# Lecture 4: Comparing Languages

**Focus Points:**

- **Overview of Linguistic Reconstruction**
- **Ways of Comparing Languages: Contrastive vs. Comparative**
  - **Sources of Linguistic Similarity**
  - **Building a Family Tree: Reconstructing the Parent Language. Types of Reconstruction.**
- **The Comparative Method:**
  - **Cognates: the Majority Principle**
  - **The Most Natural Development Principle**

## 4.1 Overview of Linguistic Reconstruction

Until Chomsky, the main preoccupation of linguists was reconstructing earlier stages of individual languages and of language families, in particular the Indo-European family, to which English, French and German belong. Reconstruction was popular for good reasons: there are many benefits and pleasures – as well as dangers – to be found in pursuing language history:

- Pleasures of reconstruction:
  - Explanation of odd archaisms in one's language
  - Picture of the prehistory of a people
  - What they were like: how they spoke, plants and foods they knew, etc.
  - Where they were
  - Who they were in contact with
  - Also: practical help in learning/understanding languages

- Dangers of reconstruction:
  - May stimulate the rise of extreme nationalism and xenophobia. For example, the development of IE linguistics played a major role in the development of nationalism in Europe, and led to the formulation of the Aryan theory. Aryan Invasion Theory was one of the central myths of Nazi ideology, preaching Aryan racial superiority. Nazi extermination policy was rooted in eliminating 'inferior' races threatening to Aryan racial purity.
  - The homeland problem (colonialist expansion)

The developers of the Aryan theory were themselves dismayed by the unforeseen power of their model, and ended up renouncing their views. This is one of the reasons why linguists today steer clear of the larger issues that occupied the 19$^{th}$ century linguists and stick to the three broad areas which, as you hopefully remember, constitute the current aims and scope of comparative linguistics:

- The study of the *history* of particular languages on the basis of existing written data.
- The study of the *prehistory* of languages by means of **comparative reconstruction**, whereby the unrecorded past is inferred on the evidence of the data that *are* available from a later period.
- The study of *ongoing changes* in a language, i.e. changes happening at the present time.

Now, before we go into the details of comparative reconstruction, let us look at different ways of comparing languages, the sources of linguistic similarity, and the general principles of language classification.

## 4.2 Ways of Comparing Languages: Contrastive vs. Comparative

Estimates of the number of world languages alive today vary, partly because of differences in the definition of 'language.' The commonly quoted figure is between 4 and 8 thousand. Some linguists carry out analysis of individual languages. Many more, however, are involved in comparing pairs, or groups of them. Languages can be compared in two different ways, depending on whether the researchers want:

- **To pinpoint the dissimilarities** between them (**contrastive linguistics**),

or

- **To identify similarities**, which may be due to **universal, genetic, areal,** or **typological** factors.

**Contrastive Linguistics** focuses on the differences between languages. It is particularly useful to know how languages differ from each other when we want to learn/teach another language. By analogy, we transfer familiar linguistic structures to the language we are learning, so if the 'target' language is very different from the one we know, then that is exactly where we are likely to find difficulties – in those differences. For example, in Hindi negation is typically expressed by placing a single negative word before the verb, which comes at the end of the sentence:

*Peter Hindustani **nahi** hai*.
Peter Indian not is – Peter is not Indian.

By analogy, Indian learners of English tend to produce sentences like '*All of these machines don't work*,' instead of the typical English way of putting it, '*None of these machines work.*'

Contrastive linguists, therefore, make detailed comparisons of pairs of languages in order to identify the areas, which are most likely to cause difficulties to the learners. This knowledge is useful for devising the most effective teaching methods.

The major task of modern linguistics, however, has been the search for language universals and reconstruction of proto languages. Language universals have proven to be rather elusive, which is why most linguists study characteristics shared by groups/types of languages, rather than all of them.

The starting point for linguistic research in this area is a similarity observed between two or more forms, either within a single language, or between two or more languages.

### 4.2.1 Sources of Linguistic Similarity

Linguistic similarity can result from different causes, which may be due to **chance, areal**, **typological,** or **genetic factors**:

- **Chance** similarity is possible because language symbols are arbitrary:

  English *bad* : Persian *bad*
  English *who* : Karabagh *hu* [hu:]
  English *hair* : Armenian *her*
  Latin *habere* : German *haben*

- **Areal** similarities result from contact between neighbouring languages (borrowings, loans, etc.):

  Japanese '*kaban*' (bag) : Kuman '*kaban*' (bag)
  Japanese '*wakai*' (good) : Kuman '*wakai*' (good)
  French '*plage*' (beach) : Turkish '*plaz*' (beach)
  Chinese: *wok, typhoon, ketchup*

Gaelic (Irish): *galore, hooligan, whiskey*
Italian: *bizarre, spaghetti, soprano,* etc.
Arabic: *zero, algebra, candy, ghoul*
Australian aborigine: *boomerang, kangaroo*
Hindi: *swastika, khaki, pyjamas, bungalow*
Aztec: *tomato, potato, chocolate*
Latin: video, data, Volvo, et cetera :)

Also: Norman vocabulary in English: mutton, beef, veal, button, glutton, etc., and French food words: *courgettes, aubergines, croissants,* etc.

Borrowing of *constructions* is more likely if languages are *structurally similar*. But even dissimilar languages can, over time, gradually absorb features from one another. If a particular characteristic has spread over a wide range, linguists sometimes talk about **linguistic areas**.

**There are two reasons for studying areal characteristics**:

- It is useful to know how languages can affect one another when trying to understand language change;
- It is important to isolate shared features caused by borrowing, so as not to confuse them with genetic and typological similarities.

- **Typological** similarities occur when languages belong to the same overall 'type': just like we can divide human beings into different racial types on the basis of their physical characteristics, we can also divide languages into different types, based on their linguistic features. General failure to find convincing numbers of language universals, apart from vague statements like ' All languages have the means to ask questions,' has channelled linguistic thought into attempts to identify the main **language types**, based on *how* their systems operate. The observation that different languages use different constructions is not a new discovery. What is new is the recent interest in **implicational universals** and implicational tendencies. That is, if a language has a particular construction, it is also likely to have further predictable characteristics. Just like if an animal has feathers and a beak, it is also likely to have wings, so if a language has the basic pattern of S/V/C, it is also likely to have prepositions (rather than postpositions).
(We shall discuss morphological and syntactic criteria for language classification in our next class.)

- **Genetic** similarities result from common ancestry (common source). What sorts of similarity demonstrate a common source? **Systematic correspondences of sounds and morphology***

  ***N.B.: Not** typological features: i.e., such as Altaic and Uralic vowel harmony.

  Note that loanwords are least common among the most frequently used lexical items.

  **<u>Two basic assumptions underlie our search for systematic correspondences</u>:**

  - <u>First</u>**, linguistic symbols are essentially arbitrary:** as a rule, there is no connection between the sound of a word and the thing it symbolizes, except in the case of occasional onomatopoeic words. Therefore, consistent similarities between languages, which cannot be explained by borrowing, may be due to common origin (source).

- Second, **sound changes are for the most part regular**: if one sound changes, then all similar sounds in the same phonetic environment and geographical area change also.

On the basis of these two assumptions, we can draw up reliable and systematic correspondences between the various related languages. The correspondences that we are looking for may be in the sounds (phonology), or, more reliably, in the morphology, since it is rare for one language to borrow another's morphology.

Look at these examples:

| German /d/ | **meaning** | English [θ] |
|---|---|---|
| dick | **fat** | thick |
| Ding | **thing** | thing |
| Bad | **bath** | bath |
| German [ʃv] | **meaning** | English [sw] |
| schwimmen | **swim** | swim |
| scwingen | **swing** | swing |
| Schwan | **swan** | swan |

These systematic sound correspondences between words with the same or similar meaning are the first clue that the languages may be related. The evidence is cumulative: the more correspondences, the more likely the languages are to be related.
**N.B.** Correspondences may result from borrowing: we might be dealing with a series of loanwords which diverged in development after having been borrowed, i.e., a superficial correspondence between:

| French | English |
|---|---|
| mouton | mutton |
| bouton | button |
| glouton | glutton |

These are all words borrowed from French at the time of the Norman invasion.

**Morphological correspondences are more reliable:**

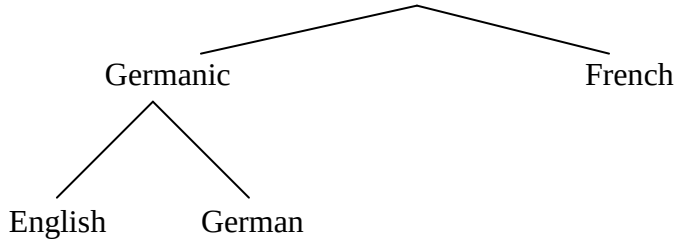
| German | | | English | | |
|---|---|---|---|---|---|
| No suffix | /ə/ | /stə/ | No suffix | /ə/ | /əst/ |
| klein | klein**er** | klein**ste** | small | small**er** | small**est** |
| schnell | schnell**er** | schnell**ste** | quick | quick**er** | quick**est** |
| reich | reich**er** | reich**ste** | rich | rich**er** | rich**est** |

Such correspondences prove 'beyond any reasonable doubt' that German and English are close relatives.
(Hittite was established as an IE language on the basis of morphological correspondences, despite its non-IE vocabulary!)

### 4.2.2 Building a Family Tree: Reconstructing the Parent Language

#### Types of Reconstruction

Once we have established that languages are related, then we have to explain *how* they are linked. If we find three related, or cognate languages, say, German, English and French, then we have to decide whether they are three offshoots from the same parent, or whether two of them diverged from one another at a later stage:




This would affect the reconstruction, because we would then have to reconstruct the ancestor of German and English before moving on to the next stage, that of reconstructing the overall ancestor (the 'grandmother').

**Reconstructing the Parent Language: Materials for Reconstruction**

When we have established a genetic relationship, and set up a family tree, we can begin to reconstruct.

What materials can we use for reconstruction?

- Living languages
- Dialects (discussion of internal reconstruction in a few minutes☺)
- Texts from the past

How can we be sure of the phonetics, phonology, semantics of a dead language?

- In some cases, we have contemporary descriptions (Varrro, Panini)
- Scripts often have modern descendants
- Semantics only controllable via
    - Comparison with related languages
    - Internal cross-reference (with a sufficient *corpus*, one can test one's hypothesis).

Recently, we have discovered a lot about language change through the study of changes in progress. However, language change is a relatively slow process, so we also need to consider how languages have altered over the centuries. Since only about 200 of the existing 6000 languages are written, and since the existing records are scarce and incomplete, historical linguists devised methods of reconstructing stages of language for which there are no written documents. Our synchronic knowledge of the language is sufficient to give us a good idea of what sort of a system a human language can have, and what sorts of changes are possible.

With these cautions in mind, we can turn to the actual **mechanics** of reconstruction. There are several types of reconstruction:

- **External Reconstruction**, which compares cognates from genetically related languages and attempts to infer conclusions about their parent language. ER requires a model of historical development: 'tree' (Schleicher; influenced by biology and Darwin, thought IE was the 'fittest'), 'wave' (Schmidt), 'crazy' (David Michaels, Greenberg), etc.
- **Internal Reconstruction**, which looks at the state of one language at a single point in time and compares elements, which are likely to have had common origin. This often provides

the basis for conclusions about their earlier forms. To take a simple example, consider the words *long* and *longer*. [lON] and [long] are both allomorphs of the morpheme *long*. This suggests that originally they were identical, and that the word *long* was once pronounced with a /g/ at the end (as it is still in some parts of England, such as Liverpool). Thus, IR typically involves looking at synchronic alternations in a language, and postulating a single source that gave rise to all of the variant forms. Another typical example: alternation between /s/ and /r/ in Latin:

- **N**. flos G. floris
- **N**. honos G. honoris

For the synchronic grammar of Latin we postulate underlying genitive forms [flosis] and [honosis] plus a phonological rule that changes /s/ to /r/ between vowels. It seems to be the case that underlying forms generally represent the historically older form; we can therefore postulate that Latin once had surface genitives *flosis and *honosis, and did not have the *s* > *r* rule. This rule is paralleled in the Germanic family:

was: were < OE wɒz, wæːɹon < IE *wes- 'be'

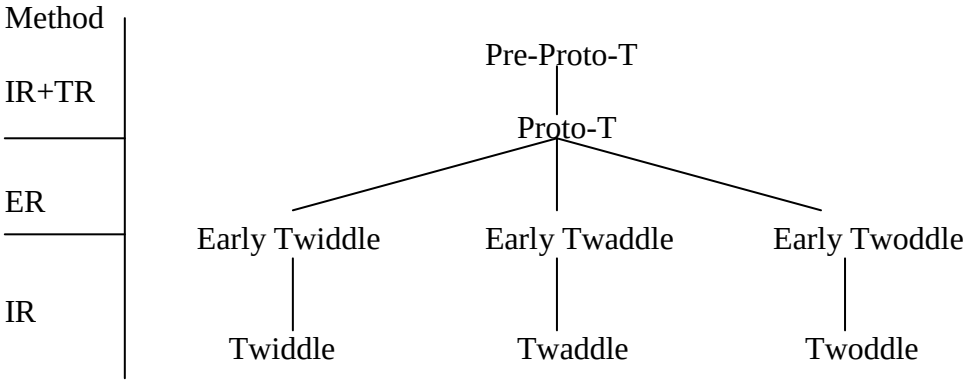
The Germanic case suggests that the change historically was *s* > *z* > *r*.

- **Typological Reconstruction**, which is somewhat newer than the other two. Linguists are beginning to be able to discern different *types* of languages, each with their own basic characteristics. For example, languages like Hindi, which have verbs after their objects, also tend to have auxiliary (modal) verbs after the main verb. On the other hand, languages like English, which have verbs before the objects, tend to have auxiliary verbs before the main verb. If therefore we found some records of a language with the auxiliaries (modals) after the main verb, we would be able to infer that it might also have its object before the verb, even if we had no direct evidence for this.

In order to do external reconstruction, one first has to do internal reconstruction to reconstruct the earliest form of each daughter language or family that one is considering.

Let us now see how we might use these different types of reconstruction. Suppose we have three related languages, Twiddle, Twaddle, and Twoddle. Let us also assume that we have no past records, only records of their present-day speech.

- First, we would use internal reconstruction (IR) to reconstruct an earlier stage of each of these languages – Early Twiddle, Early Twaddle, and Early Twoddle.
- Then we would use ER (external reconstruction) to reconstruct Proto-T, the common ancestor of these three.
- Then, once again, we would employ internal reconstruction (IR), this time combined with typological reconstruction (TR) to reconstruct an earlier form of Proto-T: Pre-Proto-T:



In this way, we might succeed in reconstructing a probable history of these languages reaching far into their past. It is not easy - it's a bit like trying to work out what the grandmother must have been like on the basis of common features in her grandchildren.
Let us now take a closer look at the **basic principles of the comparative method** used in linguistic reconstruction:

## 4.3 The Comparative Method

The two most basic principles of this procedure are:

- **The majority principle, and**
- **The most natural development principle.**

**The majority principle** is very straightforward: if, in a cognate set, three forms begin with a [p] sound and one form begins with a [b] sound, then our best guess is that the majority have retained the original sound (i.e., [p]), and the minority has changed a little through time. Example:

| Tongan reflex | Samoan reflex | Rarotongan ref. | Hawaiian reflex | gloss |
|---|---|---|---|---|
| tapu | tapu | tapu | kapu | forbidden |
| pito | pute | pito | piko | navel |
| tahi | tai | tai | kai | sea |
| ua | lua | rua | lua | two |

**The most natural development principle** is based on the fact that certain types of sound change are very common, or typical, whereas others are unlikely. We shall discuss types of sound change in greater detail later in the course – here are just a few examples of some well-documented types of sound change in the Indo-European languages:

- **Final vowels often disappear**
- **Voiceless sounds become voiced between vowels**
- **Stops become fricatives** (under certain conditions)
- **Consonants become voiceless at the end of words.**

Here are some examples from three languages – which is the most likely form of the initial sound in the original language source of the three?

| Language A | Language B | Language C | English |
|---|---|---|---|
| cavallo | caballo | cheval | horse |
| cantare | cantar | chanter | sing |
| catena | cadena | chaine | chain |
| caro | caro | cher | dear |

The initial sound in LA and LB is /k/, whereas in LC it is /ʃ/. According to the majority principle alone it is reasonable to conclude that in the ancestor language the words began with the /k/ sound. Do we have any other 'supporting evidence'?
Under the most natural development principle we have this tendency listed, for stops to become fricatives. This certainly supports our conclusion based on the majority principle, and makes it even more likely. Through this type of procedure, we have started on the comparative method of reconstruction of the common origins of Italian (LA), Spanish (LB), and French (LC). We are lucky here, because the parent language (Latin) has some written records, so we can check out our conclusions: Latin cognates for these forms are: ***Caballus, cantare, catena,*** and ***carus***!

Now try to reconstruct Proto-Romance forms from:

| Spanish | Italian | French | Portuguese | Romanian | Catalan | gloss |
|---|---|---|---|---|---|---|
| estrella | stella | etoile [etwal] | estrella | stea | estel(a) | star |
| llave | chiave | clef | chave | cheie | clau | key |
| dona | donna | dame | dona | doamna | dona | mistress |
| dedo | dito | doigt [dwa] | dedo | deget | dit | finger |
| lagrima | lagrima | larme | lagrima | lacrama | llagrema | teardrop |

N.B.:

- Easy to get epenthesis (insertion of a vowel between consonants in a consonant cluster) and loss of r, ll, and final vowel in 'star'
- Easy to explain dislike of initial cluster in 'key'; harder to go from **cave to clau
- French and Romanian show that 'mistress' had to have /m/ (*dom(i)na)
- 'finger' had to have a /g/ and a /t/ (*digito) Romanian preserves original /k/ in 'teardrop' (harder to go from **lagrima to lacrama)

Answer:

| Spanish | Italian | French | Portuguese | Romanian | Catalan | Latin | Proto-Romance |
|---|---|---|---|---|---|---|---|
| estrella | stella | etoile [etwal] | estrella | stea | estel(a) | ste:lla | strella |
| llave | chiave | clef | chave | cheie | clau | clavis | clave |
| dona | donna | dame | dona | doamna | dona | domina | domna |
| dedo | dito | doigt [dwa] | dedo | deget | dit | digitus | digito |
| lagrima | lagrima | larme | lagrima | lacrama | llagrema | lacrima | lakrima |

**Questions**

1. What is **Contrastive Linguistics**?
2. Suggest three reasons why languages might show similarities.
3. How do we determine that languages are related *genetically*?
4. What is the purpose of reconstructing a proto-language?
5. What is the difference between **Internal Reconstruction** and **External Reconstruction**?
6. What are **implicational universals**?

**Additional Reading:**

**Cognates** by Dr. Alex J. Cameron

The eighteenth century marked a sharp rise of interest in science and reason. Nearly every field of study was readjusted to reflect this interest. For example, the term *linguist* had earlier meant someone skilled in speaking or reading a number of languages and perhaps knowledgeable in the cultures that accompanied these languages. But now, there was a concerted attempt to make linguistics over into a science of language. Many speculative essays were written about the origin of language. And there was a great deal of interest in determining the exact relationships among European languages. A certain similarity in vocabulary was obvious to everyone. Consider a simple example:

| Old English | German | Dutch | Greek | Latin | French | Spanish |
|---|---|---|---|---|---|---|
| *fæder* | *vater* | *vader* | *patér* | *pater* | *pere* | *padre* |

There was enough recent historical information to know that French and Spanish had developed out of Vulgar Latin, but that knowledge led to only more questions. At some very early date, had Latin grown out of or split off from Greek? In the Middle Ages, English borrowed many words from Latin, but this fact did not explain basic vocabulary items such as English *night,* German *nacht,* and Latin *noctis.* Coincidence might explain one or two similarities, but these linguists were dealing with hundreds of examples.

During this period, an attorney named William Jones was sent to India to serve as a judge in the British Colonial courts. Jones, an expert linguist, was also an orientalist; his hobby was the study of eastern culture. Shortly after his arrival in 1783 Jones began to study ancient Sanskrit. What Jones discovered makes an interesting addition to the chart illustrated above. It now looks like this:

| Old English | German | Dutch | Greek | Latin | French | Spanish | Sanskrit |
|---|---|---|---|---|---|---|---|
| *fæder* | *vater* | *vader* | *patér* | *pater* | *pere* | *padre* | *pitar* |

The Sanskrit word for father, *pitar* (or *p'tar*), is, we must remember, only a representative of the hundreds of similarities that Jones discovered. Moreover, because religions proscriptions forbade modernizing sacred writings, the age of these examples and the geographical distance made it all but impossible to continue thinking about one known language descending from another. In 1786 Jones delivered a paper in Bengal in which he concluded that the "vertical" visualization of these languages as ancestors or descendants was simply wrong. They have to be visualized "horizontally" as "cousins" of each other. Jones hypothesized a much older ancestor language from which these others were all descended. He proposed that Sanskrit and the languages of Europe were "co-born" (*co-gnatus*); they were cognates. Once word of this suggestion began to spread, full-time, professional linguists took over the task of comparing all these languages—and the results were remarkable.

By closely comparing these cognate words and examining even the phonetic elements that make them up, linguists were able to begin constructing earlier, "ancestor" forms. These reconstructions are speculative, but recently discovered manuscripts have inspired increased confidence in their accuracy. Over the last two centuries the process of reconstruction has taken us back to a language that has come to be known as Indo-European, or Proto-Indo-European (PIE). The name merely indicates the probable geographical location of the people who spoke this language. With two or three exceptions, all the languages of Europe, as well as many of those in India, are descended from PIE.

**Five or six thousand years ago PIE was probably a simple, pragmatic language with words intended to denote objects or actions.** Over the course of time, as people encountered new experiences, some words added more abstract meanings and some became more metaphorical. New words were added if old ones could not accommodate new needs. This sequence is the ordinary progress of most languages, but it was dramatically jostled by changes in the lives of these PIE people. About 4,000 years ago, something—we don't know what—changed their lives and prompted a series of migrations. The migrations occurred periodically over many centuries, first to the south and east into India, later to the west across all of Europe.

Early migrants were generally successful at occupying land and replacing the cultures they encountered with their own values and language. Over time, languages naturally change, so later migrants meeting their distant kinsmen thought of them as strangers. An elaborate network of competition evolved as successive waves of Indo-Europeans struggled for survival and dominance. As PIE developed into different dialects and then into different languages, identical words looked less and less alike, evolved different meanings, and eventually were subject to borrowing from one language by another.

Let us take as an example the reconstructed PIE root: *STA.* It seems to have meant "stand" in the sense of "stay" or "not move." (It is only a short step to "firm" or "unyielding.") From this root we get words such as:

| | |
|---|---|
| *stand* | *circumstance* ("what stands around") |
| *standing* | *substance* ("what stands under, is present") |
| *stand up* | *constant* ["what (always) stands with"] |
| *stanch* ["to stop (the flow)"] | *substitute* ("stands in place of") |
| *staunch* ("to stand by") | *extant* ("stands forth, present") |
| *stay* ("to stand still, halt") | *distant* ("stands away from") |
| *stance* ("standing place, position") | *instance* ("what stands upon, is present") |

Old Indian has the form *STHANA,* from which we have created geographical words such as:

*Hindustan* ("where the Hindus stay")
*Kurdistan* ("where the Kurds stay")
*Afghanistan* ("where the Afghans stay")

Old Latin developed the form *STLOCUS,* from which came *local* and *location.* The Latin verb *stare* ("to stand") has participial forms *stantum* and *status* and the imperative *stet* ("let it stand").
Old French developed the form *estandard,* from which comes *standard.* Other related forms are:

| | |
|---|---|
| *state* ("how things stand") | *statistic* [from Latin *status* ("how things stand")] |
| *estate* ("where the family stands/stays") | *stale* ("that stands too long") |
| *static* ("standing still") | *staid* ("firm, fixed") |

A variant form of the PIE root was *STAK* ("stay, support"), from which English created *stay,* a word for one of the several braces or supports in a corset.

A second PIE variant was *STEL* ("put, put on, solid, firm"). From this came Greek *stellein* ("put in order, send"). From this Greek root English gets *apostle, epistle, epistolary, apostate* ("stand away"), *systole, diastole,* and *peristalsis.*
Latin *stolo* gave us *stolid* and *stultify.* Italian developments of this form yielded *pedestal, stalwart, stalk* ("what a plant stands on"), and *stout.*
Germanic *stelt* ("to cause to stand") gave us *still, stilled,* and *gestalt.*
The technical suffix *stat* appears in *rheostat* and *thermostat.*
There are many, many other cognate forms under *STA.* Here is a sampling:

| | |
|---|---|
| *stable* ("stands still; where animals stand") | *instead* ["in place (of)"] |
| *stability* ("standing still") | *statute* ("a standing law") |
| *stall* ["(place to) stand"] | *stature* ["standing (tall)"] |
| *stallion* ("horse kept in a stall") | *stationary* ("standing still") |
| *steed* ("horse kept in a stall") | *destine, destiny* ("future that is fixed") |
| *forestall* ("to stand before, intercept") | *constitution* ("what is placed, set up") |
| *install(ment)* ("to place") | *obstinate* ("standing firm") |
| stead ("standing, a place") | obstacle ("what stands in the way") |
| homestead ("where a home stands") | |

Sometimes, the *ST* construction stands at the end of a word:

| | |
|---|---|
| *arrest* ("to stay, stop") | *insist* ("to stand upon") |
| *assist* ("to stand by") | *persist* ("to stand firm") |
| *contrast* ("to stand against") | *post* ("piece of wood that 'stands' vertically") |
| *desist* ("to stand, stop") | *rest* ["(to cause) to stand back, remain"] |
| *exist* ("to stand out, be perceptible") | |

The list of cognates goes on and on. Consider these interesting relationships:
Instead of carrying possessions or supplies, we put them in a fixed place: a *store* (or *storage*).
We can "restore" in various ways. The Latin *restaurabo* ("I shall restore") gave us *restaurant.*
Oars are used to direct a rowboat, but sometimes a fixed board is added to set a fixed course: to *steer.* The place for steering is the *stern.* The side the steerboard is placed on gave us *starboard.*
A day on which the sun *(sol)* seems to stand still is the *solstice.*

Most of the examples above come from these fascinating and information-filled texts:
Shipley, Joseph. *The Origins of English Words.* The Johns Hopkins Press: Baltimore & London, 1984.
*The Oxford English Dictionary*.

Other possible sources:
Ayto, John. *Dictionary of Word Origins.* Little, Brown and Company: New York, 1990.
Claiborne, Robert. *The Roots of English.* Random House: New York, 1989.

Watkins, Calvert. *The American Heritage Dictionary of Indo-European Roots.* Houghton Mifflin: Boston, 2000.
Geoffrey Sampson

**What was the earliest ancestor of English like?**
(`Say something in Proto-Indo-European')
Languages and language families

Languages change over the centuries - the Old English (or `Anglo-Saxon') out of which modern English has evolved over the past millennium is recognizably related to present-day English, but it is so different that if people were still speaking it somewhere we would certainly count it as a separate language. We could not understand them without a course of language lessons.
Old English takes us a little over one thousand years back, and it is the earliest ancestor-language of modern English that had a written form. If we are willing to accept partial information, though, we can get far further back than that.
This is because languages belong to *families*. After a language has spread over a sizeable territory, particularly in pre-modern conditions in which travel and communication are limited, the largely random changes which happen to all languages everywhere will be different from area to area. After a time, what began as local dialects will diverge into separate languages. So, for instance, French, Italian, Spanish, Portuguese, and Rumanian all trace their ancestry back to Latin. This is a special case, because the mother language was already a written language associated with a high civilization: consequently, many of us spent years of our lives learning it at school. In other cases, the mother language was not a written language, and we do not know what its speakers called it (if indeed they had a clear concept of their language as a namable thing alongside other languages); but we can form a fairly clear picture of the mother language by comparing the earliest recorded forms of the daughter languages.
So, for instance, English, together with Dutch, German, Danish, Norwegian, Swedish, and Icelandic, go back to a language which we now call `Proto-Germanic', which was a living spoken language about the same time as Latin - broadly two thousand years ago - but which, unlike Latin, was not written.
(Of course, a language can influence another language without being its ancestor. English does not descend from Latin, but we have a huge number of words in English that derive from Latin: that is because, until very recently, learning Latin was a basic part of a European's education, so when new words needed to be coined it was natural for educated people to reach back to Latin as a source. Also, in the year 1066, England was conquered by French-speaking Normans, and for several centuries their dialect of French became the language of government in England; consequently many words flowed from French into English. But there is no difficulty in separating out the layers of Latin- and French-derived vocabulary, which are borrowings, from the `native' English vocabulary, which goes back to Old English and through that to Proto-Germanic.)

The several thousand languages now spoken in the world comprise many different families. Most linguists believe that language did not originally arise in just one place but independently among different human groups - or at least, that if all present-day languages do ultimately derive from a single shared ancestor, that ancestor language must have lain so far back in the past, and the various descendant languages must have changed so massively, that data available to us now could never show such relationships. English, Chinese, and Swahili, to take three examples, look entirely different from one another (except for borrowings that occurred in historical times - *tea* certainly comes from Chinese, because the word came to Europe with the thing). So far as anyone knows or is ever likely to know, these languages are indeed unrelated at even the remotest level.
The Indo-European language family

But English belongs to a family much wider than just the Germanic languages. Germanic - the set of languages descended from Proto-Germanic - is one branch of the `Indo-European' language family. `Italic', which covers Latin and its descendants together with a few obscure related languages of ancient Italy, is another branch (the modern languages descended from Latin are called the `Romance' languages). Altogether, the Indo-European family has about twelve main branches at this level, Germanic and Italic being two of these. The family includes almost all languages of Europe (Hungarian is one of the exceptions), together with various languages of Western Asia and Northern India. About half of the world's present population speak some IE (Indo-European) language as their mother tongue. All these languages are held to derive ultimately from a single ancestor language, which we call `Proto-Indo-European', or PIE for short.

In a sense, this is not a mere hypothesis. Saying that various languages are related to one another *means* that they share a common ancestor language. Anyone who doubted that PIE really existed would be saying that some modern languages which we take to be related are not genuinely related languages. But we think the modern data show that all these languages, including for instance Albanian, Persian, and Hindi, definitely are related languages, so we believe that there must have been a PIE language, once.

Although PIE is very remote from us in time, and there are obviously nothing like written records of even the most fragmentary kind, linguists working over the past two hundred years, since these issues were first recognized, have reconstructed quite a lot of facts about what the language was like. They have triangulated from the various descendant languages. Broadly, the logic runs: `If such-and-such aspect of language structure is like *this* in subfamily A, like *that* in subfamily B, and like *that* in C, what single kind of earlier structure could plausibly have developed in each of those directions?' Many aspects of the reconstruction are tentative or controversial - that's science for you; but the reconstruction is reasonably solid, it is certainly much more than just a collection of speculations and wild stabs in the dark.

In the past, this material has been scattered in obscure books and academic journal articles. It has recently been assembled in one place, namely a large and excellent book edited by J.P. Mallory and D.Q. Adams, *Encyclopedia of Indo-European Culture*, Fitzroy Dearborn Publishers, London and Chicago, 1997. (NB that the spelling `Encyclopedia' is not an error, Americans spell it that way.) My account here will draw heavily on Mallory and Adams's compilation.

The best guess at when PIE was spoken puts it at something like six thousand years ago, give or take a millennium or so. There is much controversy about *where* it was spoken. For a long time the most usual answer was the Southern Russia/Ukraine region, but nowadays this is just one theory among others.

## Was PIE `primitive'?

It is important to grasp that PIE is not anything like `the first human language', or even `the original ancestor of our languages'. Language, in Europe and everywhere else in the world, undoubtedly has existed far, far longer than just six thousand-odd years. PIE is simply the earliest language it is possible to reconstruct from the evidence of modern and recent IE languages. If the entire linguistic family tree of which it is a part could be spread out to view, PIE would be one node near the bottom of the entire tree, distinguished by the fact that many of the branches it dominates reach right down to the present day. There would be masses of branching higher up in the tree, but the branches which did not lead down to the PIE node would eventually end in twigs suspended in mid-air - representing languages which became extinct too long ago for us to know anything about them.

Nevertheless, PIE is sufficiently old that it may possibly have had properties that would make it seem not just `different' but somewhat `primitive', if we could encounter it as an actual spoken language today. Nobody would expect PIE to have had words for `television' or `banana' - obviously. But, more interestingly, Mallory and Adams point out for instance that the PIE word for `nine' seems to derive from the word for `new'; they suggest that `nine' may originally have been called `the new number', implying that having a name for such a big number ranked for PIE speakers as a whizzy technological breakthrough. (In English, the pronunciation of these two words

has developed rather differently, but notice that in German *neun* and *neu* are closer, and in French *neuf* has both meanings.)

For a long time, it has been suspected that PIE may have been structurally simple, relative to present-day languages, in ways that go deeper than lack of particular vocabulary items. More than a hundred years ago, Eduard Hermann argued that PIE may have had no complex sentences: all utterances would have been strings of simple clauses, with no clause subordination. Instead of saying things like `When he saw the stone he wanted, he shouted out', PIE speakers might have said things more like `He saw a stone. He wanted that stone. Then he shouted out.'

In the closing decades of the 20th century, this and similar ideas were widely rejected, not so much because of factual evidence but for ideological reasons. Many linguists wanted to think of all human languages as equal. They disliked the suggestion that languages could be ranked as more or less evolved.

However, a careful, scholarly book by Guy Deutscher (*Syntactic Change in Akkadian*, Oxford University Press, 2000) has now shown that this principle of linguistic equality is not really tenable. The most ancient languages which were recorded in writing had very limited systems of grammatical subordination; some languages spoken by simpler, tribal societies today demonstrably are less evolved than modern European languages in this respect. So it does seem quite possible that Hermann's suggestion about PIE may have been correct.

## A reconstructed specimen of PIE

The best way to show what PIE was like is to say something in it. The language is reconstructed well enough that scholars have felt reasonably confident in assembling little specimens of PIE prose. One such specimen is based on a short extract from Old Indic literature. This is material that was transmitted from generation to generation by word of mouth before first being written down, and it may represent the earliest genre of literary composition recorded in any IE language. S.K. Sen picked a simple passage in which the Old Indic vocabulary is known to correspond to PIE roots rather than later neologisms, and took a consensus view from the experts on what the passage would look like if the Old Indic structures and sounds were rolled back two or three further millennia to their PIE antecedents. Here is the passage in English translation:

Once there was a king. He was childless. The king wanted a son.
He asked his priest: `May a son be born to me!'
The priest said to the king: `Pray to the god Varuna'.
The king approached the god Varuna to pray now to the god.
`Hear me, father Varuna!'
The god Varuna came down from heaven.
`What do you want?' `I want a son.'
`Let this be so', said the bright god Varuna.
The king's lady bore a son.

When Mallory and Adams print the PIE version offered by S.K. Sen, E.P. Hamp, and others, they use the spelling system traditional among Indo-Europeanists for representing PIE sounds. This involves many letters, which contain accents and other diacritic marks, sometimes two on one letter. It is impossible to display these in an HTML Web page (even in the English translation above, I had to miss out the dot which should appear under the `n' of the name Varuna, as transcribed from Sanskrit). But in any case, this traditional spelling system looks peculiar and offputting to English-speaking readers, and it is more traditional than exact - some aspects of it represent ideas which 19th-century researchers had about PIE that everyone now agrees were probably mistaken.

I have preferred to use different spelling conventions for PIE, which capture the same information about pronunciation, but represent it using ordinary letters and letter-combinations that look as `normal' as possible. They cannot look *very* normal; this was a language extremely different from English and from the languages that English has borrowed vocabulary from, and it had sounds that in some cases were quite unlike the sounds we are familiar with.

I shall explain the spelling system used here shortly. But first, here is the passage about the childless king, as it might have been uttered by a PIE speaker - by one of our linguistic ancestors, some six millennia ago:

To rÈecs Èhest. So nnputlos Èhest. So rÈecs s™hnum Èwelt.
SÛ tÛso gceutÈrmm prrcscet: `S™hnus moi ccnnhyotaam!'
So gceutÈer tom rÈeccmm Èweuqet: `Ihkkeswo tteiwÛm WÈrunom'.
So rÈecs tteiwÛm Werunom h™po-sesore nu tteiwÛm ihkketo.
`CludÌ moi, phhter Werune!'
TteiwÛs WÈrunos kmmta ttiwÛs Èqqeht.
`QÌtt welsi?' `WÈlmi s™hnum.'
`TÛtt hÈstu', wÈuqet loukÛs tteiwos Werunos.
ReeccÛs pÛtnih s™hnum kkekkonhe.

(For four of these words, some of the scholars consulted included an extra sound, because of - for instance - different ideas about which verbal inflexion would have been used in a particular context. I certainly am not qualified to adjudicate such issues, so I have arbitrarily chosen the shortest alternative in each case.)

The word for `priest', *gceutÈer*, literally meant `pourer'; speakers of the early IE languages seem to have seen priests as men who poured libations to the gods.

**Some recognizable words**
Although this passage looks very queer at first sight, if you know a few bits and pieces of recent European languages you can quickly make links with some of the words. The second word, *rÈecs* for `king', for instance, is identical to the Latin word for king, *rex*. Latin happened to write the combination of sounds *cs* with a single letter X, as we do in English, but that is just a convention of writing (a rather irrational one). The PIE word has a double E, meaning that the vowel is long rather than short; Latin spelling did not distinguish long from short vowels, but spoken Latin did, and *rex* had a long E, not a short E. So far as we can tell, this particular word did not change at all between PIE and Latin.

In the third sentence, *s™hnum* for `son' shows a resemblance with English - and not with Latin, where the equivalent word, *filium*, is based on a quite different root. (Note that in PIE the verb usually came at the end, so in word-for-word translation this sentence runs `The king son wanted'.) In this case it is believed that the Germanic branch of the IE language family preserved the original PIE word, while the Latin branch happened to replace it with a different word (Latin *filius*, *filia* for `son', `daughter' may possibly derive from a root meaning `suck', offspring being thought of as sucklings). On the other hand, the relationship with Latin comes out in the variation of endings between *s™hnum* in `(wants) a son', object, and *s™hnus* in the next line, `(may) a son (be born)', subject - in Latin these words would be *filium*, *filius*, respectively.
What about the peculiar-looking *kmmta*, `down', in the sixth line? I am using *mm* to represent what phoneticians call a `syllabic m', an `m' which functions as the vowel of its syllable even though it is really a consonant sound. For instance, if English used this spelling system, we would write `fathom' as `fathmm' - the second syllable has no true vowel, the `m' sound acts as if it were a vowel. PIE had a lot of these syllabic consonants. This particular word corresponds to a word of Greek which is to some extent familiar to all of us, even if we have not studied Greek. A `catastrophe' was in Greek a `turning-down', `catarrh' was a `down-flow': the word *kata* was the Greek for `down'. The Greek language had no syllabic consonants; its syllables all contained true vowels, because sound-changes on the road from PIE to Ancient Greek turned the various syllabic consonants into vowels or vowel-consonant combinations. The suggestion is that in Greek *kata*, the first A is a vowel, which used to be a consonant, while the second A is a vowel that always was a

vowel. By the time the Greeks got hold of the alphabet, of course, they had no knowledge that long before their time the two A sounds had been different, so they wrote them the same.

**The spelling system**

Now to explain the PIE spelling in a bit more detail. The first point to note is that reconstructions of PIE sounds are sometimes more like abstract tokens than concrete descriptions of pronunciation. That is: if three PIE words are each reconstructed as beginning with t-, say, then this means we are fairly sure they all began with the same sound as one another; but we cannot be sure exactly how that sound was pronounced. In most cases we do know something, though. If the sound is spelled with t-, we are rather sure that the PIE sound was much more like an English T than an English F, say.

The weakest aspect of the reconstructed pronunciation has to do with the sounds I have transcribed with the letter H. PIE is believed to have had four different sounds called `laryngeals' - sounds made somewhere in the back of the mouth or throat. One of these may have been like an English H. Some of the others may have been like the sounds of modern Arabic, which are transcribed into our alphabet using apostrophes, reversed apostrophes, or the figure 9. We have hardly any clues about the precise phonetic value of the laryngeals, because they dropped out long ago from almost all the recent IE languages - they are inferred largely from the effect they had on neighbouring vowels. (When *h* occurs in Germanic words, it derives from a *k*-like sound in PIE, not from any of these laryngeals.) When we know so little about them, there seems no point in representing the laryngeals differently, so I have simply written them all alike as H.

Double vowels, as we have seen, represent long vowels. PIE had separate long and short vowels: *ii* versus *i*, *aa* versus *a*, and so on. Thus, *i* and *ii* may have been roughly similar to the vowels of English `sit' and `seat' respectively.

In the case of the `continuant' consonants *r l m n h*, doubling of the letter means that the consonant was syllabic, in the sense discussed above: *ll*, *nn* were like the second syllables of English `bottle', `button'. Even the PIE laryngeals could be syllabic, so for instance the word for `father' is *phhter*. (One guess at what a syllabic laryngeal sounded like is the obscure vowel or `shwa' which British English speakers write as *er* and Americans write *uh*.)

With so-called `stop consonants' such as *k t*, doubling means something different again. In English, these consonants come in pairs - *t* versus *d*, *p* versus *b*, *k* versus *g* - which phoneticians call voiceless and voiced respectively. PIE had a three-way rather than two-way contrast, which I am showing as *t* : *tt* : *d*, *p* : *pp* : *b*, and so on. We do not know just how these three types of stop were made. One theory is that *tt*, *pp*, and so on were `ejectives', produced with a closed glottis. That would give a perfectly plausible three-way system - modern Amharic, a non-IE language spoken in Ethiopia, is like that. But there are many other possibilities: Korean is another present-day language which has three classes of stop consonant, distinguished along rather different lines. In traditional Indo-European studies the sounds I am writing as *d*, *b*, etc. were written *dh*, *bh*, and were taken to be `voiced aspirates', such as occur in some current Indian languages. All we really know is that PIE triples such as *t*, *tt*, *d* represent three different `manners of articulation', but we don't know just what they were.

In the case of *t d p b* we are fairly sure that the `place of articulation' was similar to the English sounds written with those letters. But in the k/g area, PIE had distinctions that we have not got in English. I am arbitrarily using the letters *c*, *k*, *q* to represent three varieties of *k*-like sound. The sound written *q*, we are fairly sure, was like `kw' run together as a single sound - in PIE *q* was different from *cw* as a sequence of two sounds. It may be that *c* was the English c/k sound, while *k* was a sound like the Arabic `uvular' stop that occurs at the beginning of a name like (the Sheikhdom of) Qatar. Or (this was the traditional Indo-Europeanists' view) *k* may have been the ordinary English sound, while *c* represented `ky' run together as a single sound.

Whatever precise sounds are represented by *c k q*, each of these also enters into a three-way contrast parallel to *t tt d*: so there are sounds spelled *cc kk qq*, and also sounds which I have spelled *gc gk gq*.

(Our alphabet has not got three G-like letters in the way it has three K-like letters, so that is the best I can do.)

In case you are beginning to think that reconstructed PIE had a suspiciously large range of different sounds, notice that while there are many different stops, in some other areas there are *fewer* sounds than a typical modern European language possesses. Most present-day European languages have a series of `fricative' sounds - English *s, th, f, v,* and so on. PIE is believed to have had just the one, *s*. On a world scale, the sound pattern reconstructed for PIE does not look unusually complex or implausible. It is not very similar to patterns found in present-day European languages, but then these are fairly different from one another (for instance, French has a range of nasal vowels, English has none). Anyway, would it not be surprising if there had been no dramatic changes over a period of six millennia?

Letters that I have not mentioned were pronounced more or less as you would guess. The acute mark on some syllables shows that those syllables carried an accent of some kind - possibly, they were said on a relatively high pitch.

**Another sample: Schleicher's tale**

The passage shown above is not the only attempt that has been made to reconstruct a piece of PIE as she was spoke. The classic attempt at such an exercise was done in 1868 by the German linguist August Schleicher. Schleicher was the first man to produce a family-tree diagram for the interrelationships among the IE subfamilies (people disagree with details of his tree structure nowadays, but it seems to have been broadly along the right lines). He was also renowned for bringing linguistics into relationship with Darwin's ideas in biology. The concept of diversity of modern species resulting from descent with gradual modifications from a common ancestor had come to the fore rather earlier in language studies than in biology. When Schleicher read Darwin's *Origin of Species* he became excited by the parallels; Schleicher argued that the two domains were closer to one another than one might suppose. He urged that languages should be seen as true living organisms alongside plants and animals, not just metaphorically but in sober reality. In the event this idea did not survive, but it was the kind of `honourable error' which can sometimes be more intellectually interesting than the writings of other scholars whose ideas are not original enough to get rejected.

Rather than translate an existing piece of prose into PIE, what Schleicher did was make up a little story of his own, which allowed him to choose vocabulary whose PIE equivalents he knew. The details of PIE were much less fully worked out in Schleicher's day than they have been since, and his PIE rendering of his tale looks off-beam now. It gave what would nowadays be seen as excessive weight to the evidence from the Indo-Iranian subfamily as against all the other branches of IE. The same passage has been reworked several times by later scholars, as IE research has progressed. Again, I shall quote Mallory and Adams's up-to-date version, using the same spelling system as before.

**In English the story runs:**

On the mountain a sheep that had no wool saw horses - one pulling a heavy waggon, one a great load, and one swiftly carrying a man.

Then the sheep said to the horses: `It pains my heart to see a man driving horses'.

Then the horses said: `Listen, sheep: it pains our heart to see man, the master, making himself a warm garment from sheep's wool, when the sheep has no wool'.

On hearing this, the sheep fled into the plain.

Not quite so artistically satisfying as the childless-king passage, perhaps, but so be it. (Incidentally, Schleicher's tale did include some subordinate clauses: *... that had no wool, ... to see a man*. Schleicher composed his tale years before the paper of Eduard Hermann, mentioned earlier, was published; if Hermann was right, these particular features of Schleicher's reconstruction must have been mistaken.)

**Here is a PIE version:**
QqrrhÈei hÛwis, qÈsyo wllhnÈh ne est, hÈcwons spÈcet, hoinom gke qqrrh™m wÛgcom wÈgcontmm, hoinom-qe mÈcchmm bÛrom, hoinom-qe gcmÈnmm hÛocu bÈrontmm.
HÛwis tu hecwoibos weuqÈt: `CÈer hegknutÛr moi, hÈcwons hÈccontmm hnÈrmm wittnntÈi'.
HÈcwoos tu weuqÛnt: `CludÌ, hÛwei, cÈer gke hegknutÛr nnsmÈi wittnntbÛs: hnÈer, pÛtis, hÈwyom rr wllhnÈhm sebi qrrnÈuti nu gqÈrmom wÈstrom; nÈgci hÈwyom wllhnÈh hÈsti'.
TÛtt cecluwÛos hÛwis hÈccrom bukkÈt.

**The familiar beneath the strange**
It is perhaps a pity that my chosen spelling system requires this passage to begin with a spectacularly weird sequence of letters in *qqrrhÈei*, `on [the] mountain'. But it is not as unpronounceable as it looks. The initial consonant is the kw-said-as-one-sound, pronounced as an ejective (or whatever type of articulation the double consonants corresponded to), and it is followed by a syllabic *r*.
Many readers will know that the Balkan state which we call Montenegro, Italian for `black mountain', is called by its own people Cherna Gora - *gora* is the standard Slavonic word for `hill, mountain'. The Slavonic *g* corresponds to the PIE ejective *c*, and the PIE w-colouration explains why, when the syllabic *r* turned into a vowel + consonant sequence, the vowel that appeared was *o* (why `mountain' is *gora* rather than *gera*, say). The vowel of the second syllable is an inflexional ending for the meaning `*on* the mountain' - in Slavonic languages today, *gora* ends in *a* only when it is subject of the verb, and the *a* changes to something like an *e* (depending on which Slavonic language we are talking about) to show `place where'.
It would be tedious to go through the entire passage in this way. But notice the word for `heart', *cÈer* - in some circumstances it has an extra sound, *cÈertt*. From French one can easily see the link with *coeur* (in the intermediate language, Latin, the word was *cor*, or in inflected forms `*cord-*' - as in `cordial', which originally meant `hearty'). In Greek, `heart' is *kardia* - a heart specialist is a cardiologist. And remember that I mentioned earlier that in the Germanic subfamily, *k*-like sounds in PIE come out as *h*. Our word *heart*, too, was originally this same PIE word *cÈer(tt)*.
The Latin for `horse' was *equus*, as in English `equine', `equitation'. It is easy to recognize this root in *hÈcwons* (first line of the passage), given that the PIE laryngeals dropped out in descendant languages. The fact that this particular root shows up in many branches of IE, demonstrating that PIE speakers knew what horses were, has been a major factor in attempts to locate the PIE homeland (archaeology shows that horses were much less widely distributed six thousand years ago than they are today).
In *hecwoibos*, `to the horses' (second paragraph), Latinists will recognize a form of the distinctive Latin case ending *-ibus* for the dative plural - used for people or things that one speaks to, gives something to, etc. When I was at school, though, I certainly would not have got high marks for attaching the *-ibus* ending to the root of *equus*. `To the men' was *hominibus*, but translating `to the horses' as *equibus* would have been a major clanger. That noun belonged to a different `declension', meaning that it took a different set of endings to express the same meanings. (There were five regular declensions, plus irregular cases.) In PIE, apparently, all nouns were inflected in more or less the same way, and diversity of declensions was a feature that developed as Latin evolved out of PIE - greatly to the sorrow of generation after generation of inky-fingered English schoolboys.
As far back as we can go. Undoubtedly, many details of these reconstructions would turn out to need modification, if (impossibly) a PIE speaker could return to life and show us how his language *really* worked. On the other hand, I am not sure than any of the world's other language families provide data allowing even a tentative reconstruction of an ancestor language reaching as far back into the past as PIE.
This is not the earliest language spoken by Man - far from it. But it is probably the earliest we shall ever see.

Geoffrey Sampson (*last changed 24 May 2002*)