1 The Indo-European language family

1.1 Introduction

Indo-European (IE) is the best-studied language family in the world. For much of the past 200 years more scholars have worked on the comparative philology of IE than on all the other areas of linguistics put together. We know more about the history and relationships of the IE languages than about any other group of languages. For some branches of IE – Greek, Sanskrit and Indic, Latin and Romance, Germanic, Celtic – we are fortunate to have records extending over two or more millennia, and excellent scholarly resources such as grammars, dictionaries and text editions that surpass those available for nearly all non-IE languages. The reconstruction of Proto-Indo-European (PIE) and the historical developments of the IE languages have consequently provided the framework for much research on other language families and on historical linguistics in general. Some of the leading figures in modern linguistics, including Saussure, Bloomfield, Trubetzkoy and Jakobson, were Indo-Europeanists by training, as were many of those who taught in newly founded university departments of linguistics in the second half of the twentieth century. Despite this pedigree, IE studies are now marginalised within most university linguistics courses and departments. In most US and European institutions, Indo-Europeanists with university posts do not teach in linguistics departments but in classics, oriental studies, celtic studies or the like. Historical linguistics courses may include a section on PIE, or Saussure's work on laryngeals as an example of internal reconstruction, but few students will engage in any current work on IE in any depth.

The intention of this book is not to convert general linguists to IE studies, or to restore the discipline to the central position in linguistics that it had a hundred years ago. Rather it aims to set forth some of the areas of debate in IE studies. In recent years a number of grammars and handbooks of PIE have been published in English (Gamkrelidze and Ivanov 1984 (English translation 1995), Sihler 1995, Beekes 1995, Szemerényi 1996, Meier-Brügger 2000 (English translation 2003), Fortson 2004). Most of these works are excellent, but sometimes the apodeictic style of the presentation leaves the reader uncertain about whether what is presented is actually hypothesis or 'fact'. One explanation for a historical change may be preferred over another, but the author may not make clear what is at stake in the choice between the alternatives. This book takes a different approach. It is

deliberately not intended to be a grammar of IE, or a survey of the developments that have taken place between PIE and the daughter IE languages, but rather to be a survey of some current debates and topics of more general interest in the reconstruction of PIE, and a guide to the ways in which some of these issues have been addressed. The material throughout the book is selective and illustrative, and the reader who wants to find out more will be advised to follow the further reading sections at the end of each chapter.

1.2 The IE languages

The IE language family is extensive in time and space. The earliest attested IE language, Hittite, is attested nearly 4,000 years ago, written on clay tablets in cuneiform script in central Anatolia from the early second millennium BC. We have extensive textual remains, including native-speaker accounts of three more IE languages from 2,000 years ago: Ancient Greek, Latin and Sanskrit. Also from the beginning of the Christian Era we have much more limited corpora of many more IE languages. The stock of recorded IE languages further increases as we move forward in time. In 2003, over 2.5 billion people spoke an IE language as their first language, and there were at least seventy codified varieties, each spoken by a million or more native speakers. Four hundred years ago nearly all speakers of IE lived in Europe, Iran, Turkey, Western Asia and the Indian subcontinent, but migrations have now spread speakers to every part of the world. The wealth of historical material makes IE the best-documented language family in the world.

What is it that makes an IE language IE? What does it mean to be classed as an IE language? It is usual at the opening of books on IE to repeat the famous words of Sir William Jones in 1786 which are traditionally taken to have inaugurated the discipline. Jones remarked on the similarity of Sanskrit to Latin and Greek, stating that they all bore '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 philologer could examine them all three, without believing them to be sprung from some common source, which, perhaps, no longer exists'. Jones also noted that Gothic, Celtic and Persian could be added to the same family. Since 1786, a considerable methodology has been established to qualify and quantify Jones' notion of 'affinity' between the grammars and lexicons of the IE languages, and to work out a hypothetical model of the 'common source', PIE. But there has been no advance on Jones' criterion for relatedness between languages of the family: greater similarity in verbal roots and morphological paradigms than might be expected by chance. Languages which belong to the IE family do so either because the similarity between them and other IE languages is so strong as to be self-evident, or because they can be clearly related to languages which do obviously belong to the family. For a language which has textual remains sufficient

for the linguist to extract lexical and grammatical information, it is possible to apply the techniques of reconstruction, such as the comparative method, to build a picture of its development from PIE. However, the operation of the comparative method does not guarantee a language's place in the family; only the initial recognition that two or more languages are related can do that. (We shall return to examine the implications of this point more fully in section 1.6.)

When does a linguist decide that there is enough material to relate a language to the IE family? There is no absolute set of criteria beyond the general rule that the evidence must convince both the individual linguist and the majority of the scholarly community. A language which only survives in a very limited corpus may contain sufficient IE features to be generally agreed to be IE. As an example, take the case of Lusitanian. Lusitanian is known from a handful of inscriptions from the west of the Iberian peninsular, written in the Latin alphabet around the first century of the Christian Era. One of these inscriptions, from Lamas de Moledo in Portugal, reads as follows (the slash / signals the end of the line in the original inscription):

RVFINVS. ET
TIRO SCRIP/SERVNT
VEAMINICORI
DOENTI
ANGOM
LAMATICOM
CROVCEAIMAGA
REAICOI. PETRANIOI. T
ADOM. PORGOM IOVEAI
CAELOBRIGOI

The first four words are Latin: 'Rufus and Tiro wrote (this).' But the remainder of the inscription is not Latin. The inscription is taken to refer to the sacrifice of animals by a people called the *Veaminicori* to gods who are also addressed with their cult titles. Not all the words are understood, although the structure is clear: *Veaminicori* is nominative plural, *doenti* is a verb meaning 'they give'. The rest of the inscription has nouns in the accusative singular, denoting what is given: *angom lamaticom*, *tadom porgom*; and the names of the recipients in the dative singular: *petranioi*, *caelobrigoi*. This is not much, but enough that no Indo-Europeanist doubts that Lusitanian is a member of the IE family. Several of the word-forms are very similar to Latin. For example, the dedicated item *porgom* is very likely to mean 'pig' (Latin accusative singular *porcum* 'pig'), and *angom* to mean 'lamb' (Latin accusative singular *agnum* 'lamb'). The verb-form *doenti* 'they give' contains the root *do-* 'give', familiar from the equivalent forms in Greek (*dō-*), Latin (*da-*) and Sanskrit (*dā- / d-*). More importantly, it shows a third person plural ending *-enti* which is also found in these languages (dialectal Greek *-enti*, Archaic Latin *-nti* and Sanskrit *-anti*). Furthermore, the ending *-oi* coincides with a dative singular marker elsewhere (Greek *-ōi*, Archaic Latin *-oi*

and Sanskrit *-ai*), and the nominative plural ending *-i* accords with the nominative plural *-i* of one Latin noun declension. The interpretation of this inscription rests entirely on the identification of its language as IE, but most scholars have found it hard to believe all these similarities are entirely due to chance.

Compare with Lusitanian the case of Tartessian, another language from Ancient Spain which is known only from short inscriptions. Tartessian is better attested than Lusitanian, and from a period 600–800 years earlier. Unfortunately, we are not confident about our reading of the Tartessian script, and we do not have the helpful marks which are usually present in the Lusitanian inscriptions indicating where words begin or end. We consequently do not know a lot about the morphology of the language. However, some scholars have identified in Tartessian repeated patterns of (what they take to be) verbal endings. Consider the following inscription, reproduced in its entirety:

botieanakertorobatebarebanarkenti

The final nine letters, *narkenti*, occur elsewhere in the inscriptional corpus, as do the similar forms *narken*, *narkenii*, *narke*, *narkenai*. Here again we see a final element *-nti* that could represent the third person plural of a verbal ending in an IE language, just as in Lusitanian above. However, there is no obvious connection in the older IE languages to what would appear to be the verbal 'stem' *nark-*. Moreover, if we try to use what we know of IE morphology and vocabulary to interpret the rest of the inscription, we do not get very far. In Lusitanian, the assumption that the language was IE yielded vocabulary and morphology. In Tartessian, we have nothing more than the ending *-enti*. We do not even know enough about the morphological structure of the language to be confident that *narkenti* should be analysed as stem *nark-* + affix *-enti*. Accordingly, the general consensus is that Tartessian should not be included among the IE family.

The status of languages as IE or not may change in the light of an increase in our knowledge of the family. This is the case with the languages Lydian and Lycian, spoken in Anatolia in the first millennium BC, and known from inscriptions written in modified forms of the Greek alphabet. Before the discovery and accurate description of older IE languages in the Anatolian family, Hittite and Luwian, written in cuneiform and hieroglyphic scripts hundreds of years earlier, Lydian and Lycian could not be securely included in the IE family. However, their affinity to the earlier Anatolian languages is now patent, and since these show clear morphological and vocabulary similarities with the rest of IE, there is no doubt that Lydian and Lycian belong in the family as well. If we did not have any Anatolian languages other than Lydian and Lycian, we would not now be so certain of their ancestry. Indeed, we would not be able to make much sense of them at all, since it is only through the knowledge of how Anatolian languages are structured that headway has been made with the interpretation of the surviving inscriptions. It is, consequently, conceivable that a language such as Tartessian could come into the IE fold, if we were to have some intermediate steps to show the link between the rest of the family and the inscriptional remains that we have.

1.3 The branches of the IE tree

It follows from the remarks about Lydian and Lycian that the sub-families of IE are vitally important in determining the membership of the family. Whereas the affinity of the oldest IE languages declares itself as stronger than could be produced by chance (to most of those who study them), the affinity of languages attested more recently is sometimes only discernible through first relating them to sub-families of IE. Thus, to take an example of two languages at the far ends of the historical IE speech area, Modern Irish and Sinhala would not strike a linguist who was fluent in each, but unacquainted with their history, as *necessarily* related. It is only through relating Modern Irish to Old Irish, and Sinhala to Sanskrit, that the connection between the two languages becomes clear.

The majority of IE languages currently spoken belong to six large sub-groups of IE. Modern Irish and Old Irish are members of the Celtic sub-group, which also includes Welsh, Scots Gaelic, Breton, Cornish and Manx. Sinhala is part of the large Indic family, comprising most of the languages currently spoken in North India and Pakistan, Sanskrit and the Middle Indian Prakrits. English is a member of the Germanic branch; this includes Dutch, German and the Scandinavian languages among living languages, as well as earlier stages of these languages, such as Old English, Old High German and Old Norse, and other extinct varieties such as Gothic, once spoken in south-east Europe and southern Russia. The other large sub-groups are Romance and Slavic in Europe, and Iranian in Asia. All of these sub-groups of IE were themselves recognised as linguistic families before Jones' identification of the larger IE family cited above. The traditional criterion for grouping these languages was, in general terms, analogous to the criterion Jones used for IE. The members of a sub-group are so much more similar to each other than they are to other IE languages that the similarity cannot be put down to chance. Now, however, there are firmer criteria for membership of a sub-group. Two languages grouped together in a sub-group are assumed to have derived from a language, the 'sub-group parent language', which is chronologically earlier than either of the grouped languages, but which was spoken after PIE. The relationship can be represented diagrammatically as a family tree, with the historically prior languages situated at higher nodes in the tree. In figure 1.1, languages A and B constitute a sub-group, since they derive from a single language intermediate between them and the parent. Languages C and D do not constitute a sub-group between each other or with either A or B.

The family tree model has been very influential in IE studies, and we shall consider it in more detail below. In some cases, as in the Romance language sub-group of IE or the Indic sub-group, we have records of an early language variety which either can be identified with the sub-group parent, or which is very close to the sub-group parent (Latin and Sanskrit in the two cases respectively). But for some other sub-groups we do not have an attested parent, and it has to be reconstructed using the comparative method. It is now generally agreed among

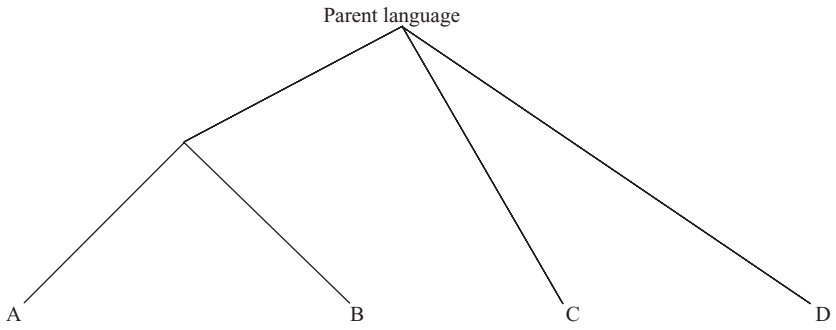

Figure 1.1 *A language family tree*

linguists that the most certain sub-groups are constructed on the basis of unique shared morphological innovations. That is, where there is no attested parent for a group of languages, they may be reckoned to belong to the same sub-group if they share a significant number of new developments in their morphology, particularly inflectional morphology. If, for example, two languages have constructed a new morphological category with a new morphological marker, and the marker is not found in other IE languages, this is reckoned to be a significant morphological innovation. It is only through morphological changes of this sort that we can be sure that there is a reconstructed sub-group parent: lexical and phonological developments are too easily shared through linguistic convergence, and we do not generally have enough information about reconstructed syntax to be certain that syntactic changes are innovations.

Using this methodology of sub-grouping it is possible to identify further sub-groups of Indo-European beyond the six large sub-groups identified above. Lithuanian and Latvian are only attested from the Early Modern period, and together with the now extinct Old Prussian they form the Baltic sub-group. Two sub-groups are no longer extant: Anatolian, mentioned in section 1.2 above, which was widespread in central and western Anatolia before the Christian Era, and Tocharian, known from the textual remains of two separate languages (now known as Tocharian A and Tocharian B) spoken in central Asia in the sixth to eighth centuries AD. Sub-grouping methodology also makes it clear that the Indic branch and the Iranian branch are more closely related to each other than to any other branch, and they are now recognised as an Indo-Iranian sub-group. Baltic and Slavic are usually also assumed to stem from a single Balto-Slavic branch, but in this case we cannot be so sure, since the languages are attested so much later.

A few IE varieties still spoken are not allocated to sub-groups, but are usually represented as separate 'branches' of the IE family tree. The languages in question are all spoken around the Eastern Mediterranean: Greek, Albanian and Armenian. Greek, as we have seen, has a long history, but the other two languages are more recent: Armenian dates from the middle of the first millennium, Albanian

from the second millennium of the Christian Era. Greek, Albanian and Armenian are thought by some scholars to comprise a 'Balkan IE' sub-group, but this hypothesis is disputed, since Albanian and Armenian have undergone so much linguistic change that their morphological developments are difficult to identify with confidence. Finally, there are varieties of IE no longer spoken which are not securely allocated to sub-groups. These are sometimes called 'fragmentary IE languages', since most are known from only a small corpus of material. Lusitanian, discussed in section 1.2 above, is an example of such a language.

It is a curious paradox of IE linguistics that the languages which are attested earliest are often the most difficult to assign to any sub-group. Of the IE languages spoken today, only Greek, Armenian and Albanian do not have close relatives in the same way that English compares to Dutch and German, or French to Italian and Spanish. Two thousand years ago, the linguistic map was different. Many of the languages spoken around the Mediterranean in 500 BC were superseded by Latin and its descendants following the Roman Conquest. As far as we can tell from the scanty textual remains of these languages, most were independent branches of IE, and not part of a sub-group. Lusitanian is one example of such a language, and Messapic provides another. Messapic is the name given to the language of around 300 short inscriptions from the heel of Italy, which were written in the Greek alphabet between the fifth and second century BC. Like Lusitanian, it is generally recognised to be IE, but it is not securely associated with any other IE language. The difficulty of assigning Messapic to any branch of IE is not just a problem of interpretation of a scanty corpus; the language shows significant divergences from the IE branches which are attested closest to it: Greek, Latin and the Sabellian languages of Italy, and Albanian. Other scantily attested Mediterranean languages which do not fit into a sub-group include: Phrygian, attested in central Asia Minor in two different varieties (Old Phrygian, from the eighth to the fourth century BC, and New Phrygian, from the second and third century AD); Venetic, attested in north-east Italy in nearly 300 short inscriptions from around the sixth to the second century BC; Thracian, the name given to the language of a text of sixty-one letters inscribed on a gold ring found at Ezerovo in Bulgaria and some short inscriptions on coins. Of the languages attested in the last 200 years, the only good candidates for a new branch of IE are the Nuristani languages spoken in remote valleys in eastern Afghanistan, which are thought to represent a third branch of the Indo-Iranian sub-group beside Indic and Iranian.

Table 1.1 is intended to illustrate the point about sub-groups; it shows first attestations of language and language groups by date and place, dividing the IE speech area into four different zones. Northern Europe comprises the area north of the Alps stretching from Ireland in the west to the Urals. The western Mediterranean comprises Spain, southern France and Italy. The eastern Mediterranean comprises Greece, Anatolia and the Black Sea area. The fourth zone includes Asia east of the Urals, the Indian sub-continent, and Iran and neighbouring countries to the east. The table gives the first appearance of languages in lower case and IE sub-groups or languages which represent independent branches of IE in

Table 1.1 *IE languages by date and place of first attestation.*

Date	Northern Europe	Western Mediterranean	Eastern Mediterranean	Iran / Central Asia / India
1800 BC			Old Hittite (ANATOLIAN)	
1400 BC			Mycenaean Greek (GREEK) Mittani (INDIC)	
500 BC		Latin (ROMANCE) South Picene (SABELLIAN) VENETIC Lepontic (CELTIC) MESSAPIC	PHRYGIAN THRACIAN MACEDONIAN	Old Persian (IRANIAN)
1 AD	LUSITANIAN			
500 AD	Rune inscriptions (GERMANIC)		ARMENIAN	
1000 AD	Old Church Slavonic (SLAVIC)			TOCHARIAN
1500 AD	Old Prussian (BALTIC)	ALBANIAN		
2000 AD	NURISTANI			

SMALL CAPS. The information in the table relies on dated texts, which means that the Indic family is attested first through the existence of some personal names and words relating to horse-training which occur in Hittite, Hurrian and Babylonian records from 1400 BC on, and not through the orally transmitted Vedic hymns. A similar problem surrounds the dating of the Iranian languages: Gathic Avestan, the language of the central portion of the sacred books of the Zoroastrians, certainly reflects an earlier stage of Iranian than the Old Persian inscriptions, but its transmission history does not allow us to date it securely. In the table, once one member of a sub-group is attested the sub-group is not recorded again, even when later representatives of the family occur in a different zone.

The order of attestation of different languages is reliant on the transmission of scripts and literacy. Unfortunately, the social and cultural changes which brought about an increase in literacy in much of the area where IE varieties are spoken also led to the spread of a few dominant languages at the expense of others. Table 1.1 shows the effect this has on the attestation of different languages. In the western and eastern Mediterranean zones at the onset of literacy in the first millennium BC a number of different languages are attested. In the early centuries of the Christian Era most of these languages were replaced by Latin and Greek and their descendants. The spread of these languages, and of the other

large sub-groups, is not surprising. Most of the area where the IE languages are spoken are classic 'spread zones' in the terminology of Nichols (1992). That is to say, they are areas where large-scale population movement is possible, and where one social group may readily achieve dominance over its neighbours. The IE languages for which we have fairly extensive records from before 1000 AD – Latin, Greek, Germanic, Iranian and Indic – have been the carriers of cultures which have in time predominated over other indigenous groups, with resultant language shift. Populations which once spoke Messapic, Venetic and Lusitanian eventually shifted to speaking Latin, Phrygians adopted Greek and Thracian lost out to overlapping waves of Greek, Latin, Germanic (Gothic) and Slavic. In the Mediterranean area, the early adoption of literacy allows us to know of a range of IE varieties. In northern and eastern Europe, where the first written records appear considerably later, we do not know whether there was a similar diversity in the territories later occupied by speakers of Celtic, Germanic, Slavic and Baltic languages. We shall consider further the question of how we can assess the evidence for the early relationship of the IE family, considering what we have lost, in the next section.

1.4 Cladistics: constructing family trees

The family tree model of IE is over 150 years old. The model was first put forward in the nineteenth century, and the first tree diagram was produced by the German Indo-Europeanist August Schleicher (reproduced in figure 1.2). Schleicher's tree does not include Armenian, which was not then recognised as a separate branch of IE, nor Anatolian or Tocharian, which were not then known. As our understanding of the IE languages has increased and changed, so also the tree has changed. In Schleicher's tree, the first split is made between Germanic, Baltic and Slavic and the other language groups. This split reflects the fact that the three sub-groups spoken in the north of the IE area form dative-ablative and instrumental plural cases in some noun paradigms with a marker involving the original phoneme *m, whereas the other languages use a marker with *b^h, as shown in table 1.2, which gives the instrumental plural markers in various IE languages (note that all reconstructed, as opposed to attested, sounds, morphs and words are preceded by * throughout this book and in most works on PIE).

This divergence between the languages is still unexplained – it may be that the two plural cases which use *m or *b^h, the dative-ablative and instrumental, originally took separate markers, but some languages generalised *m to both of them, others *b^h. Modern scholars do not see the distinction between the use of *b^h and *m in these cases as sufficient evidence for a fundamental split between two parts of the IE language family. Furthermore, there are other features which unite the languages of the western IE zone: Celtic, Germanic and Latin and the Sabellian languages. In constructing a family tree, the shape of the tree depends on what the linguist sees as important.

Table 1.2 *Instrumental plural markers in various IE languages.*

PIE **-mis*	Germanic: Gothic *-m*
	Slavic: Old Church Slavonic *-mi*
	Baltic: Lithuanian *-mis*
PIE **-b^{h}is*	Indo-Iranian: Sanskrit *-bhis*
	Greek: Mycenaean Greek *-pi* (/-p^{h}i/)
	Celtic: Old Irish *-^{i}b*

In recent years, the advance of statistical techniques and the use of computers to process very large amounts of data have allowed the construction of family trees from a much wider data set and a resurgence in interest in drawing a family tree for IE. Since computer analysis allows for such a large amount of discrete data to be handled, trees can be constructed using hundreds of different features. The new technology brings with it a new terminology, and now linguists are beginning to talk not of family trees, but *phylogenies*, and to use the term *cladistics* for referring to the techniques of constructing family trees. Two recent phylogenies of PIE are given in figure 1.3 (the 'New Zealand' tree constructed by Gray and Atkinson 2003: 437) and figure 1.4 (the 'Pennsylvania tree' taken from Ringe, Warnow and Taylor 2002: 90). The two phylogenies use different features in order to rank languages against one another.

The New Zealand tree, figure 1.3, relies upon vocabulary items only, following in a long tradition of language surveys which rely upon word lists or 'basic

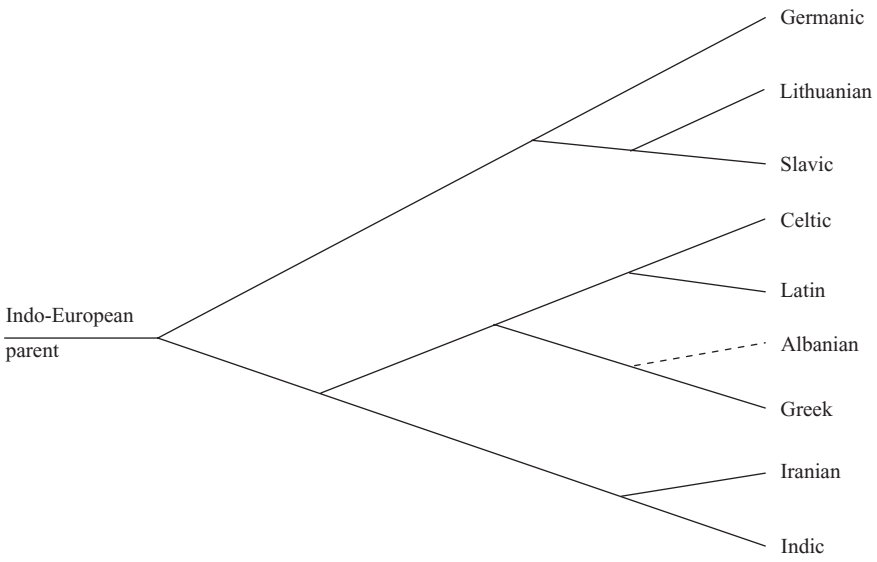

Figure 1.2 *Schleicher's Indo-European family tree*