hat [which I believed [that Otto was wearing]] was red'
- * Otto ga kabutte ita to iu syutoyoo o watakusi ga sinzita boosi wa akai
  Otto wearing was that say claim I believed hat red
  'The hat [which I believed [the claim [that Otto was wearing]]] was red'

It takes some work, but it can be argued that this is a complexity bound too, a bound deriving from the way memory can be accessed during the computation of sentence structure (Marcus, 1980; Berwick and Weinberg, 1984; Hawkins, 2001; Hawkins, 1999; Hawkins, 1990) Consequently, this would not need to be a special property imposed by some external influence, but is an internal property, a fact about the way language users work that follows from other basic properties.

7.4 Strange conclusions, and some related issues

This week we considered some ideas about how language change could be regarded as an evolutionary process, as the Lamarckian evolution of a cultural artifact. This seems possible since it exhibits variation; it is reproduced and transmitted by learning; and there is even a kind of selection in the sense that some languages (and structures within languages) arise and persist, while others arise and then disappear.

However, many of the proposals in this chapter are quite speculative. Cultural transmission (by learning) is rapid, and languages in contact can influence each other with only brief exposures. Though there may be atoms of variation, basic parameters of variability, their interactions are complex and we have no analog of the chemical basis of heredity to guide our exploration.

The evolutionary origins of language remain mysterious, but evolutionary theory is rather new, and rigorous scientific studies of human language even newer. I think there is cause for optimism. Our understanding of the range of human languages and their structural peculiarities is increasing rapidly. (Many languages are becoming extinct in our generation, but still there is a very wide variety that will survive for some time.) Our understanding of human learning and of learning in general is proceeding in leaps and bounds, with a tremendous surge of interest recently and new technologies that allow recording and analysis of teacher-learner interactions of a kind that has never been possible before. And our understanding of the biological bases of language abilities is advancing too, with the discovery of FOXP2, and new sophistication in technologies for studying neural activity.

In this drive to understand the origins of language, we have discovered many fundamental things that are already quite clear. We saw that languages have different kinds of patterns that can be classified according to their complexities. Strangely, of all the languages discussed here, it is the language of DNA/RNA and human language that turn out to be rather high in the complexity hierarchy, because they have both nested and crossing dependencies. (This is noted in the remark from Hauser, Chomsky and Fitch quoted on page 171, but they never offer an explanation of why it would be so.) This is almost as weird as the discovery that the same rhodopsin-like proteins are used in both light-sensitive bacteria and human vision, the sophistication of insect navigation, and the complexity of neural codes.

Why would DNA and human language be the most sophisticated discrete, symbolic languages that we see anywhere in the universe? Both of them have both nested and crossing dependencies, placing them rather high in the Chomsky hierarchy (as we saw on page 179). I conjecture that this is not because of some external force applying in each case, but for "self-organizing" reasons. DNA and human languages are high in the hierarchy because they both exploit a whole range of patterns, patterns that require rather sophisticated mechanisms to recognize (but patterns that can still be feasibly identified). The Chomsky hierarchy classifies all collections of structured objects, and so it should be no surprise to find a language that has evolved over millions of years, or languages developed much more quickly by creatures with remarkable pattern recognizing abilities, to be as expressive as possible but still efficiently recognizable.

If we had extra time, it would be interesting to explore some of the connections between what we have done and some other topics. I will briefly mention a couple that seem especially relevant.

7.4.1 Sociobiology and culture

The claim that the grammars of human languages have some of their properties for biological reasons has stirred a great deal of controversy. (Wilson, 2000) gets into similar controversial areas when he argues that a surprising range of social behavior seems to be at least partly under biological control:

> The optimistic prospect for sociobiology can be summarized briefly as follows. In spite of the phylogenetic remoteness of vertebrates and insects and the basic distinction between their respective personal and impersonal systems of communication, these two groups of animals have evolved social behaviors that are similar in degree of complexity and convergent in many important details. This fact conveys a special promise that sociobiology can eventually be derived from the first principles of population and behavioral biology and developed into a single, mature science. (Wilson, 1971, p460).

7.4.2 Language and thought

(Whorf, 1941) is well known for suggesting that language shapes the way we think and perceive the world. It is tricky to provide good support for this idea, because it is obvious that we develop our languages to describe the things we are thinking about and perceiving, and this is hard to disentangle from a causal influence going the other direction, from the language to our perceptual and cognitive abilities. Whorf himself took an extreme view which now has been pretty well debunked. If you use a language with few color terms, does this decrease your ability to perceive differences between two colors (like when you choose which shade of white to paint your house)? The short answer is: no. But really you have to be careful about exactly what question you are asking (Kay et al., 1997; Lucy and Shweder, 1988). Do languages affect the way we conceptualize our positions in space and time? Although there are significant differences among languages, the cognitive consequences are not so clear (Li and Gleitman, 2002).

7.4.3 Analytic and simulation studies of evolutionary language change

Since the factors influencing language development and language change are complex, one strategy is to study extremely simple mathematical models analytically, and another strategy is to study slightly more complex but still simplified models with simulation studies of the sort mentioned only briefly above (Niyogi, 1999; Nowak, Komarova, and Niyogi, 2002; Yang, 2000; Kirby, 1999a; Steels, 1996; Niyogi and Berwick, 1994). Though this kind of work abstracts away from very many (and possibly very relevant) details, it can provide illuminating suggestions about how different factors interact, and of the conditions under which evolutionary mechanisms could cause 'optimal' combinations of features to emerge.

Exercises

1. Consider these 3 perspectives on the emergence of recursive, compositional human language. Premack (1985) writes:

   > I challenge the reader to reconstruct the scenario that would confer selective fitness on recursiveness. Language evolved, it is conjectured, at a time when humans or protohumans were hunting mastodons...Human language is an embarrassment for evolutionary theory because it is vastly more powerful than one can account for in terms of selective fitness. A semantic language with simple mapping rules, of a kind one might suppose the chimpanzee would have, appears to confer all the advantages one normally associates with discussions of mastodon hunting or the like. For discussions of that kind, syntactic classes, structure-dependent rules, recursion and the rest, are overly powerful devices, absurdly so.

   Pinker (1994) takes up the challenge:

   > First, bear in mind that selection does not need great advantages. Given the vastness of time, tiny advantages will do...Second, if contemporary hunter-gatherers are any guide, our ancestors were not grunting cave men with little more to talk about than which mastodon to avoid. Hunter-gatherers are accomplished toolmakers and superb amateur biologists with detailed knowledge of the life cycles, ecology, and behavior of the plants and animals they depend on. Language would surely have been useful in anything resembling such a lifestyle....And grammatical devices designed for communicating precise information about time, space, objects, and who did what to whom are not like the proverbial thermonuclear flyswatter. Recursion in particular is extremely useful; it is not, as Premack implies, confined to phrases with tortuous syntax. Without recursion, you can't say *the man's hat* or *I think he left*....Third, people everywhere depend on cooperative efforts for survival, forming alliances by exchanging information and commitments. This too puts complex grammar to good use....But could these exchanges really produce the rococo complexity of human grammar? Perhaps. Evolution often produces spectacular abilities when adversaries get locked into an "arms race," like the struggle between cheetahs and gazelles....outwitting and second-guessing an organism of approximately equal mental abilities with non-overlapping interests, at best, and malevolent intentions, at worst, makes formidable and ever-escalating demands on cognition....Thus there could have been selection for any edge in the ability to frame an offer so that it appears to present maximal benefit and minimal cost to the negotiating partner, and in the ability to see through such attempts and to formulate attractive counterproposals.

   But maybe recursiveness in language is not biologically determined at all, not biologically specified either by selection or exaptation. Maybe the recognition of recursive structure is not a language-specific trait but quite general in cognition, and instead we ought to consider how and why it emerges in each particular human language. (So then we should not have

considered it in last week's, but in this week's lecture!) For example, Kirby argues that "we should not rush into a biological evolutionary explanation for such universals." He says:

> The picture that emerges…is of the language of the population acting as an adaptive system in its own right. Initially, the rules [of grammar] are minimally general, each pairing one string with one meaning. At some point, a chance invention or random noise will lead a learner to "go beyond the data" in making a generalization that the previous generation had not made. This generalisation will then compete with the idiosyncratic rule(s) for the same meaning(s). Given that generalisations are better replicators, the idiosyncratic meanings will be pushed out over time. The competition will then be replayed among the generalisations, always with the more general rules surviving. The inevitable end state of this process is a language with a syntax that supports compositionally derived semantics in a highly regular fashion. The grammar for such a language appears to be the shortest (in terms of numbers of rules) that can express the entire meaning space. (Kirby, 1999b)

Some questions to consider:

a. Explain why Pinker says that recursion happens even in simple phrases like *the man's hat* or *I think he left*

b. The last point is odd, because it is about English, which has lots of recursion. When Pinker says, "Without recursion, you can't say *the man's hat* or *I think he left*" could he really mean that something like the man's had or the proposition that you think he left could not be expressed in any language without a recursive structure? (explain why or why not)

c. Kirby suggests that even if every human language is recursive, we might be able to explain that because it is useful, in the sense that culturally transmitted systems of communication with this property will tend to survive. But doesn't this presuppose a biologically determined ability to perceive and produce recursive structures? How do you think Kirby would respond to this question?

d. What kind of evidence (think of anything that, hypothetically, could be available) could decide among these 3 perspectives?

2. Lamarck (1809) thought that people were descended from ape-like animals that he called quadrumanes (because he regarded them has having 4 hands instead of 4 feet), and he speculated about the origin of language in them:[6]

> The animals don't acquire new needs and so their ideas remain few and do not change. And among these ideas are very few which they would need to communicate to other individuals of the same species. Thus, they need only very few different signs to make themselves understood to those like them. Hence, some movements of the body or some of its parts, a few hissings and cries varied by simple vocal inflections are enough for them. By contrast, the individuals of the dominant race ["quadrumanes"] already mentioned, having had a need to multiply the signs to communicate quickly their ideas (which have become more and more numerous), and not resting content with pantomime signs or the possible inflections of their voice, in order to represent this multitude of signs which has become necessary would have succeeded, by different efforts, in forming articulated sounds. At first they would have used only a small number, combined with the inflections of their voice. Afterwards, they would have increased them, varied them, and perfected them, according to the growth of their needs and to the extent that they would have made more effort to produce them. In fact, the habitual exercise of the throat, tongue, and lips to articulate sounds would have really developed this faculty in them. From that would come, for this particular race, the origin of the admirable capability of talking. And since the distance between places where the individuals making up this race would have widened and encouraged the corruption of the signs agreed upon in order to convey each idea, from that would have originated languages, which have been diversified everywhere.

   What are the main things here that Lamarck has gotten wrong? And which things has he gotten right?

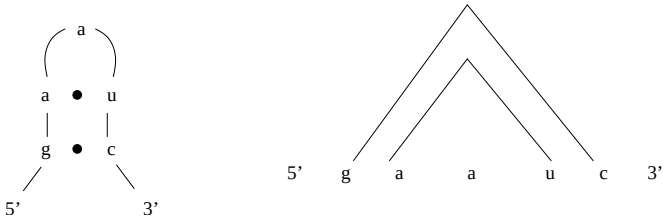
[6] The original French text: "ces animaux ne se forment plus de nouveaux besoins; n'acquièrent plus d'idées nouvelles; n'en ont qu'un petit nombre, et toujours les mêmes qui les occupent; et parmi ces idées, il y en a très-peu qu'ils aient besoin de communiquer aux autres individus de leur espèce. Il ne leur faut donc que très-peu de signes différens pour se faire entendre de leurs semblables; aussi quelques mouvemens du corps ou de certaines de ses parties, quelques sifflemens et quelques cris variés par de simples inflexions de voix leur suffisent. Au contraire, les individus de la race dominante, déjà mentionnée, ayant eu besoin de multiplier les signes pour communiquer rapidement leurs idées devenues de plus en plus nombreuses, et ne pouvant plus se contenter, ni des signes pantomimiques, ni des inflexions possibles de leur voix, pour représenter cette multitude de signes devenus nécessaires, seront parvenus, par différens efforts, à former des sons articulés: d'abord ils n'en auront employé qu'un petit nombre, conjointement avec des inflexions de leur voix; par la suite, ils les auront multipliés, variés et perfectionnés, selon l'accroissement de leurs besoins, et selon qu'ils se seront plus exercés à les produire. En effet, l'exercice habituel de leur gosier, de leur langue et de leurs lèvres pour articuler des sons, aura éminemment développé en eux cette faculté. De là, pour cette race particulière, l'origine de l'admirable faculté de parler; et comme l'éloignement des lieux où les individus qui la composent se seront répandus favorise la corruption des signes convenus pour rendre chaque idée, de là l'origine des langues, qui se seront diversifiées partout."

Review questions

1. **Population:** A fruitfly lays about 1200 eggs, while a tsetse fly only produces about 6-12 larvae. But unlike the fruitfly, the tsetse fly is "viviparous" – meaning its offspring are born almost fully developed. The tsetse larva develops inside the uterus, so that it can, in effect be born as a "teenager." It is often said that vivipary often evolves with a relatively low reproduction rate. Why would this be true?
2. **Mendel and Hardy from a modern perspective:** Suppose you observe that some animals have trait X.
   a. In the naturally occurring population of the animal, would you expect X to be distributed according to Hardy's ratios of some genotypes aa and aA? Explain why not.
   b. How could you tell whether trait X is genetically determined? (Hint: Remember some of the methods used to show that languages abilities are genetically determined at least in part)
3. **Hardy:** Suppose that the relative proportions of genotypes in a population are these:

   $$\text{a } \tfrac{1}{3} \qquad \text{A } \tfrac{2}{3}$$

   How frequent is the dominant phenotype?
4. **The language of RNA.** Use a tree diagram to show the steps in building an RNA molecule with the structural dependencies shown here, using the rules given below.



RNA Parts:	5’	3’	a	u	g	c
	Begin	End	Base	Base	Base	Base
RNA-rule0:	x	$\mapsto$	x			**‘start with any base’**
	Base		Start			
RNA-rule1:	x	y		$\mapsto$	xy	**‘extend to the right’**
	Start	Base			Start	
RNA-rule2:	x	y		$\mapsto$	xy	**‘extend to the left’**
	Base	Start			Start	
RNA-rule3:	x	y	z	$\mapsto$	xyz	**‘add Begin & End’**
	Begin	Start	End		RNA	
RNA-loop:	x	y	z	$\mapsto$	xyz	**<u>if x & z are attracting</u>**
	Base	Start	Base		Start	

5. **Genetic code.** Suppose that instead of 20 amino acids, there were 62.
   a. In this case, how long would a codon have to be to provide a unique code for each amino acid?
   b. Do you think the mechanisms of heredity would work differently in this situation? (explain)
6. **Human language and DNA/RNA.** In the languages of DNA and RNA, there are units of structure (nucleotides, codons, genes). What are the main units of structure in human languages?

			manner	voice	place
1.	[p]	s**p**ot	stop	−voice	labial
2.	[p^h]	**p**op	stop	−voice	aspirated labial
3.	[p']	–	stop	−voice	glottalized labial
4.	[t]	s**t**uck	stop	−voice	alveolar
5.	[t^h]	**t**ick	stop	−voice	aspirated alveolar
6.	[t']	–	stop	−voice	glottalized alveolar
7.	[q]	–	stop	−voice	uvular
8.	[q^h]	–	stop	−voice	aspirated uvular
9.	[q']	–	stop	−voice	glottalized uvular
10.	[k]	s**k**ip	stop	−voice	velar
11.	[k^h]	s**k**ip	stop	−voice	aspirated velar
12.	[k']	s**k**ip	stop	−voice	glottalized velar
13.	[ʧ]	**ch**ip	affricate	−voice	alveopalatal
14.	[ʧ^h]	–	affricate	−voice	aspirated alveopalatal
15.	[ʧ']	–	affricate	−voice	glottalized alveopalatal
16.	[m]	**m**oat	nasal stop	+voice	labial
17.	[n]	**n**ote	nasal stop	+voice	alveolar
18.	[ñ]	–	nasal stop	+voice	palatalized alveolar
19.	[s]	**s**ip	fricative	−voice	alveolar
20.	[ʃ]	**sh**ip	fricative	−voice	alveopalatal
21.	[x]	–	fricative	−voice	velar
18.	[h]	**h**at	fricative	−voice	glottal
22.	[ɾ]	butter	flap	+voice	alveolar
22.	[r]	reef	(central) approximant	+voice	retroflex
21.	[l]	leaf	lateral approximant	+voice	alveolar
21.	[ʎ]	–	lateral approximant	+voice	palatal
24.	[j]	yet	(central) approximant	+voice	palatal
25.	[w]	weird	(central) approximant	+voice	labiovelar

			tongue body height	tongue body backness	lip rounding	tongue root tense (+ATR) or lax (−ATR)
26.	[i]	b**ea**t	high	front	unrounded	+ATR
27.	[u]	b**oo**t	high	back	rounded	+ATR
28.	[o]	r**oa**d	mid	back	rounded	+ATR
29.	[e]	**a**te	mid	front	unrounded	+ATR
30.	[a]	p**o**t	low	back	unrounded	+ATR

7. **Human language structure (Quechua).** One of the main cities in Bolivia is spelled "Cochabamba", but it is really the Quechua word that is phonetically spelled [q^{h}otʃapampa], where [q^{h}otʃa] means 'lake' and [pampa] means 'treeless plain', or 'ground', Draw the syllable structure for [q^{h}otʃapampa].

English Free Morphemes:	(determiner)	D:	the, some, no, a, every, one, two,…
	(names)	DP:	John, Mary, Bill, Sue,…
	(noun)	N:	student, penguin, cat, yard, kid, school, summer, winter, quarter,…
	(verb)	V:	laugh, cry, fall, sing, dance,…
	(transitive verb)	Vt:	like, praise, sing, teach, see,…
	(tense,modal)	T:	will,would,can,could,…
	(adjective)	A:	happy, sad, probable, rare,…
	(adverb)	Adv:	always, sometimes,…
	(preposition)	P:	in, on, with, about, near, by, from, to,…
English Bound Morphemes:	(progressive)	Prog:	(be,-ing)
	(perfective)	Perf:	(have,-en), (have,-ed)
	(tense)	T:	-s, -ed

Syntax, with "Movement":

	x X	↦	x XP	**X to XP, for X=N,A,Adv**	(1)
x D	y NP	↦	xy DP	**D takes NP object**	(2)
x Vt	y DP	↦	x,y VP	**Vt takes DP object (2 parts!)**	(3*a*)
	x V	↦	x,ϵ VP	**V to VP (no object, but still 2 parts)**	(3*b*)
x,y X	z,w VP	↦	x,zyw XP	**if X=Perf,Prog, XP has 2 parts**	(4*a*)
x,y Perf	z,w ProgP	↦	x,zyw PerfP	**PerfP has 2 parts**	(4*b*)
x T	y,z X	↦	x,yz T'	**if T is a <u>word</u> and X=VP,ProgP,PerfP**	(5*a*)
x T	y,z X	↦	ϵ,yxz T'	**if T is a <u>suffix</u> and X=VP**	(5*b*)
x T	y,z X	↦	yx,z T'	**if T is a <u>suffix</u> and X=ProgP,PerfP**	(5*c*)
x DP	y,z T'	↦	xyz TP	**Sentence: build TP as usual**	(6*a*)
x DP	y,z T'	↦	yxz TP	**Y/N Question: if y not empty**	(6*b*)

8. **Human language structure (English).** Using the English grammar above, draw a tree showing the derivation of *John will praise Mary*

9. **Human language structure.** Using the English grammar above, draw a tree showing the derivation of *John be -s praise -ing Mary*

10. **Human language and DNA/RNA.**

    a. The language of RNA has "crossing" and "nested" dependencies. What are they, and why do they occur?

    b. English also has "crossing" and "nested" dependencies. What are they, and why do they occur?

11. Why does Noam Chomsky say that English and other languages have structure-dependent rules? And why is he so interested in this fact?

12. Explain why Chomsky and others think that human language emerged by exaptation. Why does Pinker disagree?

13. Does the Baldwin effect help explain why human languages show so much plasticity? (Explain why or why not)

14. Explain why Pinker and others think that human language emerged by selection. Why does Chomsky disagree?

15. In *The Major Transitions in Evolution*, John Maynard Smith and Eörs Szathmáry proposed that there have been 8 major transitions:

    1. from replicating molecules to protocells
    2. from independent replicators to chromosomes
    3. from RNA as gene and enzyme to DNA genes and protein enzymes
    4. from bacterial cells to cells with nuclei and organelles
    5. from single-celled to multi-celled organisms
    6. from non-sexual to sexual reproduction
    7. from solitary individuals to colonies with non-reproductive castes (termites, ants, bees)
    8. from primate to human societies, possible only with the advent of language abilities

    In an interview about this work in *Natural History*, Smith said

    > These transitions are all concerned with the storage and passage of information. It's fascinating, for instance, to try to model the evolution of the genes together with the evolution of human language – not identical events but at bottom very similar.

    What are the most significant similarities between the evolution of genes (steps 1-3) and the evolution of (not particular human languages but) human language abilities (step 8)?[7]

[7] If you do not include something like Smith's "storage and passage of information" in your list of significant similarities, explain why you didn't, and if you do include it, explain how the information is stored and what it is information about.

Glossary

(some terms introduced in the text are not here: use the index too)

affix Bound morphemes, morphemes that can only occur together with others, are called affixes. Familiar prefixes and suffixes are affixes, but some languages seem also to have infixes – morphemes that go into the middle of another word – and circumfixes – morphemes that have a prefix and a suffix part.

AIDS Acquired immune deficiency syndrome, caused by HIV, and transmitted by bodily fluids such as blood or semen.

allele An allele is any one of a number of alternative forms of the same gene occupying a given locus (position) on a chromosome.

amino acids A molecule that contains an amino and carbolic acid. 20 of these molecules compose the proteins in every living thing from bacteria to humans.

angstrom 10^{-10} meter

A-P axis Anterior-posterior axis, important in early development of eukaryotes.

archaebacteria This variety of prokaryotes was discovered in the 1970's by Woese on the basis of genetic distinctions. Some are single-celled, while others form filaments or aggregates. Most live in high-temperature, anaerobic, hypersaline environments.

arithmetic progression A sequence of numbers where each one is k more than the previous one. For example, the sequence $3, 6, 9, 12, 15, \ldots$ is arithmetic because each number is 3 more than the previous one. Contrast *geometric progression*.

AZT Azidothymidine, an anti-retroviral drug. Developed initially for cancer, this was the first antiviral treatment to be approved for use against HIV. It blocks the enzyme that HIV uses to replicate its RNA for splicing into the DNA of a target cell.

base In organic chemistry, the nitrogenous bases are the building blocks of DNA and RNA. There are five of them: Adenine, Thymine, Guanine, Cytosine, and Uracil. Uracil is found only in RNA, in place of Thymine in DNA.

bovine Of or related to cattle or oxen.

catalyst A chemical that facilitates the occurrence of some reaction.

cephalopod A subclass of *Mollusca*, typically with bilateral body symmetry, a prominent head, and a modification of the mollusc foot into tentacles. It contains the octopus, squid, cuttlefish and the shelled nautilus.

chromosome A chromosome is a large contiguous DNA strand that contains many genes. In eukaryotes, the DNA exists inside the nucleus. Just prior to cell division the chromosomes coalesce and become visible with an optical microscope. They were discovered by Karl von Nägeli in 1842.

clade A group of organisms with a common ancestor.

codon A sequence of 3 bases in RNA that specifies an amino acid.

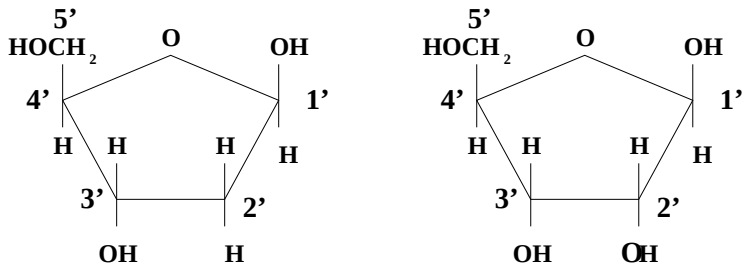
complementizer A morpheme like *that* in the sentence *She knows that grass is green,* or like *whether* in *She wonders whether grass is green.* A complementizer is a morpheme that introduces an embedded sentence or sentence-like phrase.

diploid Cells that have two copies of their genome: two copies of each chromosome. In higher organisms, most cells are diploid except for reproductive cells which are haploid.

DNA Deoxyribonucleic acid, the molecule that controls the synthesis of proteins in living things (except viruses). In 1953, Watson and Crick proved that genes were composed of DNA. The structure is double-stranded: two nucleotide polymers pair up, matching nucleotide to nucleotide along their length.

deme a group or aggregate, typically related.

deoxyribose In DNA, a nucleotide consists of a base, a 5-carbon sugar molecule called deoxyribose, and a phosphate group. The carbon atoms in deoxyribose are labeled with "primes" according to their position in the structure: 1', 2',...5'. In RNA, a very similar molecule called ribose (with one more extra oxygen atom) plays a similar role:



deoxyribose **ribose**

Down syndrome A chromosomal abnormality in humans that typically causes mental impairment.

ecology The study of relations between organisms and their environments.

ectopical Proceeding in an abnormal locality.

epidemiology The branch of medicine that studies disease and epidemics in populations.

epistasis The name for an interaction between genes at different loci, where the individual contributions of each gene in the resulting phenotype are not recognizable: for example, when one determines a trait that conceals the effects of the other.

ethology In biology, the study of animal behavior in their natural environments.

eubacteria Most bacteria are eubacteria. Some bacteria were discovered to be genetically distinct in the 1970's, and are now called archaebacteria.

eukaryote Organisms with cells that contain a well-defined nucleus.

eukaryote crown group The rapid diversification of eukaryotes that we see in the phylogenetic tree on page 3 is sometimes called a "crown:" a cluster of relatively closely related organisms.

exaptation A term invented by (Gould and Vrba, 1982) for features not specifically selected for ("spandrels"), or features previously designed for another function, which have been coopted for their current use.

fitness In evolutionary biology and artificial life: average number of offspring per individual, or closely related quantity.

gamete A reproductive cell.

gene In classical genetics, genes are the basic units of material passed to offspring during reproduction: an atom of heredity. In 1910, Thomas Hunt Morgan proved chromosomes to be the carriers of genes, and in recent genetics, the term "gene" is applied to segments of DNA that are transcribed into RNA and translated into proteins. (In many species, genes are separated by long sequences of so-called "junk DNA" which does not encode proteins.)

genome The complete set of genes in an organism.

genotype An organism's genetic endowment, its genome.

geometric progression A sequence of numbers where each one is k times the previous one. For example, the sequence $3, 6, 12, 24, 48, \ldots$ is a geometric progression because each number is 2 times the previous one. Contrast *arithmetic progression*.

haploid Cells that have just one copy of their genome: one copy of each chromosome. In higher organisms, most cells are diploid except for reproductive cells which are haploid.

haplotype means 'haploid genotype'. Even with diploid organisms, it often makes sense to look at just half the whole genome. Humans have a 23 chromosome haplotype.

HIV human immunodeficiency virus. This retrovirus attacks the human immune system and can cause AIDS.

homeodomain A distinctive structural domain with a DNA-binding function in proteins involved in the regulation of transcription and translation.

homology In biology, traits of two organisms, or segments of DNA sequences of two organisms, are said to be homologous if they are an inheritance from some common ancestor.

homophone, homophony Two words or expressions of a language that sound the same are called homophonous. For example, the noun *bank* meaning a river edge and the verb *bank* meaning to conduct financial affairs are homophones. And the expressions there, their, they're are homophones.

homoplasy Similarity of structure produced independently by the operation of similar external circumstances.

homozygote A cell or other organism with two copies of the same allele (e.g. AA or aa in the example of §1.3)

heterozygote A cell or other organism with two copies of two different alleles (e.g. aA in the example of §1.3)

hybrid The offspring of two different varieties of a species.

meiosis Chromosomal and nuclear division where each new nucleus receives just half of the genome, for example in the formation of reproductive cells of higher organisms.

messenger RNA, mRNA Transcribed from DNA, carries instructions out of the nucleus to other places in the cell.

metazoa The phylum of multicellular animals, from jellyfish to humans.

missense mutation A point mutation that changes the resulting amino acid in a polypeptide. This kind of mutation changes the "meaning" of the sequence. In contrast, a "nonsense" mutation results in a sequence that truncates protein transcription or is otherwise unreadable.

mitochondria In the cytoplasm of eukaryotic cells, these lipid and protein membrane enclosed complexes produce ATP (adenosine triphosphate) from fatty acids and glucose. The high-energy bonds of ATP provide the energy for most growth and metabolism.

mitosis Chromosomal and nuclear division in eukaryotic cells.

morphology In biology, the study of the forms, the shapes of organisms. In linguistics, the study of word formation.

nanometer a billionth (10^{-9}) of a meter

nanomachine, nanobot a molecular-sized machine. DNA, protein molecules, ribosomes, and mitochondria are sometimes described as natural nanomachines. Chemists and engineers are now building artificial ones too.

nonsense mutation a mutation resulting in protein truncation or other sorts of unreadability. Contrasts with **missense mutation**.

nucleotide A certain kind of organic molecule, consisting of a phosphate, a sugar (deoxyribose in DNA and ribose in RNA), and one of the bases Adenine, Thymine (or Uracil in RNA), Guanine, and Cytosine.

paleontology The study of what fossils tell us about the organisms and ecology.

panmictic population A population with random mating, approximating conditions of the Hardy-Weinberg law.

peptide A molecule formed from linking amino acids

phage, or bacteriophage A virus-like particle infecting prokaryotes.

phenotype the properties of an organism which may be determined in part by the organism's genetic endowment, its genotype

phonetics The study of the basic vocal gestures of human languages.

phonology The study of the how sound patterns are organized in human languages.

phrase structure rule A principle that governs how phrases are assembled from words.

phyllotaxis The arrangement or order of leaves or other parts (e.g. scales of a pine-cone, florets of a cauliflower, etc.) upon an axis or stem.

phylogeny The origin and evolution of species

phylum Group of evolutionarily related organisms.

picogram a trillionth (10^{-12}) of a gram

polyhedron A solid made up of flat faces, like a cube or a pyramid. A sphere or any other solid with a curved face is not a polyhedron.

polymer A molecule built from a large number of similar units bonded together

polypeptide A single chain of amino acids. Proteins are long polypeptide chains.

primordium An organ in its earliest stages of development.

prion infectious self-reproducing protein structures. They are implicated in "mad cow disease" (Bovine Spongiform Encephalopathy)

prokaryote Single celled organisms lacking a well-defined nucleus. In the late 1970s, Carl Woese at the University of Illinois discovered that the prokaryotes fall into two genetically different classes: eubacteria and archaea.

protein A complex of one or more polypeptides, more than about 50 amino acids long. These chains fold up into distinctive shapes. Proteins control many of the fundamental cellular processes.

purine The bases Adenine and Guanine are purines.

pyrimidine The bases Cytosine, Uracil and Thymine are pyrimidines.

Quechua Quechua is the language of the Incas, now spoken by approximately 7 million people in South America.

retrovirus, or RNA virus A virus with a genome composed of RNA.

ribose In RNA, a nucleotide consists of a base, a 5-carbon sugar molecule called ribose, and a phosphate group. See deoxyribose.

ribosome a complex of RNA and proteins found in the fluid of all cells that translates mRNA into a protein

RNA Ribonucleic acid. A nucleic acid similar to DNA, but less stable. Except in viruses, it is almost always appears in a single strand. In retroviruses, RNA is the only genetic material.

semantics The study of the meaning of linguistic expressions.

spandrel In a rectangular structure that is supported by an arch, the space outside the arch, between the arch and the corner of the rectangular structure, is a spandrel. Gould uses these as an example of "nonadaptive sequelae of prior structural decisions" (Gould, 2002, p43). See *exaptation*.

speciation The emergence of new species.

species There are a number of different ideas about how this should be defined. Probably the most familiar idea is that two organisms are of the same species if they are sufficiently similar – but the question of just how similar they must be can be debated. Another idea is that two organisms are of the same species if they can mate successfully, but this idea only applies if the organisms reproduce sexually, and the criterion of success can be debated (successful every time? in nature, or in a lab with extensive medical support?...). Yet another idea, sometimes added to one of the previous proposals, is that two organisms of the same species must have a common ancestor.

syntax The study of patterns of elements. In linguistics: how words are combined to form intelligible phrases of various kinds.

vestigial Remaining or surviving in a reduced or degenerate form.

viroid Infectious, single stranded RNA, occurring in higher plants, as in potatoes (e.g. the potato spindle tuber viroid), and can cause deformation.

virulence In biology and medicine, the property of a pathogen that it damages its host.

virus In biology, a virus is a particle which infects eukaryotes, typically composed of either DNA or RNA surrounded by some form of protective "coat" consisting of protein, or protein and lipid. They rely on enzymes of their host cells to reproduce.
In computer science, a virus is a program that attaches itself to another program and relies on that other program to reproduce.

wild type A naturally occurring genotype, in contrast to an induced mutation or artificially cross-bred genotype.

zygote A cell resulting from the union of two reproductive cells.

References

Adleman, Leonard M. 1994. Molecular computations of solutions to combinatorial problems. Science, 266(5187):1021–1024.

Anderson, J.R., M. Montant, and D. Schmitt. 1996. Rhesus monkeys fail to use gaze direction as an experimenter-given cue in an object choice task. Behavioural Processes, 37:47–55.

Au, Terry Kit-Fong, Leah M. Knightly, Sun-Ah Jun, and Janet S. Oh. 2001. Overhearing a language during childhood. Psychological Science, 13(3):238–243.

Baayen, Harald. 2001. Word Frequency Distributions. Kluwer, NY.

Baker, Mark C. 2001. The Atoms of Language. Basic Books, NY.

Baldwin, J. Mark. 1896. A new factor in evolution. American Naturalist, 30:441–451, 536–553.

Barbieri, Marcello. 2003. The Organic Codes: An Introduction to Semantic Biology. Cambridge University Press, Cambridge.

Berwick, Robert C. 1997. Feeling for the organism: Are living things nothing more than the sum of their gradually-evolved parts? Boston Review, 21(6):23–27.

Berwick, Robert C. and Amy S. Weinberg. 1984. The Grammatical Basis of Linguistic Performance: Language Use and Acquisition. MIT Press, Cambridge, Massachusetts.

Bloom, Paul and L. Markson. 1998. Capacities underlying word learning. Trends in Cognitive Science, 2:67–73.

Bloomfield, Leonard. 1933. Language. University of Chicago Press, Chicago.

Boyd, Robert and Peter J. Richerson. 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.

Boyd, Robert and Joan B. Silk. 2000. How Humans Evolved. Norton, NY.

Bradbury, Jack W. and Sandra L. Vehrencamp. 1998. Principles of Animal Communication. Sinauer, Sunderland, Massachusetts.

Breen, Tricia L., Joe Tien, Scott R.J. Oliver, Tanja Hadzik, and George M. Whitesides. 1999. Design and self-assembly of open, regular, 3D mesostructures. Science, 284:948–951.

Brent, Michael R. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. Machine Learning, 34:71–105.

Brody, Michael. 2000. Mirror theory: syntactic representation in perfect syntax. Linguistic Inquiry, 31:29–56.

Brown, Roger. 1973. A First Language. Harvard University Press, Cambridge, Massachusetts.

Byrnes, W. Malcolm, Wenjue Hu, Ezzat S. Younathan, and Simon H. Chang. 1995. A chimeric bacterial phosphofructokinase exhibits cooperativity in the absence of heterotropic regulation. Journal of Biological Chemistry, 270:3828–3835.

Calvin, William H. and Derek Bickerton. 2000. Lingua ex Machina : Reconciling Darwin and Chomsky with the Human Brain. MIT Press, Cambridge, Massachusetts.

Carey, Susan. 1977. The child as word learner. In M. Halle, J. Bresnan, and G.A. Miller, editors, Linguistic Theory and Psychological Reality. MIT Press, Cambridge, Massachusetts.

Cartwright, John. 2000. Evolution and Human Behavior. MIT Press, Cambridge, Massachusetts.

Casado, Concepción, Soledad García, Carmen Rodríguez, Jorge del Romero, Gonzalo Bello, and Cecilio López-Galíndez. 2001. Different evolutionary patterns are found within human immunodeficiency virus type 1-infected patients. Journal of General Virology, 82:2495–2508.

Cassares, F. and R.S. Mann. 1998. Control of antennal vs. leg development in Drosophila. Nature, 392:723–726.

Cavalli-Sforza, L.L. 1997. Genes, peoples, languages. Proceedings of the National Academy of Sciences of the United States of America, 94:7719–7724.

Chalfie, M., Y. Tu, G. Euskirchen, W. Ward, and D. Prasher. 1994. Green fluorescent protein as a marker for gene expression. Science, 263:802–805.

Chambers, Craig G., Michael K. Tanenhaus, Kathleen M. Eberhard, Hana Filip, and Greg N. Carlson. 2002. Circumscribing referential domains during real-time language comprehension. Journal of Memory and Language, 47:30–49.

Chang, Catherine C.Y., Chi-Yu Gregory Lee, Ellen T. Chang, Jonathan C. Cruz, Marc C. Levesque, and Ta-Yuan Chang. 1998. Recombinant acyl-coa:cholesterol acyltransferase-1 (acat-1) purified to essential homogeneity utilizes cholesterol in mixed micelles or in vesicles in a highly cooperative manner. Journal of Biological Chemistry, 273:35132–35141.

Cheney, Dorothy L. and Robert M. Seyfarth. 1990. The representation of social relations by monkeys. Cognition, 37(1-2):167–196.

Cheney, Dorothy L. and Robert M. Seyfarth. 1992. How Monkeys See the World. University of Chicago Press, Chicago.

Cheney, Dorothy L. and Robert M. Seyfarth. 1996. How Monkeys See the World. University of Chicago Press, Chicago.

Cheney, Dorothy L. and Robert M. Seyfarth. 1997. Reconciliatory grunts by dominant female baboons influence victims' behavior. Animal Behaviour, 54:409–418.

Cheney, Dorothy L. and Robert M. Seyfarth. 2005. Constraints and preadaptations in the earliest stages of language evolution. Linguistic Review, 22:135–159.

Cheney, Dorothy L., Robert M. Seyfarth, and Ryne Palombit. 1996. The function and mechanisms underlying baboon contact barks. Animal Behaviour, 52:507–518.

Cheng, Ken. 1986. A purely geometric module in the rat's spatial representation. Cognition, 23:149–178.

Chomsky, Noam. 1956. Three models for the description of language. IRE Transactions on Information Theory, IT-2:113–124.

Chomsky, Noam. 1971. Problems of Knowledge and Freedom: The Russell Lectures. Vintage, NY.

Chomsky, Noam. 1975. Reflections on Language. Pantheon, NY.

Chomsky, Noam. 1986. Knowledge of Language. Praeger, NY.

Chomsky, Noam. 1991. Universal grammar (letter). New York Review of Books, 38(21).

Clark, Eve V. 1993. The Lexicon in Acquisition. Cambridge University Press, Cambridge.

Clark, Ross. 1990. Austronesian languages. In Bernard Comrie, editor, The World's Major Languages. Oxford University Press, Oxford.

Cleland, Carol E. and Christopher F. Chyba. 2002. Defining 'life'. Origins of Life and Evolution of the Biosphere, 32:387–393.

Clowes, M.B. 1971. On seeing things. Artificial Intelligence, 2:79–116.

Cope, Edward D. 1987. Origin of the Fittest: Essays on Evolution and the Primary Factors of Organic Evolution. Ayer, NY.

Cormen, Thomas H., Charles E. Leiserson, and Ronald L. Rivest. 1991. Introduction to Algorithms. MIT Press, Cambridge, Massachusetts.

Crockford, C. and C. Boesch. 2003. Context-specific calls in wild chimpanzees, pan troglodytes verus: Analysis of barks. Animal Behaviour, 66:115–125.

Darwin, Charles. 1859. On The Origin of Species by Means of Natural Selection, or The Preservation of Favoured Races in the Struggle for Life. Modern Library, NY.

Davis, R.L. 1996. Physiology and biochemistry of *drosophila* learning mutants. Physiological Reviews, 76:299–317.

Dawkins, Richard. 1976. The Selfish Gene. Oxford University Press, Oxford.

Dawkins, Richard. 1986. The Blind Watchmaker. Norton, NY.

Deacon, Terrence W. 1997. The Symbolic Species: The Co-Evolution of Language and the Brain. Norton, NY.

Demuth, Katherine. 1986. Prompting routines in the language socialization of Basotho children. In B.B. Schieffelin and E. Ochs, editors, Language Socialization across Cultures. Cambridge University Press, Cambridge, pages 51–79.

Dermitzakis, Emmanouil T., Alexandre Reymond, Robert Lyle, Nathalie Scamuffa, Catherine Ucla, Samuel Deutsch, Brian J. Stevenson, Volker Flegel, Philipp Bucher, C. Victor Jongeneel, and Stylianos E. Antonarakis. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome. Nature, 420:578–581.

Donahue, Kathleen and Johanna Schmitt. 1999. The genetic architecture of plasticity to density in *impatiens capensis*. Evolution, 53(5):1377–1386.

Douady, S. and Y. Couder. 1996. Phyllotaxis as a dynamical self organizing process (parts i, ii, iii). Journal of Theoretical Biology, 178(3):255–312.

Drake, John W., Brian Charlesworth, Deborah Charlesworth, and James F. Crow. 1998. Rates of spontaneous mutation. Genetics, 148:1667–1686.

Dretske, Fred. 1981. Knowledge and the Flow of Information. Cambridge University Press, Cambridge.

Dretske, Fred. 1983. Précis of knowledge and the flow of information. Behavioral and Brain Sciences, 6(1):55–90.

Duda, Richard O., Peter E. Hart, and David G. Stork. 2001. Pattern Classification. Wiley, NY.

Dudley, Andrew T., María A. Ros, and Clifford J. Tabin. 2002. A re-examination of proximodistal patterning during vertebrate limb development. Nature, 418:539–544.

Durrett, Rick. 2002. Probability Models of DNA Sequence Evolution. Springer, NY.

Dyer, Fred C. 2002. The biology of the dance language. Annual Reviews Entomology, 47:917–949.

Dyson, Freeman. 1979. Disturbing the Universe. Harper, NY.

Egholm, M., O. Buchardt, L. Christensen, P.E. Nielson, and R.H. Berg. 1992. Peptide nucleic acids (pna). Journal of the American Chemical Society, 114:1895–1897.

Eigen, Manfred. 1992. Steps toward Life: A Perspective on Evolution. Oxford University Press, Oxford.

Embick, David, Alec Marantz, Masushi Miyashita, Wayne O'Neil, and Kuniyoshi L. Sakai. 2001. A syntactic specialization for broca's area. Proceedings of the National Academy of Sciences, 97:6150–6154.

Enard, Wolfgang, Molly Przeworski, Simon E. Fisher, Cecilia S.L. Lai, Victor Wiebe, Takashi Kitano, Anthony P. Monaco, and Svante Pääbo. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. Nature, 418:869–872.

Engh, Anne L., Rebekah R. Hoffmeier, Dorothy L. Cheney, and Robert M. Seyfarth. 2006. Who me? Can baboons infer the target of vocalizations? Animal Behaviour, 71(2):381–387.

Ervin, S. 1964. Imitation and structural change in children's language. In E. Lenneberg, editor, New Directions in the Study of Language. MIT Press, Cambridge, Massachusetts.

Eyre-Walker, Adam and Peter D. Keightly. 1999. High genomic deleterious mutation rate in hominids. Nature, 397:344–347.

Fischer, Julia, Markus Metz, Dorothy L. Cheney, and Robert M. Seyfarth. 2001. Baboon responses to graded bark variants. Animal Behaviour, 61:925–931.

Fisher, R.A. 1934. Adaptation and mutations. School Science Review, 15:294–301.

Flemming, Edward. 2002. Contrast and perceptual distinctiveness. In B. Hayes, R. Kirchner, and D. Steriade, editors, The Phonetic Bases of Markedness. Cambridge University Press, NY.

Fodor, J.A., M.F. Garrett, E.C.T. Walker, and C.H. Parkes. 1980. Against definitions. Cognition, 8:263–367.

Frege, Gottlob. 1923. Gedankengefüge. Beträge zur Philosophie des deutschen Idealismus, 3:36–51. Translated and reprinted as 'Compound thoughts' in *Mind* **72**(285): 1-17, 1963.

French, R. and A. Messinger. 1994. Genes, phenes and the Baldwin effect. In Rodney Brooks and Patricia Maes, editors, Artificial Life IV. MIT Press, Cambridge, Massachusetts.

Fukson, Olga I., Michael B. Berkinblit, and Anatoly G. Feldman. 1980. The spinal frog takes into account the scheme of its body during the wiping reflex. Science, 209:1261–1263.

Gallistel, C.R. 1990. The Organization of Learning. MIT Press, Cambridge, Massachusetts.

Gallistel, C.R., Rochel Gelman, and Sara Cordes. 2003. The cultural and evolutionary history of the real numbers. In S. Levinson and P. Jaisson, editors, Culture and Evolution. MIT Press, Cambridge, Massachusetts. forthcoming.

Gehring, W.J. 1996. The master control gene for morphogenesis and evolution of the eye. Genes to Cells, 1:11–15.

Gifford, David K. 1994. On the path to computation with DNA. Science, 266:993–994.

Godwin, John, J. Adam Luckenbach, and Russell J. Borski. 2003. Ecology meets endocrinology: environmental sex determination in fishes. Evolution and Development, 5:40–49.

Gold, E. Mark. 1967. Language identification in the limit. Information and Control, 10:447–474.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. Computational Linguistics, 27(2):153–198.

Gopnik, Myrna. 1990. Feature-blind grammar and dysphasia. Nature, 344:715.

Gopnik, Myrna. 1992. Talking genes (letter). New York Review of Books, 39(7).

Gould, James L. 1986. The locale map of honeybees: Do insects have cognitive maps? Science, 232(4752):861–863.

Gould, James L. and Carol Grant Gould. 1988. The Honey Bee. Freeman, NY.

Gould, Stephen Jay. 2002. The Structure of Evolutionary Theory. Harvard University Press, Cambridge, Massachusetts.

Gould, Stephen Jay and Elisabeth S. Vrba. 1982. Exaptation – a missing term in the science of form. Paleobiology, 8(1):4–15.

Grant, P.R. and B.R. Grant. 2002. Unpredictable evolution in a 30-year study of Darwin's finches. Science, 296:707–711.

Grant, P.R. and B.R. Grant. 2006. Evolution of character displacement in Darwin's finches. Science, 313:224–226.

Griffiths, R. 1970. The abilities of young children. Technical report, Child Development Research Centre, London.

Gunter, Chris and Ritu Dhand. 2002. Human biology by proxy. Nature, 420:509.

Guntrip, J. and R.M. Sibly. 1998. Phenotypic plasticity, genotype-by-environment interaction and the analysis of generalism and specializatin in *callosobruchus maculatus*. Heredity, 81:198–204.

Guo, Ximing, , Dennis Hedgecock, William K. Hershberger, Kenneth Cooper, and Standish K. Allen. 1998. Genetic determinants of protandric sex in the Pacific Oyster, *crassostrea gigas thunberg*. Evolution, 52(2):394–402.

Haas, J., E.C. Park, and B. Seed. 1996. Codon usage limitation in the expression of hiv-1 envelope glycoprotein. Current Biology, 6:315–324.

Halder, Georg, Patrick Callaerts, and Walter J. Gehring. 1995. Induction of ectopic eyes by targeted expression of the eyeless gene in drosophila. Science, 267:1788–1792.

Hamblin, Martha T., Emma E. Thompson, and Anna Di Rienzo. 2002. Complex signatures of natural selection at the Duffy blood group locus. American Journal of Human Genetics, 70:369–383.

Hardy, Godfrey H. 1908. Mendelian proportions in a mixed population. Science, 28:49–50. www.esp.org/foundations/genetics/classical/hardy.pdf.

Harkema, Henk. 2000. A recognizer for minimalist grammars. In Sixth International Workshop on Parsing Technologies, IWPT'00.

Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie, editors. 2005. The World Atlas of Language Structures. Oxford University Press, Oxford.

Hastie, T., R. Tibshirani, and J.H. Friedman. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer, NY.

Hauser, Marc D. 1989. Ontogenetic changes in the comprehension and production of vervet monkey *cercopithecus aethiops* vocalizations. Journal of Comparative Psychology, 103:149–158.

Hauser, Marc D. 2000. The Evolution of Communication. MIT Press, Cambridge, Massachusetts.

Hauser, Marc D., Noam Chomsky, and W. Tecumseh Fitch. 2002. The faculty of language: what it is, who has it, and how did it evolve? Science, 298:1569–1579.

Hauser, Marc D., Elissa L. Newport, and Richard N. Aslin. 2001. Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. Cognition, 78(1):B53–B64.

Hauser, Marc D. and Bailey Spaulding. 2006. Wild rhesus monkeys generate causal inferences about possible and impossible physical transformations in the absence of experience. Proceedings of the National Academy of Sciences, 103(18):7181–7185.

Hawkins, John A. 1990. A parsing theory of word order universals. Linguistic Inquiry, 21:223–262.

Hawkins, John A. 1994. A Performance Theory of Order and Constituency. Cambridge University Press, NY.

Hawkins, John A. 1999. Processing complexity and filler-gap dependencies across grammars. Language, 75:244–285.

Hawkins, John A. 2001. Why are categories adjacent? Journal of Linguistics, 27:1–34.

Heim, R., A.B. Cubitt, and R.Y. Tsein. 1995. Improved green fluorescence. Nature, 373:663–664.

Herrero, Joaquín and Federico Sánchez de Lozada. 1978. Gramatica Quechua: Estructura del Quechua Boliviano Contemporaneo. Editorial Universo, Cochabamba, Bolivia.

Hinton, G.E. and S.J. Nolan. 1987. How learning can guide evolution. Complex Systems, 1:495–502.

Horwich, Paul. 1998. Meaning. Oxford University Press, Oxford.

Huber, Sarah K. and Jeffrey Podos. 2006. Beak morphology and song features covary in a population of Darwin's finches (Geospiza fortis). Biological Journal of the Linnean Society, 88(3):489–498.

Huffman, D.A. 1971. Impossible objects as nonsense sentences. Machine Intelligence, 6:295–323.

Hull, David L. 1980. Individuality and selection. Annual Review of Ecology and Systematics, 11:311–332,1141–1144.

Hull, David L. 1994. Darwin and his Critics: The Reception of Darwin's Theory of Evolution by the Scientific Community. University of Chicago Press, Chicago.

Hurst, J.A., M. Baraitser, E. Auger, F. Graham, and S. Norell. 1990. An extended family with a dominantly inherited speech disorder. Developmental Medecine and Child Neurology, 32:347–355.

Hutchinson, Janis Faye. 2001. The biology and evolution of HIV. Annual Review of Anthropology, 30:85–108.

Indefrey, P., C.M. Brown, P. Hagoort, H. Herzog, and R. Seitz. 2001. Left prefrontal cortex processes syntax independent of lexical meaning. NeuroImage, 12.

Jablonsky, D. 1987. Heritability at the species level: analysis of geographic ranges of Cretaceous mollusks. Science, 238:360–363.

Jain, Sanjay, Daniel Osherson, James S. Royer, and Arun Sharma. 1999. Systems that Learn: An Introduction to Learning Theory (second edition). MIT Press, Cambridge, Massachusetts.

Janzen, D.H. 1971. Euglossine bees as long-distance pollinators of tropical plants. Science, 171(3967):203–205.

Jorde, Lynn B., Michael Bamshad, and Alan R. Rogers. 1998. Using mitochondrial and nuclear DNA markers to reconstruct human evolution. BioEssays, 20:126–136.

Joseph, Brian. 2000. Historical linguistics. In Mark Aronoff and Janie Rees-Miller, editors, The Handbook of Linguistics. Blackwell, Oxford.

Joshi, Aravind K. 2002. Some new directions for applications of NLP techniques for modeling biological sequences. Presentation, Utrecht University.

Josselyn, Sheena A., Satoshi Kida, Sandra Peña de Ortiz, and Alcino J. Silva. 2003. CREB, plasticity and memory. In L. Kaczmarek and H.J. Robertson, editors, Handbook of Chemical Neuroanatomy, Vol.19: Immediate Early Genes and Inducible Transcription Factors in Mapping of the Central Nervous System Function and Dysfunction. Elsevier, NY.

Joyce, G.F., A. Schwartz, S.L. Miller, and L. Orgel. 1997. The case for an ancestral genetic system involving simple analogs of nucleotides. Proceedings of the National Academy of Sciences, 84:4398–4402.

Kaminen, N., K. Hannula-Jouppi, M. Kestilä, P. Lahermo, K. Muller, M. Kaaranen, B. Myllyluoma, A. Voutilainen, H. Lyytinen, J. Nopola-Hemmi, and J. Kere. 2003. A genome scan for developmental dyslexia confirms linkage to chromosome 2p11 and suggests a new locus on 7q32. Journal of Medical Genetics, 40(5):340–345.

Kauffman, Stuart A. 1993. The Origins of Order. Oxford University Press, NY.

Kauffman, Stuart A. 1995. At Home in the Universe. Oxford University Press, NY.

Kay, P., B. Berlin, L. Maffi, and W. Merrifield. 1997. Color naming across languages. In C.L. Hardin and L. Maffi, editors, Color Categories in Thought and Language. Cambridge University Press, NYp.

Kearns, Michael J. and Umesh V. Vazirani. 1994. An Introduction to Computational Learning Theory. MIT Press, Cambridge, Massachusetts.

Keenan, Edward L. 1997. The historical development of the English anaphora system. UCLA manuscript, forthcoming.

Kendall, Bruce E. 1991. Chaos and cycles. In Harold A. Mooney and Josep G. Canadell, editors, The Earth System: Biological and Ecological Dimensions of Global Environmental Change, Encyclopedia of Global Environmental Change. Wiley, NY, pages 209–215.

Kettlewell, H.B.D. 1958. Industrial melanism in the Lepidoptera and its contribution to our knowledge of evolution. In Proceedings of the 10th International Congress on Entomology, volume 2, pages 831–841.

Kim, Chul-Hyun, C. Cheng Kao, and Ignacio Tinoco. 2000. RNA motifs that determine specificity between a viral replicase and its promoter. Nature Structural Biology, 7:415–423.

Kimura, Motoo. 1985. The neutral theory of molecular evolution. New Scientist, pages 41–46.

Kirby, Simon. 1999a. Function, Selection and Innateness. Oxford University Press, Oxford.

Kirby, Simon. 1999b. Learning, bottlenecks, and the evolution of recursive syntax. In E.J. Briscoe, editor, Linguistic Evolution Through Language Acquisition: Formal and Computational Models, Cambridge. Cambridge University Press.

Kirousis, Lefteris M. 1990. Effectively labeling planar projections of polyhedra. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(2):123–130.

Kirousis, Lefteris M. and Christos H. Papadimitriou. 1988. The complexity of recognizing polyhedral scenes. Journal of Computer and System Sciences, 37:14–38.

Klavins, Eric. 2002. Automatic synthesis of controllers for assembly and formation forming. In International Conference on Robotics and Automation.

Kobele, Gregory M. 2002. Formalizing mirror theory. Grammars, 5:177–221.

Komarova, Natalia L. and Martin A. Nowak. 2001. Evolutionary dynamics of the lexical matrix. Bulletin of Mathematical Biology, 63(3):451–485.

Korb, Judith and Karl E. Linsenmair. 2000. Ventilation of termite mounds: new results require a new model. Behavioural Ecology, 11:486–494.

Kreitman, Martin. 2002. Methods to detect selection in populations with applications to the human. Annual Review of Genomics and Human Genetics, 1:539–559.

Kroch, Anthony. 1987. Function and grammar in the history of English: Periphrastic *do*. In R.W. Fasold and D. Schiffrin, editors, Language Change and Variation. John Benjamins, pages 133–172.

Lackner, J.A. 1968. A developmental study of language behavior in retarded children. Neuropsychologia, 6:301–320.

Lai, Cecilia S.L., Simon E. Fisher, Jane A. Hurst, Faraneh Vargha-Khadem, and Anthony P. Monaco. 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature, 413:519–523.

Lamarck, Jean-Baptiste. 1809. Philosophie zoologique ou exposition des considérations relatives à l'histoire naturelle des animaux; à la diversité de leur organisation et des facultés qu'ils en obtiennent; aux causes physiques qui maintiennent en eux la vie et donnent lieu aux mouvements qu'ils exécutent; enfin, à celles qui produisent, les unes le sentiment et les autres l'intelligence de ceux qui en sont doués. Dentu, Paris. Translated as *Philosophical Zoology, or An Exposition of the Considerations Concerning the Natural History of Animals, the diversity of their organic structure and the faculties which they derive from it, the physical causes which maintain life in them and give rise to movements which they carry out, finally, those causes which produce, in some animals feeling, and in others the intelligence of those who are endowed with it.*

Lamarck, Jean-Baptiste. 1815. Histoire naturelle des animaux sans vertèbres. Verdiere, Paris.

Lenneberg, E.H., I.A. Nichols, and E.F. Rosenberger. 1964. Primite stages of language development in mongolism. In D.M. Rioch, L.C. Kolb, and J. Ruesch, editors, Disorders of Communication, Volume XLII: Research Publications of the Association for Research in Nervous and Mental Disease, Vol. XLII. Williams and Wilkins, Baltimore, Maryland.

Leung, Siu-wai, Chris Mellish, and Dave Robertson. 2001. Basic gene grammars and DNA-chart parser for language processing of Escherichia coli promoter DNA sequences. Bioinformatics, 17:226–236.

Lewontin, Richard. 2002. Directions in evolutionary biology. Annual Review of Genetics, 36(1):1–18.

Lewontin, Richard. 2003. Science and simplicity. New York Review of Books, 50(7):39–42.

Lewontin, Richard C. 1983. The organism as subject and object of evolution. Scientia, 118:65–82.

Lewontin, Richard C. 1998. The evolution of cognition: Questions we will never answer. In D. Scarborough and S. Sternberg, editors, An invitation to cognitive science, Volume 4: Methods, models, and conceptual issues. MIT Press, Cambridge, MA.

Li, Peggy and Lila Gleitman. 2002. Turning the tables: language and spatial reasoning. Cognition, 83(3):265–294.

Li, Wentian. 1992. Random texts exhibit Zipf's law-like word frequency distribution. IEEE Transactions on Information Theory, 38:1842–1845.

Lieberman, Philip. 2000. Human Language and Our Reptilian Brain. Harvard University Press, Cambridge, Massachusetts.

Lightfoot, David. 1982. The language lottery: Toward a biology of grammars. MIT Press, Cambridge, Massachusetts.

Lightfoot, David. 1999. The Development of Language: Acquisition, Change and Evolution. Blackwell, Oxford.

Lindblom, B., P. MacNeilage, and M. Studdert-Kennedy. 1984. Self-organizing processing and the explanation of phonological universals. In B. Butterworth, B. Comrie, and O. Dahl, editors, Explanations for Language Universals. Mouton, NY.

Lindblom, Björn. 1998. Phonetic universals in vowel systems. In James R. Hurford, Michael Studdert-Kennedy, and Chris Knight, editors, Approaches to the Evolution of Language. Cambridge University Press, New York.

Lindblom, Björn and Ian Maddieson. 1988. Phonetic universals in consonant systems. In Larry M. Hyman and Charles N. Li, editors, Language, Speech, and Mind. Routledge, NY.

Loomis, Lynn. 1975. Calculus. Addison-Wesley, NY.

Lorenz, Edward. 1972. Predictability: Does the flap of a butterfly's wings in Brazil set off a tornado in texas? In AAAS Convention of the Global Atmospheric Research Program, MIT. unpublished.

Loritz, Donald. 1999. How the Brain Evolved Language. Oxford University Press, Oxford.

Lucy, John and Richard Shweder. 1988. The effect of incidental conversation on memory for focal colors. American Anthropologist, 81:923–931.

MacDermot, Kay D., Elena Bonora, Nuala Sykes, Anne-Marie Coupe, Cecilia S. L. Lai, Sonja C. Vernes, Faraneh Vargha-Khadem, Fiona McKenzie, Robert L. Smith, Anthony P. Monaco, and Simon E. Fisher. 2005. Identification of FOXP2 truncation as a novel cause of developmental speech and language deficits. American Journal of Human Genetics, 76:1074–1080.

MacLeod, Colin M. 1991. Half a century of research on the Stroop effect: An integrative approach. Psychological Bulletin, 109:163–203.

Mandelbrot, Benoit. 1961. On the theory of word frequencies and on related Markovian models of discourse. In Roman Jakobson, editor, Structure of Language in its Mathematical Aspect, Proceedings of the 12th Symposium in Applied Mathematics. American Mathematical Society, Providence, Rhode Island, pages 190–219.

Manser, Marta B., Robert M. Seyfarth, and Dorothy L. Cheney. 2002. Suricate alarm calls signal predator class and urgency. Trends in Cognitive Sciences, 6(1):55–57.

Marcus, Mitchell. 1980. A Theory of Syntactic Recognition for Natural Language. MIT Press, Cambridge, Massachusetts.

Marler, P. 1977. The structure of animal communication sounds. In T. H. Bullock, editor, Recognition of Complex Acoustic Signals. Dahlem Konferenzen, Berlin, pages 17–35.

Marslen-Wilson, William. 1975. Sentence perception as an interactive parallel process. Science, 189:226–228.

Masataka, N. and K. Fujita. 1989. Vocal learning of Japanese and rhesus monkeys. Behaviour, 109:191–199.

May, Robert M. 1976. Simple mathematical models with very complicated dynamics. Nature, 261:459–467.

May, Robert M. 2002. The best possible time to be alive. In Graham Farmelo, editor, It Must Be Beautiful: Great Equations of Modern Science. Granta, Oxford, pages 212–229.

Mayley, Giles. 1996. Landscapes, learning costs and genetic assimilation. Evolutionary Computation, 4:213–234.

Mayley, Giles. 1997. Guiding or hiding: Explorations into the effects of learning on the rate of evolution. In P. Husbands and I. Harvey, editors, Proceedings of the Fourth European Conference on Artificial Life, ECAL97.

McEvedy, Colin and Richard Jones. 1978. Atlas of World Population History. Facts on File, NY.

McKinney, M.L. and K.J. McNamara. 1991. Heterochrony: The Evolution of Ontogeny. Plenum, NY.

Mehotra, Kishan, Chilukuri K. Mohan, and Sanjay Ranka. 1997. Elements of Artificial Neural Networks. MIT Press, Cambridge, Massachusetts.

Meinhardt, Hans. 1982. Models of Biological Pattern Formation. Academic, NY.

Meinhardt, Hans. 2001. Organizer and axes formation as a self-organizing process. The International Journal of Developmental Biology, 45:177–188.

Meinhardt, Hans and Alfred Gierer. 2000. Pattern formation by local self-activation and lateral inhibition. BioEssays, 22:753–760.

Michaelis, Jens. 1998. Derivational minimalism is mildly context-sensitive. In Proceedings, Logical Aspects of Computational Linguistics, LACL'98, NY. Springer.

Miller, George A. and Noam Chomsky. 1963. Finitary models of language users. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter, editors, Handbook of Mathematical Psychology, Volume II. Wiley, NY, pages 419–492.

Miller-Ockhuizen, Amanda. 2001. Grounding Ju'Hoansi Root Phonotactics: The Phonetics of Gutteral OCP and other Acoustic Modulations. Ph.D. thesis, Ohio State University, Columbus, Ohio.

Mitani, J. 1993. Contexts and social correlates of long-distance calling by male chimpanzees. Animal Behaviour, 45:735–746.

Morré, D. James, Pin-Ju Chueh, Jake Pletcher, Xiaoyu Tang, Lian-Ying Wu, and Dorothy M. Morré. 2002. Biochemical basis for the biological clock. Biochemistry, 41(40):11941–11945.

Müller, Gerd B. 2003. Homology: The evolution of morphological organization. In Gerd B. Müller and Stuart A. Newman, editors, Origination of Organismal Form: Beyond the Gene in Developmental and Evolutionary Biology. MIT Press, Cambridge, Massachusetts, pages 51–69.

Nachman, M.W. and S.L. Crowell. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics, 156:297–304.

Nakanishi, Ryuichi, Keita Takada, and Hiroyuki Seki. 1997. An efficient recognition algorithm for multiple context free languages. In Proceedings of the Fifth Meeting on Mathematics of Language, MOL5.

Neuhauser, Claudia. 2000. Calculus for Biology and Medecine. Prentice-Hall, New Jersey.

Niyogi, Partha. 1999. Models of cultural evolution and their application to language change. In E.J. Briscoe, editor, Language Evolution through Language Acquisition. Cambridge University Press, Cambridge.

Niyogi, Partha and Robert C. Berwick. 1994. A dynamical systems model for language change. A.I. technical report no. 1515, Massachusetts Institute of Technology.

Niyogi, Partha and Robert C. Berwick. 1995. A note on Zipf's law, natural languages and non-coding DNA regions. A.I. technical report no. 1530, Massachusetts Institute of Technology.

Nowak, Martin A., Natalia Komarova, and Partha Niyogi. 2002. Computational and evolutionary aspects of language. Nature, 417:611–617.

Nowak, Martin A. and Robert M. May. 2000. Virus Dynamics. Oxford University Press, Oxford.

Ochs, Elinor and Bambi B. Schieffelin. 1999. Language acquisition and socialization: Three developmental stories. In R. Shweder and R. LeVine, editors, Culture Theory: Essays in Mind, Self and Emotion. Cambridge University Press, Cambridge, pages 276–320.

Ogura, Mieko. 1993. The development of perphrastic do: a case of lexical diffusion in syntax? Diachronica, 10:51–85.

Ohala, John J. 1990. The phonetics and phonology of aspects of assimilation. In J. Kingston and M. Beckman, editors, Papers in Laboratory Phonology I: Between the grammar and physics of speech. Cambridge University Press, Cambridge.

Oró, Juan. 1961. Mechanism of synthesis of adnenine from hydrogen cynanide under plausible primitive earth conditions. Nature, 191:1193–1194.

Orr, H. Allen. 1995. Dennett's dangerous idea. Evolution, 50:467–472.

Osawa, S., T.H. Jukes, K. Watanabe, and A. Muto. 1992. Recent evidence for evolution of the genetic code. Microbiological Reviews, 56:229–264.

Palumbi, Stephen R. 2001. Humans as the world's greatest evolutionary force. Science, 293(5536):1786–1790.

Paun, G., G. Rozenberg, and A. Salomaa. 1998. DNA Computing: New Computing Paradigms. Springer, NY.

Pennartz, Cyrial M.A., Marcel T.G. de Jeu, Nico P.A. Bos, Jeroen Schaap, and Alwin M.S. Guertsen. 2002. Diurnal modulation of pacemaker potentials and calcium current in the mammalian circadian clock. Nature, 416:286–290.

Perline, Richard. 1996. Zipf's law, the central limit theorem, and the random division of the unit interval. Physical Review E, 54:220–223.

Pinker, Steven. 1994. The Language Instinct. William Morrow, NY.

Pinker, Steven. 2000. Survival of the clearest. Nature, 401:442–443.

Pinker, Steven. 2001. Talk of genetics and vice-versa. Nature, 413:465–466.

Pinker, Steven. 2002. The Blank Slate: The Modern Denial of Human Nature. Viking, NY.

Pinker, Steven and Paul Bloom. 1990. Natural language and natural selection. Behavioral and Brain Sciences, 13:707–784. Reprinted in J. Barkow, L. Cosmides, and J. Tooby, eds., 1991, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York: Oxford University Press.

Platt, J.R. and D.M. Johnson. 1971. Localization of position within a homogenous behavior chain: Effects of error contingencies. Learning and Motivation, 2:386–414.

Premack, David. 1985. 'gavagai' or the future history of the animal language controversy. Cognition, 19:207–296.

Przeworski, Molly. 2002. The signature of positive selection at randomly chosen loci. Genetics, 160:1179–1189.

Rambow, Owen. 1994. Formal and computational aspects of natural language syntax. Ph.D. thesis, University of Pennsylvania. Computer and Information Science Technical report MS-CIS-94-52 (LINC LAB 278).

Rendall, Drew, Dorothy L. Cheney, and Robert M. Seyfarth. 2000. Proximate factors mediating 'contact' calls in adult female baboons *papio cynocephalus ursinus* and their infants. Journal of Comparative Psychology, 114:36–46.

Rendall, Drew, Robert M. Seyfarth, Dorothy L. Cheney, and Michael J. Owren. 1999. The meaning and function of grunt variants in baboons. Animal Behaviour, 57:583–592.

Richerson, Peter J. and Robert Boyd. 2005. Not By Genes Alone: How Culture Transformed Human Evolution. University of Chicago Press, Chicago.

Rieke, Fred, David Warland, Rob de Ruyter van Steveninck, and William Bialek. 1997. Spikes: Exploring the Neural Code. MIT Press, Cambridge, Massachusetts.

Roberts, Ian and Anna Roussou. 2002. The history of the future. In David W. Lightfoot, editor, Syntactic Effects of Morphological Change. Oxford University Press, NY.

Roberts, Ian G. 1993. Verbs and Diachronic Syntax. Kluwer, Boston.

Robinson, J. G. 1984. Syntactic structures in the vocalizations of wedge-capped capuchin monkeys, cebus nigrivittatus. Behaviour, 90:46–79.

Ronshaugen, Matthew, Nadine McGinnis, and William McGinnis. 2002. Hox protein mutation and macroevolution of the insect body plan. Nature, 415:914–917.

Ross, John R. 1967. Constraints on Variables in Syntax. Ph.D. thesis, Massachusetts Institute of Technology.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. 2006. Positive natural selection in the human lineage. Science, 312:1614–1620.

Sackett, G.P. 1970. Unlearned responses, differential rearing experiences, and the development of social attachments by rhesus monkeys. In L.A. Rosenblum, editor, Primate Behavior, Volume 1. Academic, NY, pages 81–98.

Sakakibara, Y., M. Brown, R. Hughey, I. S. Mian, K. Sjolander, R. C. Underwood, and D. Haussler. 1994. Stochastic context-free grammars for tRNA modeling. Nucleic Acids Research, 22(23):5112–5120.

Salvini-Plawen, L.V. and E. Mayr. 1977. On the evolution of photoreceptors and eyes. In M.K. Hecht, W.C. Steere, and B. Wallace, editors, Evolutionary Biology. Plenum, NY, pages 207–263.

Sandler, Wendy and Diane Lillo-Martin. 2001. Natural sign languages. In Mark Aronoff and Janie Rees-Miller, editors, The Handbook of Linguistics. Blackwell, Oxford, pages 533–562.

Schrödinger, Erwin. 1945. What is Life? The Physical Aspect of the Living Cell, Mind and Matter. Cambridge University Press, Cambridge.

Searles, David B. 2002. The language of genes. Nature, 420:211–217.

Seki, Hiroyuki, Takashi Matsumura, Mamoru Fujii, and Tadao Kasami. 1991. On multiple context-free grammars. Theoretical Computer Science, 88:191–229.

Seyfarth, Robert M. and Dorothy L. Cheney. 1986. Vocal development in vervet monkeys. Animal Behaviour, 34:1640–1658.

Shannon, Claude E. 1948. The mathematical theory of communication. Bell System Technical Journal, 127:379–423. Reprinted in Claude E. Shannon and Warren Weaver, editors, *The Mathematical Theory of Communication*, Chicago: University of Illinois Press.

Shen, Ling X. and Ignacio Tinoco. 1995. The structure of an RNA pseudoknot that causes efficient frameshifting in mouse mammary tumor virus. Journal of Molecular Biology, 247(5):963–978.

Sherman, Gavin and P. Kirk Visscher. 2002. Honeybee colonies achieve fitness through dancing. Nature, 419:920–922.

Slobodchikoff, C.N., J. Kiriaz, C. Fischer, and E. Creef. 1991. Semantic information distinguishing individual predators in the alarm calls of Gunnison's prairie dogs. Animal Behaviour, 42:713–719.

Snedeker, Jesse and Lila R. Gleitman. 2004. Why it is hard to label our concepts. In D.G. Hall and S.R. Waxman, editors, Weaving a Lexicon. MIT Press, Cambridge, Massachusetts, pages 257–293.

Snowdon, C.T. 1990. Language capacities of nonhuman primates. Yearbook of Physical Anthropology, 33:215–243.

Sogin, Mitchell L. and David J. Patterson. 1992. Eukaryote origins and protistan diversity. In H. Hartman and K. Matsuno, editors, The Origin And Evolution of the Cell. World Scientific, New Jersey, pages 13–46.

Stabler, Edward P. 1994. The finite connectivity of linguistic structure. In Charles Clifton, Lyn Frazier, and Keith Rayner, editors, Perspectives on Sentence Processing. Lawrence Erlbaum, Hillsdale, New Jersey, pages 245–266.

Stabler, Edward P. 2001. Minimalist grammars and recognition. In Christian Rohrer, Antje Rossdeutscher, and Hans Kamp, editors, Linguistic Form and its Computation. CSLI Publications, Stanford, California. (Presented at the SFB340 workshop at Bad Teinach, 1999).

Steels, Luc. 1996. Synthesizing the origins of language and meaning using co-evolution, self-organisation and level formation. In J. Hurford, C. Knight, and M. Studdert-Kennedy, editors, Evolution of Human Language. Edinburgh University Press, Edinburgh, pages 161–165.

Steriade, Donca. 1995. Positional neutralization. ms.

Stewart, James. 2003. Single variable calculus. Thomson, Belmont, California.

Stojanovic, Milan N. and Darko Stefanovic. 2003. A deoxyribozome-based molecular automaton. Nature Biotechnology, Advance online publication: 17 August 2003, doi:10.1038/nbt862.

Stromswold, Karin. 1994. Language comprehension without language production: Implications for theories of language acquisition. In 18th Annual Boston University Conference on Language Development, Boston. Boston University.

Stromswold, Karin. 1998. The heritability of language: A review and metaanalysis of twin, adoption and linkage studies. Human Biology, 70:297–324.

Stroop, J. Ridley. 1935. Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 12:643–662.

Sun, Xin, Francesca V. Mariani, and Gail R. Martin. 2002. Functions of FGF signalling from the apical ectodermal ridge in limb development. Nature, 418:501–508.

Tattersall, Ian. 2003. Once we were not alone. In New Look at Human Evolution. Scientific American (Special Edition).

Teahan, W.J. 1998. Modelling English Text. Ph.D. thesis, University of Waikato.

Thompson, D'Arcy Wentworth. 1917. On Growth and Form. Dover (revised edition 1992), NY.

Tickle, Cheryll and Lewis Wolpert. 2002. The progress zone – alive or dead? Nature Cell Biology, 4(9):E216–E217.

Tinoco, Ignacio and Carlos Bustamante. 2000. How RNA folds. Journal of Molecular Biology, 293:271–281.

Tomasello, Michael. 2003. Constructing a language: A Usage-based Theory of Language Acquisition. Harvard University Press, Cambridge, Massachusetts.

Tomasello, Michael and Josep Call, editors. 1997. Primate Cognition. Oxford University Press, NY.

Turing, Alan M. 1952. The chemical basis of morphogenesis. Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences, pages 37–72.

Turing, Alan M. ca. 1952. Outline of the development of the daisy. In P.T. Saunders, editor, Morphogenesis: Collected Works of A.M. Turing, Volume 3. North-Holland, 1994, Amsterdam.

Tyers, Mike and Matthias Mann. 2003. From genomics to proteomics. Nature, 422:193–197.

United Nations Secretariat, Population Division of the Department of Economic and Social Affairs. 2001. World population prospects: The 2002 revision and world urbanization prospects: The 2001 revision. Technical report, UN.

van Kampen, Jacqueline. 1997. First steps in wh-movement. Ph.D. thesis, University of Utrecht.

Varela, Francisco J. 1994. On defining life. In G. Fleischaker and M. Colonna, editors, Self-reproduction of Supramolecular Structures, NATO ASI Series. Plenum, NY, pages 23–33.

Vargha-Khadem, F., K. Watkins, K. Alcock, P. Fletcher, and R Passingham. 1995. Praxic and non-verbal cognitive deficits in a large family with a genetically transmitted speech and language disorder. Proceedings of the National Academy of Sciences, 92:930–933.

Voet, A.B. and A.W. Schwartz. 1982. Uracil synthesis via HCN oligomerization. Origins of Life and Evolution of the Biosphere, 12:45–49.

Waltz, David. 1975. Understanding line drawings of scenes with shadows. In P. Winston, editor, The Psychology of Computer Vision. McGraw-Hill, NY, pages 19–91.

Warren, R.M. and G. Sherman. 1974. Phonemic restorations based on subsequent context. Perception and Psychophysics, 16:150–156.

Warren, R.M. and R.P. Warren. 1970. Auditory illusions and confusions. Scientific American, 223:30–36.

Watkins, K.E., N.F. Dronkers, and F. Vargha-Khadem. 2002. Behavioural analysis of an inherited speech and language disorder. Brain, 125:452–464.

Watson, James D. 2003. DNA: The Secret of Life. Knopf, NY. With Andrew Barry.

Weir, David. 1988. Characterizing mildly context-sensitive grammar formalisms. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Whorf, Benjamin L. 1941. The relation of habitual thought and behavior to language. In Leslie Spier, editor, Language, culture, and personality, essays in memory of Edward Sapir. Sapir Memorial Publication, Menasha, Wisconsin.

Williams, George C. 1966. Adaptation and Natural Selection. Oxford University Press, Oxford.

Wilson, Edward Osborne. 1971. The Insect Societies. Harvard University Press, Cambridge, Massachusetts.

Wilson, Edward Osborne. 2000. Sociobiology: The New Synthesis, Twenty-fifth Anniversary Edition. Harvard University Press, Cambridge, Massachusetts.

Winter, P., P. Handley, W. Ploog, and D. Schott. 1973. Ontogeny of squirrel monkey calls under normal conditions and under acoustic isolation. Behaviour, 47:230–239.

Wittgenstein, Ludwig. 1958. Philosophical Investigations. MacMillan, NY. This edition published in 1970.

Woese, Carl R. 1997. Bacterial evolution. Microbiological Reviews, 51:221–271.

Wonnacott, Elizabeth. 2000. Investigating the trajectory of language change. University of Edinburgh M.A. thesis.

Yang, Charles. 2000. Knowledge and learning in natural language. Ph.D. thesis, Massachusetts Institute of Technology.

Zelditch, M.L., editor. 2001. Beyond Heterochrony: The Evolution of Development. Wiley, NY.

Zipf, George K. 1949. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Houghton-Mifflin, Boston.

Zuberbühler, Klaus. 2002. A syntactic rule in forest monkey communication. Animal behavior, 419:920–922.

Zuberbühler, Klaus, Dorothy L. Cheney, and Robert M. Seyfarth. 1999. Conceptual semantics in a nonhuman primate. Journal of Comparative Psychology, 113:33–42.

Index