

## 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  $s$ ,  $z$  and  $u$  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; so strong, indeed, that no philologist could examine them all three, without believing them to have sprung from some common source, which, perhaps, no longer exists. There is similar reason, though not quite so forcible, for supposing that both the Gothic and the Celtic, though blended with a different idiom, had the same origin with the Sanskrit; and the old Persian might be added to the same family, if this were the place for 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 Boixhorn 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.

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 Conrad Malte-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.

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 are traditionally divided into two main groups — **Continental** and **Insular** Celtic.

The Continental Celtic is represented by **Gaulish**, the language of Gaul. Gaul is the name given by the Romans to the territories where the Celtic Gauls lived, including present France, Belgium, Luxemburg and parts of the Netherlands, Switzerland, Germany.

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**. Breton originated in Britain and was carried from southern Britain to Brittany (region in the north-west of France) 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.

### **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 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.

## **Germanic**

### **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 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 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 is 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 **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 are: **Polish, Czech, and Slovak**.

### **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 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.

### **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.

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.

In south and west Anatolia we find evidence of the other main Anatolian language, **Luvian**. Luvian was written in two scripts: the cuneiform which was also employed for Hittite and a hieroglyphic script created in Anatolia itself.

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.

### **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.

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.

The oldest 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 first millennium BC we find evidence for the vernacular

languages of India which are designated Prakrit. The earliest attested Indo-Aryan documents are in Prakrit and these provide the bases of the modern Indo-Aryan languages, e.g. **Hindi, Urdu, Bengali, Punjabi, Gujarati, Marathi, Sinhalese.**

### **Iranian**

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.

Part of East Iranian tribes settled in the Caucasus where they survive today as the **Ossetes**, and **Ossetic** provides a valuable source for East Iranian. 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. 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. 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 and Tocharian B, otherwise West Tocharian.