Indo-European languages and branches

Language Relations

One of the first hurdles anyone encounters in studying a foreign language is learning a new vocabulary. Faced with a list of words in a foreign language, we instinctively scan it to see how many of the words may be like those of our own language.We can provide a practical example by surveying a list of very common words in English and their equivalents in Dutch, Czech, and Spanish. A glance at the table suggests that some words are more similar to their English counterparts than others and that for an English speaker the easiest or at least most similar vocabulary will certainly be that of Dutch. The similarities here are so great that with the exception of the words for 'dog' (Dutch hond which compares easily with English 'hound') and 'pig' (where Dutch zwijn is the equivalent of English 'swine'), there would be a nearly irresistible temptation for an English speaker to see Dutch as a bizarrely misspelled variety of English (a Dutch reader will no doubt choose to reverse the insult). When our myopic English speaker turns to the list of Czech words, he discovers to his pleasant surprise that he knows more Czech than he thought. The Czech words bratr, sestra,and syn are near hits of their English equivalents. Finally, he might be struck at how different the vocabulary of Spanish is (except for madre) although a few useful correspondences could be devised from the list, e.g. English pork and Spanish puerco.

The exercise that we have just performed must have occurred millions of times in European history as people encountered their neighbours' languages.

Table 1.1. Some common words in English, Dutch, Czech, and Spanish

English	Dutch	Czech	Spanish
mother	moeder	matka	madre
father	vader	otec	padre
brother	broer	bratr	hermano
sister	zuster	sestra	hermana
son	zoon	syn	hijo
daughter	dochter	dcera	hija
dog	hond	pes	perro
cow	koe	kra´va	vaca
sheep	schaap	ovce	oveja

pig	zwijn	prase	puerco
house	huis	dum	casa

The balance of comparisons was not to be equal, however, because Latin was the prestige language employed both in religious services and as an international means of communication. A medieval monk in England, employing his native Old English, or a scholar in medieval Iceland who spoke Old Norse, might exercise their ingenuity on the type of wordlist displayed in Table 1.2 where we have included the Latin equivalents.
The similarities between Latin and Old English in the words for 'mother', 'father', and 'pig', for example, might be explained by the learned classes in terms of the influence of Latin on the other languages of Europe. Latin, the language of the Roman Empire, had pervaded the rest of Europe's languages, and someone writing in the Middle Ages, when Latin words were regularly being imported into native vernaculars, could hear the process happening with their own ears. The prestige of Latin, however, was overshadowed by that of Greek as even the Romans acknowledged the antiquity and superior position of ancient Greek. This veneration for Greek prompted a vaguely conceived model in which Latin had evolved as some form of degraded Greek. Literary or chronological prestige then created a sort of linguistic pecking order with Greek at the apex and most ancient, then the somewhat degenerate Latin, and then a series of debased European languages that had been influenced by Latin.

Comparable words in Old English, Old Norse, and Latin

English	Old English	Old Norse	Latin
mother	modor	moira	mater
father	fader	faðir	pater
brother	broþor	broðir	frater
sister	sweostor	systir	soror
son	sunu	sunr	fılius
daughter	dohtor	dottir	fılia
dog	hund	hundr	canis
pig	swın	svın	suınus
house	hus	hus	domus

What about the similarities between Old English and Old Norse? Our English monk might note that all ten words on the list appeared to correspond with one another and in two instances the words were precisely the same ('pig' and 'house'). We have no idea whether any Englishman understood why the two languages were so similar. But in the twelfth century a clever Icelandic scholar, considering these types of similarities, concluded that Englishmen and Icelanders 'are of one tongue, even though one of the two (tongues) has changed greatly, or both somewhat'. In a wider sense, the Icelander believed that the two languages, although they differed from one another, had 'previously parted or branched off from one and the same tongue'. The image of a tree with a primeval language as a trunk branching out into its various daughter languages was quite deliberate—the Icelander employed the Old Norse verb greina 'to branch'. This model of a tree of related languages would later come to dominate how we look at the evolution of the Indo-European languages.

The similarities between the languages of Europe could then be accounted for in two ways: some of the words might be explained by diffusion or borrowing, here from Latin to the other languages of Europe. Other similarities might be explained by their common genetic inheritance, i.e. there had once been a primeval language from whence the current languages had all descended and branched away. In this latter situation, we are dealing with more than similarities since the words in question correspond with one another in that they have the same origin and then, as the anonymous Icelander suggests, one or both altered through time.

Speculation as to the identity of the primeval language was largely governed by the Bible that provided a common origin for humankind. The biblical account offered three decisive linguistic events. The first was the creation of Adam and Eve that provided a single ancestral language which, given the authority and origin of the Bible, ensured that Hebrew might be widely regarded as the 'original' language from which all others had descended.

The Indo-European Languages

The idea of an 'Indo-European' family of languages grew out of the discovery that the oldest language of the Indian subcontinent, Sanskrit, was related to the European languages. The discovery of Sanskrit provided the key which opened the door to the possibility of comparing the Indo-European languages with each other. Sanskrit was helpful in a number of ways: it was older than all other known languages (its oldest text goes back to before 1000 B.C.), and it was relatively transparent because its forms could be easily analyzed: the original structure of its forms was well-preserved. In Greek, on the other hand, the inherited sounds *s*, *i̯* and *u̯* had disappeared at an early stage, followed by the contraction of adjacent vowels which masked the structure of the original forms. A consequence of the transparent structure of Sanskrit, as opposed to Greek, was that the Sanskrit grammarians had been able to describe the way its forms were constructed: this proved to be of enormous importance for the work of Western scholars.

In 1498 Vasco de Gama discovered the sea route to India, and it was not long after that Europeans began to settle there. They quickly heard about Sanskrit, the holy language of India, which was comparable in many respects with regard to its social position to Latin in Europe in the Middle Ages. Almost immediately, in the period between 1500 and 1550, it was noticed that there were close similarities between individual Sanskrit words and the words of the languages of Europe. As knowledge of Sanskrit increased, such relationships were more frequently noticed. It was Sir William Jones who, in 1786, publicly acknowledged this relationship and correctly explained it. He was the Chief Magistrate of Calcutta, the capital of English India, and founder of the Asiatic Society, which encouraged scholarly research into all aspects of Indian culture and history. In a speech given to the Society, he said:

The Sanskrit language, whatever be its antiquity, is of a wonderful structure; more perfect than the Greek, more copious than the Latin, and more exquisitely refined than either, yet bearing to both of them a stronger affinity, both in the roots of verbs and in the forms of grammar, than could possibly have been produced by accident; them to have sprung from some common source, which, perhaps, no longer exists: there is similar reason, though no quite so forcible, for supposing that both the Gothic and the Celtic, though blended with a different idiom, had the same origin with

theSanskrit; and the old Persian might be added to the same family, if this were the placefor discussing any question concerning the antiquities of Persia.

The reasoning that we see here assumes that so great a number of similarities cannot be explained by the borrowing of words between languages, and that it is therefore more likely that the languages in question must all have a common ancestor which relates them to each other. This analysis goes back to Van Boxhorn and had been passed on by a number of Dutch and English scholars before Jones, but the latter's authority was such that his statement is considered to mark the birth of Indo-European linguistics.
In his speech Sir William Jones did not go into further detail. For this reason we will look at a more extensive report on the subject that was prepared by the French priest Coeurdoux in 1767 (but which was not published until 1808 because the learned scholar who received it failed to realize its value!). Coeurdoux compared words with each other (his spelling of the Sanskrit words is not completely accurate), as for example:

Sanskrit:	*devah* 'god'	Latin *deus*	Greek *theos*
	padam 'foot'	*pes*, *ped-is*	*pous*, *pod-os*
	maha 'great'		*megas*
	viduva 'widow'	*vidua*	

These similarities seem to be quite obvious. Yet *theos* does not belong in the list: not everything which seems to be self-evident and trustworthy is therefore true. And this mistake is not only made by beginners!
But Coeurdoux was not satisfied with words alone. He noticed that Sanskrit had a dual number (a separate form next to the plural for groups of two), just as Greek had; that the numerals were basically the same, as well as the pronouns, the negating prefix *aa•*, and the verb 'to be'.
His list of similarities certainly shows insight, but not everything in it is correct. We must not forget that comparative linguistics did not yet exist! The dual does not provide an argument, because there are many languages in the world which have it (this was unknown at the time), so that what seemed to be an exceptional similarity is not really that exceptional after all. Moreover, for similarities to really count, they must take into consideration the form as well as the meaning of words: it is as if one should note that English makes use of an article and language X does, too, and conclude that a genetic relation must therefore exist between them, without paying attention to the *form* of the article in question.

Thus by 1800 a preliminary model for the relationship between many of the languages of Europe and some of those of Asia had been constructed. The language family came to be known as Indo-Germanic (so named by ConradMalte-Brun in 1810 as it extended from India in the east to Europe whose westernmost language, Icelandic, belonged to the Germanic group of languages) or Indo-European (Thomas Young in 1813). Where the relationships among language groups were relatively transparent, progress was rapid in the expansion of the numbers of languages assigned to the Indo-European family. Between the dates of the two early great comparative linguists, Rasmus Rask (1787–1832) and Franz Bopp (1791–1867), comparative grammars appeared that solidified the positions of Sanskrit, Iranian, Greek, Latin, Germanic, Baltic, Slavic, Albanian, and Celtic within the Indo-European family. Some entered easily while others initially proved more difficult. The Iranian languages, for example, were added when comparison between Iran's ancient liturgical texts, the Avesta, was made with those in Sanskrit. The similarities between the two languages were so great that some thought that the Avestan language was merely a dialect of Sanskrit, but by 1826 Rask demonstrated conclusively that Avestan was co-ordinate with Sanskrit and not derived from it. He also showed that it was an earlier relative of the modern Persian language.

The Celtic languages, which displayed many peculiarities not found in the classical languages, required a greater scholarly effort to see their full incorporation into the Indo-European scheme. Albanian had absorbed so many loanwords from Latin, Greek, Slavic, and Turkish that it required far more effort to discern its Indo-European core vocabulary that set it off as an independent language.

After this initial phase, which saw nine major language groups entered into the Indo-European fold, progress was more difficult. Armenian was the next major language to see full incorporation. It was correctly identified as an independent Indo-European language by Rask but he then changed his mind and joined the many who regarded it as a variety of Iranian. This reticence in seeing Armenian as an independent branch of Indo-European was due to the massive borrowing from Iranian languages, and here the identification of Armenian's original Indo-European core vocabulary did not really emerge until about 1875.

The last two major Indo-European groups to be discovered were products of archaeological research of the late nineteenth and early twentieth centuries.

Western expeditions to oasis sites of the Silk Road in Xinjiang, the westernmost province of China, uncovered an enormous quantity of manuscripts in the first decades of the twentieth century. Many of these were written in Indic or Iranian but there were also remains of two other languages which are now known as Tocharian and by 1908 they had been definitely shown to represent an independent group of the Indo-European family.

It was archaeological excavations in Anatolia that uncovered cuneiform tablets which were tentatively attributed to Indo-European as early as 1902 but were not solidly demonstrated to be so until 1915, when Hittite was accepted into the Indo-European fold. Other Indo-European languages, poorly attested in inscriptions, glosses in Greek or other sources, or personal and place names in classical sources, have also entered the Indo-European family. The more important are Lusatian in Iberia, Venetic and Messapic in Italy, Illyrian in the west Balkans, Dacian and Thracian in the east Balkans, and Phrygian in central Anatolia.

If we prepare a map of Eurasia and depict on it the various major groups of Indo-European languages, we find that they extend from the Atlantic to western China and eastern India; from northernmost Scandinavia south to the Mediterranean and the Indian Ocean. The family consists of languages or language groups from varying periods. As we are currently painting our Indo-European world with a broad brush, we can divide the Indo-European groups into those in which there are languages still spoken today and those that are extinct. In some cases the relationship between an ancient language such as Illyrian and its possible modern representative, Albanian, is uncertain.

The map of the surviving Indo-European groups masks the many changes that have affected the distribution of the various language groups. Celtic and Baltic, for example, once occupied territories vastly greater than their attenuated status today and Iranian has seen much of its earlier territory eroded by the influx of other languages.

The map of the Indo-European languages is not entirely continuous as there are traces of non-Indo-European languages in Europe as well. Even before a model of the Indo-European family was being constructed, scholars had begun observing that another major linguistic family occupied Europe. Before 1800 the Hungarian linguist S. Gya´rmathi (1751–1830) had demonstrated that Hungarian, a linguistic island surrounded by a sea of Indo-European languages, was related to Finnish (Hungarian did not take up its

historical seat until the Middle Ages). He accomplished this primarily on the basis of grammatical elements, rightly realizing that vocabulary offers the least trustworthy evidence because it may be so easily borrowed. Linguists, including the irrepressible Rask, established the constituent elements of the Uralic language family. In Europe this comprises Finnish, Karelian, Lapp (Saami), Estonian, Hungarian, and a number of languages spoken immediately to the west of the Urals such as Mordvin and Mari. Its speakers also occupy a broad region east of the Urals and include the second major Uralic branch, the Samoyedic languages.

The Caucasus has yielded a series of non-Indo-European languages that are grouped into several major families. Kartvelian, which includes Georgian in the south and two northern varieties, Northern and North-Eastern Caucasian, both of which may derive from a common ancestor. What has not been demonstrated is a common ancestor for all the Caucasian languages.
In Anatolia and South-West Asia Indo-Europeans came into contact with many of the early non-Indo-European civilizations, including Hattic and Hurrian in Anatolia, the large group of Semitic languages to the south, and Elamite in southern Iran. The Indo-Aryans shared the Indian subcontinent with two other language families, most importantly the Dravidian family.
The major surviving non-Indo-European language of western Europe is Basque, which occupies northern Spain and southern France. The other spoken non-Indo-European languages of Europe are more recent imports such as Maltese whose origins lie in the expansion of Arabic. There are also poorly attested extinct languages that cannot be (confidently) assigned to the Indo-European family and are generally regarded as non-Indo-European. These would include Iberian in the Iberian Peninsula and Etruscan in north-central Italy.
We have seen that speculations concerning the similarities between languages led to the concept of an Indo-European family of languages comprised of twelve main groups and a number of poorly attested extinct groups. This language family was established on the basis of systematic correspondence in grammar and vocabulary among its constituent members. The similarities were explained as the result of the dispersal or dissolution of a single ancestral language that devolved into its various daughter groups, languages, and dialects. We call this ancestral language Proto-Indo-European.

The Indo-European languages

Celtic

The Celtic languages represent one of the more attenuated groups of Indo-European. In the first centuries BC Celtic languages could be found from Ireland in the west across Britain and France, south into Spain, and east into central Europe. Celtic tribes raided the Balkans, sacked Delphi in 279 bc, and some settled in Anatolia in the same century to become the Galatians. The expansion of the Roman Empire north and westwards and the later movement of the Germanic tribes southwards saw the widespread retraction of Celtic languages on the Continent.

The Celtic languages are traditionally divided into two main groups—Continental and Insular Celtic. The Continental Celtic languages are the earliest attested. Names are found in Greek and Roman records while inscriptions in Celtic languages are found in France, northern Italy, and Spain. The Continental evidence is usually divided into Gaulish, attested in inscriptions in both southern and central France, Lepontic, which is known from northern Italy in the vicinity of Lake Maggiore, and Ibero-Celtic or Hispano-Celtic in the north-western two-thirds of the Iberian peninsula. The inscriptions are very heavily biased toward personal names and do not present a particularly wide-ranging reservoir of the Celtic language. The earliest inscriptions are in the Lepontic language. Celtic inscriptions may be written in the Greek script, modified versions of the Etruscan script, the Roman script, or, in Iberia, in a syllabic script employed by the non-Indo-European Iberians. Where the inscriptions do have value is illustrating the earliest evidence for Celtic speech in its most primitive form. This latter point is quite significant as most of the Insular Celtic languages have suffered such a brusque restructuring that many of the original grammatical elements have either been lost or heavily altered.

The evidence of Celtic

Continental Celtic
Gaulish (c. 220–1 bc)
Lepontic (c. 600–100 bc)
Ibero-Celtic (c. 200–1 bc)

Insular Celtic
Ancient British (c. ad 1–600)
Welsh
Archaic (c. ad 600–900),
Old Welsh (900–1200),
Middle Welsh (1200–1500)
Modern Welsh (1500–)
Cornish
Old Cornish (c. ad 800–1200)
Middle Cornish (1200–1575)
Late Cornish (1575–1800)
Breton
Primitive Breton (c. ad 500–600)
Old Breton (600–1000)
Middle Breton (1000–1600)
Modern Breton (1600–)
Irish
Ogam Irish (c. ad 400–700)
Old Irish (c. ad 700–900)
Middle Irish (c. ad 900–1200)
Modern Irish (1200–)

The Insular Celtic languages, so named because they were spoken in Britain and Ireland, are divided into two main groups—Brittonic and Goidelic. The first comprises the languages spoken or originating in Britain. The early British language of the first centuries BC, known primarily from inscriptions and Roman sources, evolved into a series of distinct languages—Welsh, Cornish, and Breton. Welsh developed a rich literary tradition during the Middle Ages

and the main body of Welsh textual material derives from the Middle Welsh period. Cornish, which became extinct by the end of the 18th century, yields a much smaller volume of literature, and most of our Cornish data derives from the Middle Cornish period (which also serves as the basis of the Modern Cornish revival). Breton originated in Britain and was carried from southern Britain to Brittany during the fifth to seventh centuries where, some argue, it may have encountered remnant survivors of Gaulish.

The Goidelic languages comprise Irish and two languages derived from Irish—Scots Gaelic and Manx—that were imported into their historical positions in the early Middle Ages.

From a linguistic standpoint, the most important of the Celtic languages is Old and Middle Irish, as the quantity of output for these periods was quite large (the dictionary of early Irish runs to more than 2,500 pages). There is also inscriptional evidence of Irish in Ireland dating to c. ad 400–700. These inscriptions are written in the Ogam script, notches made on the edges of an upright stone, hence the language of the inscriptions is termed Ogam Irish, and although they are largely confined to personal names, they do retain the fuller grammatical complement of the Continental Celtic inscriptions,

which presents some of the Continental and Insular inscriptional evidence compared with the equivalent words in Old Irish, indicates something of the scale of change in Old Irish compared with the earlier evidence for Continental Celtic languages.

Italic

Latin is the principal Italic language but it only achieved its particular prominence with the expansion of the Roman state in the first centuries bc. It is earliest attested in inscriptions that date from c. 620 BC onwards and are described as Old Latin. The main source of our Latin evidence for an Indo-Europeanist derives from the more familiar Classical Latin that emerges about the first century bc. The closest linguistic relation to Latin is Faliscan, a language (or dialect) spoken about 40 km north of Rome and also attested in inscriptions from c. 600 BC until the first centuries BC when the region was assimilated entirely into the Latin language.

South of Rome lay the Samnites who employed the Oscan language, attested in inscriptions, including graffiti on the walls of the destroyed city of Pompeii, beginning about the fifth century BC. There are also about two hundred other documents, usually quite short, in the Oscan language. Oscan finds a close

relation in Umbrian, which was spoken north of Rome, and, after Latin, provides the next largest corpus of Italic textual material. Although there are a number of short inscriptions, the major evidence of Umbrian derives from the Iguvine Tablets, a series of seven (of what were originally a total of nine) bronze tablets detailing Umbrian rituals and recorded between the third and first centuries BC. In addition to these major Italic languages, there are a series of inscriptions in poorly attested languages such as Sabine, Volscian, and Marsian. While these play a role in discussions of Italic languages, it is largely Latin and occasionally Oscan and Umbrian that play the greatest role in Indo-European studies.

The so-called Vulgar Latin of the late Roman Empire gradually divided into what we term the Romance languages. The earliest textual evidence for the various Romance languages begins with the ninth century for French, the tenth century for Spanish and Italian, the twelfth century for Portuguese, and the sixteenth century for Romanian. As our knowledge of Latin is so extensive, comparative linguists rarely require the evidence of the Romance languages in Indo-European research.

The evidence of the Italic languages

Latin-Faliscan
Latin
Old Latin (c.620–80 bc)
Classical Latin (c.80 bc–ad 120)
Late Latin (ad 120–c.1000)
Faliscan (600–100 bc)
Osco-Umbrian
Oscan (500–1 bc)
Umbrian (300–1 bc)

Germanic

The collapse of the Roman Empire was exacerbated by the southern and eastern expansion of Germanic tribes. The Germans first emerge in history occupying the north European plain from Flanders in the west to the Vistula river in the east; they also occupied at least southern Scandinavia.

The Germanic languages are divided into three major groups: eastern, northern, and western. Eastern Germanic is attested by a single language, Gothic, the language of the Visigoths who settled in the Balkans where the Bible in the Gothic language (only portions of which survive) was prepared by the Christian missionary Wulfilas. This fourth-century translation survives primarily in a manuscript dated to c. ad 500. Eighty-six words of the language of the Ostrogoths were recorded in the Crimea by Oguier de Busbecq, a western diplomat to the Ottoman Empire, in the sixteenth century. Because of its early attestation and the moderately large size of the text that it offers, survives primarily in a manuscript dated to c. ad 500. Eighty-six words of the language of the Ostrogoths were recorded in the Crimea by Oguier de Busbecq, a western diplomat to the Ottoman Empire, in the sixteenth century. Because of its early attestation and the moderately large size of the text that it offers, Gothic plays a significant part of the Germanic set of languages in comparative linguistics.

The northern group of Germanic languages is the earliest attested because of runic inscriptions that date from c. ad 300 onwards. These present an image of Germanic so archaic that they reflect not only the state of proto-Northern Germanic but are close to the forms suggested for the ancestral language of the entire Germanic group. But the runic evidence is meagre and the major evidence for Northern Germanic is to be found in Old Norse. This comprises a vast literature, primarily centred on or composed in Iceland. The extent of Old Norse literature ensures that it is also regarded as an essential comparative component of the Germanic group. By c.1000, Old Norse was dividing into regional east and west dialects and these later provided the modern Scandinavian languages. Out of the west dialect came Icelandic, Faeroese, and Norwegian and out of East Norse came Swedish and Danish.

The main West Germanic languages were German, Frankish, Saxon, Dutch, Frisian, and English. For comparative purposes, the earliest stages of German and English are the most important. The textual sources of both German and English are such that Old High German and Old English provide the primary comparative evidence for their respective languages (cf. Mallory–Adams where only 23 Middle English words contribute what could not be found among the 1,630 Old English words cited). Incidentally, the closest linguistic relative to English is Frisian followed by Dutch.

Baltic

The Baltic languages, now confined to the north-east Baltic region, once extended over an area several times larger than their present distribution indicates. The primary evidence of the Baltic languages rests with two subgroups:
West Baltic attested by the extinct Old Prussian, and East Baltic which survives today as Lithuanian and Latvian.
The evidence for Old Prussian is limited primarily to two short religious tracts (thirty pages altogether) and two Prussian wordlists with less than a thousand words. These texts date to the sixteenth–seventeenth centuries and were written by non-native speakers of Old Prussian.

The evidence of the Baltic languages

West Baltic
Old Prussian (c.1545–1700)
East Baltic
Lithuanian (1515–)
Latvian (c.1550–)

The evidence for the East Baltic languages is also tied to religious proselytization and it might be noted that the Lithuanians, beginning to convert to Christianity only in the fourteenth century, were among the last pagans in Europe. Unlike Old Prussian, however, both Lithuanian and Latvian survived and have full national literatures. There is considerable evidence that Latvian spread over an area earlier occupied by Uralic speakers, and within historic times an enclave of Uralic-speaking Livonians has virtually disappeared into their Latvian environment. Although attested no more recently than Albanian, the Baltic languages, especially Lithuanian, have been far more conservative and preserve many features that have disappeared from many much earlier attested Indo-European languages. For this reason, Lithuanian has always been treated as a core language in comparative Indo-European reconstruction.

Slavic

In the prehistoric period the Baltic and Slavic languages were so closely related that many linguists speak of a Balto-Slavic proto-language. After the two groups had seen major division, the Slavic languages began expanding over

territory previously occupied by speakers of Baltic languages. From c. ad 500 Slavic tribes also pushed south and west into the world of the Byzantine Empire to settle in the Balkans and central Europe while other tribes moved down the Dnieper river or pressed east towards the Urals and beyond.
The initial evidence for the Slavic language is Old Church Slavonic which tradition relates to the Christianizing mission of Saints Cyril and Methodius in the ninth century. Their work comprises biblical translations and was directed at Slavic speakers in both Moravia and Macedonia. The language is regarded as the precursor of the earliest South Slavic languages but it also quite close to the forms reconstructed for Proto-Slavic itself. The prestige of Old Church Slavonic, so closely associated with the rituals of the Orthodox Church, ensured that it played a major role in the development of the later Slavic languages.
The Slavic languages are divided into three main groups—South, East, and West Slavic. The South Slavic languages comprise Bulgarian, Macedonian, Serbo-Croatian, and Slovenian. The earliest attestations of these languages, as distinct from Old Church Slavonic, begin about ad 1000–1100.
The East Slavic languages comprise Russian, Byelorussian, and Ukrainian, and their mutual similarity to one another is closer than any other group. Here too the prestige of Old Church Slavonic was such that the three regional developments were very slow to emerge, generally not until about 1600.
The West Slavic languages were cut off from their southern neighbours by the penetration of the Hungarians into central Europe. The language that

The evidence of the Slavic languages

South Slavic
Old Church Slavonic (c. 860–)
Macedonian (1790–)
Bulgarian
Old Bulgarian (900–1100)
Middle Bulgarian (1100–1600)
Modern Bulgarian (1600–)
Serbo-Croatian (1100–)
Slovenian (1000–)
East Slavic
Russian
Old Russian (c.1000–1600)

Russian (c.1600–)
Byelorussian (c.1600–)
Ukrainian (c.1600–)
West Slavic
Polish (c.1270–)
Czech (c.1100–)
Slovak (c.1100–)

Polish, Czech, and Slovak replaced Latin, not Old Church Slavonic.
Unlike the case with East and South Slavic, Church Slavonicisms are almost entirely absent from West Slavic.
The abundance of Old Church Slavonic material, its conservative nature, and the fact that subsequent Slavic languages appear to evolve as later regional developments means that linguists generally find that Old Church Slavonic will suffice for Indo-European comparative studies although its evidence can be augmented by other Slavic languages.

Albanian

The earliest reference to an Albanian language dates to the fourteenth century but it was not until 1480 that we begin to recover sentence-length texts and the first Albanian book was only published in 1555. The absorption of so many foreign words from Greek, Latin, Turkish, and Slavic has rendered Albanian only a minor player in the reconstruction of the Indo-European vocabulary, and of the 'major' languages it contributes the least number of Indo-European cognates. However, Albanian does retain certain significant phonological and grammatical characteristics .

Greek

The earliest evidence for the Greek language comes from the Mycenaean palaces of mainland Greece (Mycenae, Tiryns, Pylos) and from Crete (Knossos). The texts are written in the Linear B script, a syllabary, i.e. a script whose signs indicate full syllables (ra, wa, etc.) rather than single phonemes, and are generally administrative documents relating to the palace economies of Late Bronze Age Greece. With the collapse of the Mycenaean civilization in the twelfth century BC, evidence for Greek disappears until the

emergence of a new alphabetic writing system, based on that of the Phoenicians, which developed in the period c.825–750 BC. The early written evidence indicates the existence of a series of different dialects that may be assigned to Archaic Greek. One of these, the Homeric dialect, employed in the Iliad and Odyssey, was an eastern dialect that grew up along the coast of Asia Minor and was widely employed in the recitation of heroic verse. The Attic dialect, spoken in Athens, became the basis of the classical standard and was also spread through the conquests of Alexander the Great. This established the line of development that saw the later emergence of Hellenistic, Byzantine, and Modern Greek. The literary output of ancient Greece is enormous and the grammatical system of Greek is sufficiently conservative that it plays a pivotal role in Indo-European comparative studies.

The evidence of the Greek language

Mycenaean (c. 1300–1150 bc)
Greek
Archaic Greek (c. 800–400 bc)
Hellenistic Greek (c. 400 bc–ad 400)
Byzantine Greek (c. ad 400–1500)
Modern Greek (1500–)

Anatolian

The earliest attested Indo-European languages belong to the extinct Anatolian group. They first appear only as personal names mentioned in Assyrian trading documents in the centuries around 2000 BC. By the mid second millennium texts in Anatolian languages are found in abundance, particularly in the archives of the Hittite capital at Hattus in central Anatolia.

The Anatolian languages are divided into two main branches: Hittite-Palaic and South/West Anatolian. The first branch consists of Hittite and Palaic. Hittite is by far the best attested of the Anatolian languages. There are some 25,000 clay tablets in Hittite which deal primarily with administrative or ritual matters, also mythology. The royal archives of the Hittite capital also yielded some documents in Palaic, the language of the people of Pala to the north of the Hittite capital. These are of a ritual nature and to what extent Palaic was even spoken during the period of the Hittites is a matter of speculation.

It is often assumed to have become extinct by 1300 BC if not earlier but we have no certain knowledge of when it ceased to be spoken.
In south and west Anatolia we find evidence of the other main Anatolian language, Luvian. Excepting the claim that the earliest references to Anatolians in Assyrian texts refer explicitly to Luvians, native Luvian documents begin about 1600 bc. Luvian was written in two scripts: the cuneiform which was also employed for Hittite and a hieroglyphic script created in Anatolia itself.
Primarily along the south-west coast of Anatolia there was a string of lesser known languages, many if not all believed to derive from the earlier Luvian language or, if not derived directly from attested Luvian, derived from unattested varieties of Anatolian closely related to attested Luvian. These include Lycian which is known from about 200 inscriptions on tombs, Lydian, also known from tombs and some coins as well, Pisidian, which supplied about thirty tomb inscriptions, Sidetic about half a dozen, and Carian, which is not only found in Anatolia but also in Egypt where it occurs as graffiti left by Carian mercenaries.
Anatolian occupies a pivotal position in Indo-European studies because of its antiquity and what are perceived to be extremely archaic features of its grammar; however, the tendency for Anatolian documents to include many items of vocabulary from earlier written languages, in particular Sumerian and Akkadian, has militated against a comparable importance in contributing to the reconstruction of the Proto-Indo-European vocabulary. All too often we do not know the actual Hittite word for a concept because that concept is always expressed as a Sumerian or Akkadian phonogram (which the Hittite speaker would have pronounced as the proper Hittite word much in the way an English speaker says ‘pound’ when confronted with the Latin abbreviation lb).

The evidence of the Anatolian languages
Hittite-Palaic
Hittite
Old Hittite (1570–1450 bc)
Middle Hittite (1450–1380 bc)
New Hittite (1380–1220 bc)
Palaic (?–?1300 bc)
South/West Anatolian
Luvian
Cuneiform Luvian (1600–1200 bc)

Hieroglyphic Luvian (1300–700 bc)
Lycian (500–300 bc)
Milyan (500–300 bc)
Carian (500–300 bc)
Lydian (500–300 bc)
Sidetic (200–100 bc)
Pisidian (ad 100–200)

Armenian

As with many other Indo-European languages, it was the adoption of Christianity that led to the first written records of the Armenian language. The translation of the Greek Bible into Armenian is dated by tradition to the fourth century, and by the fifth century there was a virtual explosion of Armenian literature. The earliest Armenian records are in Old or Classical Armenian which dates from the fourth to the tenth century. From the tenth to nineteenth century Middle Armenian is attested mainly among those Armenians who had migrated to Cilicia. The modern literary language dates from the early nineteenth century.

As we have seen, the Armenian vocabulary was so enriched by neighboring Iranian languages—the Armenian-speaking area was regularly in and out of Iranian-speaking empires—that its identification as an independent Indo-European language rather than an Iranian language was not secured until the 1870s. It has been estimated that only some 450 to 500 core words of the Armenian vocabulary are not loanwords but inherited directly from the Indo-European proto-language.

Indo-Aryan

The ancient Indo-European language of India is variously termed Indic, Sanskrit, or Indo-Aryan. While the first name is geographically transparent (the people of the Indus river region), Sanskrit refers to the artificial codification of the Indic language about 400 bc, i.e. the language was literally 'put together' or 'perfected', i.e. samkrta, a term contrasting with the popular or natural language of the people, Prakrit. Indo-Aryan acknowledges that the Indo-Europeans of India designated themselves as Aryans; as the Iranians also termed themselves Aryans, the distinction here is then one of Indo-Aryans

in contrast to Iranians (whose name already incorporates the word for 'Aryan'). The earliest certainly dated evidence for Indo-Aryan does not derive from India but rather north Syria where a list of Indo-Aryan deities is appended to a treaty between the Mitanni and the Hittites. This treaty dates to c.1400–1330 bc and there is also other evidence of Indo-Aryan loanwords in Hittite documents.

These remains are meagre compared with the vast religious and originally oral traditions of the Indo-Aryans. The oldest such texts are the Vedas (Skt veda 'knowledge'), the sacred writings of the Hindu religion. The Rigveda alone is about the size of the Iliad and Odyssey combined and this single work only begins a tradition of religious literature that runs into many volumes. These religious texts, however, were not edited and written down until the early centuries bc, and dating the composition of the Vedas has been a perennial problem.

Most dates for the Rigveda fall within a few centuries on either side of c.1200 bc. Because of the importance of the Vedas in Indic ritual and the attention given to the spoken word, the texts have probably not suffered much alteration over the millennia. A distinction may be made between Vedic Sanskrit, the earliest attested language, and later Classical Sanskrit of the first millennium BC and more recently. Sanskrit literature was by no means confined to religious matters but also included an enormous literary output, including drama, scientific treatises, and other works, such that the volume of Sanskrit documents probably exceeds that of ancient Greece and Rome combined.

By the middle of the Wrst millennium BC we Wnd evidence for the vernacular languages of India which, as we have seen above, are designated Pra¯krit. The earliest attested Indo-Aryan documents are in Pra¯krit and these provide the bases of the modern Indo-Aryan languages, e.g. Hindi-Urdu, Gujarati, Marathi, Sinhalese.

Iranian

In the first millennium BC the distribution of the Iranian languages was truly enormous and not only comprised Iran and Afghanistan but also all of central Asia and the entire Eurasian steppe from at least the Dnieper east to the Yenisei river. The Iranian languages are divided into two major groups, Eastern and Western.

The Eastern branch is earliest attested in the form of Avestan, the liturgical language of the religion founded by Zarathustra, or Zoroaster as he was known

to the Greeks. The Avesta is a series of hymns and related material that was recited orally and not written down prior to the fourth century ad. Unlike the Rigveda, the integrity of its oral transmission was not nearly so secure and there are many difficulties in interpreting the earlier passages of the document. These belong to the Gathas, the hymns reputedly composed by Zarathustra himself; there is also much later material in the Avesta. The dates of its earliest elements are hotly disputed but generally fall c.1000 BC and are presumed to be roughly contemporary with the Rigveda.

Eastern Iranian offers many other more recently attested languages that belong to the Middle Iranian period. In central Asia, Bactrian, Sogdian, and Choresmian were all spoken and occasionally recorded from about the fourth century ad onwards until the Turkish conquest of the region. The European steppelands were occupied by the nomadic Scythians in the west and the Saka in the east, and what little evidence survives indicates that these all spoke an East Iranian language as well. The Saka penetrated what is now western China and settled along the southern route of the Silk Road in the oasis town of Khotan where they have left more abundant documents known as Khotanese Saka. Most of these East Iranian languages have disappeared except for those spoken by peoples who occupied mountainous regions and have survived into the New Iranian period. On the European steppe, East Iranian tribes settled in the Caucasus where they survive today as the Ossetes, and Ossetic provides a valuable source for East Iranian. Sogdian has a distant descendant in the Yaghnobi language of Tadjikistan while the remnants of the Saka languages survive in the Pamirs. The most important modern East Iranian language is Pashto, the state language of modern Afghanistan.

The West Iranian languages were carried into north-west Iran by the Persians and Medes. Old Persian is attested primarily in a series of cliff-carved inscriptions in cuneiform. This material is not particularly abundant and is often repetitively formulaic but it does offer significant additional evidence to Avestan for the early stages of Iranian. By the Middle Iranian period we find Middle Persian, markedly changed from the earlier language. After the Arab conquests of the region (and a major Arabic impact on the Persian language), New Persian arose by the tenth century.

Iranian is closely related to Indo-Aryan and because the latter is far better represented in the earliest periods, there is a greater emphasis on Indo-Aryan among comparativists than on Iranian. Within the wider context of Iranian itself, there are far more languages than have been summarized here.

Because the Avesta and the Old Persian documents are meagre compared to the volume of Sanskrit material, scholars often exploit the vocabularies of the Middle and even the Modern Iranian languages in order to fill out the range of Iranian vocabulary.

Tocharian

At the end of the nineteenth century, western expeditions to Xinjiang, the westernmost province of China, began to uncover remains of what are known as the Tocharian languages. The documents date from the fifth century ad until Tocharian was replaced by Uyghur, a Turkic language, by the thirteenth century AD. There are approximately 3,600 documents in Tocharian but many of these are excruciatingly small fragments. The documents are primarily translations of Buddhist or other Indic texts, monastery financial accounts, or caravan passes. There are two Tocharian languages. Tocharian A, also known as East Tocharian or Agnean, is recovered exclusively from around Qarashahr (the ancient Agni) and Turfan and gives some the impression that it may have been a 'dead' liturgical language by the time it was recorded. Tocharian B, otherwise West Tocharian or Kuchean, was spoken from the oasis town of Kucha east across Tocharian A territory. It is better attested and more conservative than Tocharian A. The application of the name 'Tocharian' to the remains of the documents is controversial: the Tocharians of classical sources were one of the peoples who occupied Bactria, and the presumption that these were the same people (or a closely related group) as those who lived in the Tarim and Turfan basins derives from several manuscript readings which have been rejected as often as they have been accepted. For convenience sake, Tocharian has remained the common designation for this group by most but not all linguists.