1   Introduction: how are languages shown to be related to one another?

> I fear great evil from vast opposition in opinion on all subjects of classification.
> (Charles Darwin 1838 [Bowlby 1990:225])

1.1   Introduction

How are languages shown to be related to one another? How are language families established? Judging from media attention, it might be thought that this is one of the "hottest" questions in contemporary linguistics. Proposals of distant linguistic kinship such as Amerind, Nostratic, Eurasiatic, and Proto-World have been featured in *Atlantic Monthly, Nature, Science, Scientific American*, *US News and World Report, The New York Times*, and in BBC and PBS television documentaries. Nevertheless, these same proposals have been rejected by the majority of practicing historical linguists. The difference of opinion is reflected in much debate and considerable confusion about the methods for demonstrating family (phylogenetic) relationships among languages as yet not known to be related, and about the ways that language families have come to be established. Some enthusiasts of long-range linguistic relationships, disappointed that proposed language connections they favor have not been accepted in the profession, have at times responded bitterly. For example, we read charges that these rejections are just "clumsy and dishonest attempts to discredit deep reconstructions [proposed macro-families]" (Shevoroshkin 1989a:7, also 1989b:4). Nor is the strong rhetoric just from fans of distant language relationships; for example, Dixon (2002:23) lambasts the so-called "long-rangers": "at a different level – which transcends scientific worth to such an extent that it is at the fringe of idiocy – there have in recent years been promulgated a number of far-fetched ideas concerning 'long-distance relationships,' such as 'Nostratic,' 'Sino-Caucasian' and 'Amerind.'"

Some proponents of remote relationships have sometimes attempted to give their claims a sense of legitimacy by associating their methods with those of revered founding figures in historical linguistics, in particular those credited with contributions to establishing the Indo-European family of languages. Given

1

these claims and the importance assigned to methods in the debates, one of the purposes of this book is to review and evaluate the methods which have been used throughout the history of linguistics to establish language families, that is, the methods for demonstrating genetic relationships among related languages.[1] The successful methods employed in the past provide important lessons for current practice. The picture which emerges from this historical survey of how major language families came to be established is very different from that which is often claimed. Therefore, one goal of this book is to set the historical record straight. Our aim, however, is not to dwell on past linguistics, but to assess current claims about past methods and in so doing to contribute by refining current practice.[2] The contribution to linguistic historiography is a fortunate by-product of this, but is not our main goal.

The broader and more central aim of this book is to contribute to language classification, to research in distant genetic relationships generally by: (1) showing how the methods have been employed, (2) revealing which methods, techniques, strategies, rules of thumb and the like have proven successful and which ones have proven ineffective, (3) finding out how particular language families were established – that is, what methods were utilized and proved successful, (4) evaluating a number of the most prominent and more controversial proposals of distant genetic relationship in the light of the methods which prove most adequate, and (5) making recommendations for practice in future research. In brief, we hope to contribute significantly to understanding of language classification in general.

Several scholars – outside the mainstream of historical linguistics – have in recent years made pronouncements in favor of very remote language relationships, and several of them (though not all) have urged methodological points of view with which most historical linguists do not agree. Their claims figure in

---

[1] In linguistics, "genetic relationship" among languages means a phylogenetic, genealogical relationship, that is, descent from a common ancestor. While there is potential for confusion with "genetic" in reference to biological species, in this book we continue the common linguistic practice of speaking of languages as being genetically related. In contexts where potential confusion may arise, we speak of "human genetic" to distinguish it from the use of "genetic" for linguistic kinship. Also, in general, we speak of languages being "related," of "related languages," and of "language relationships" in reference to genetic relationships. Of course, languages can be "related" through borrowing and other means, but "relationship" in the context of historical linguistics usually refers to a phylogenetic relationship.

[2] We concur with Hoenigswald (1974:346) and Wilbur (1977:ix) that linguists have been too obsessed with the nineteenth century. The claim of historical antecedents in attempts to justify particular methods used today is an example of this. As Wilbur (1977:ix) says:

Henry M. Hoenigswald's undisguised complaint [is] that we are "obsessed" with the nineteenth century. He suggests to the linguist a therapeutic program by insisting that he face up to this presumably morbid preoccupation. With a certain amount of pathos, Hoenigswald (1974:346) also deplores our "extreme working dependence" upon that century. It would be intellectually more honest and psychologically more salubrious to admit that, as far as historical linguistics is concerned, we are, to a very great degree, still in the nineteenth century.

discussions in this book. While understanding of the methods of the founding figures in historical linguistics does offer object lessons for us today, as will be seen, interpretations of this part of the history of linguistics have often been mistaken.

In Chapters 2, 3, and 4, we examine the methodological pronouncements and practices of these founding figures, and in the process we discover their real outlook on language relationships and how they have contributed to understanding of language classification. Jones', Rask's, Bopp's, and the others' methods owe much to their predecessors, discussed in Chapter 2, and therefore we approach our historical survey by considering events roughly in chronological order, first earlier developments (Chapter 2), then Sir William Jones' much misconstrued role (Chapter 3), followed by a consideration of the work of Rask, Bopp, and others influential in the development of comparative linguistics, including the claims of the Neogrammarians and their contemporaries (Chapter 4). In Chapter 5 we look at how several languages were shown to belong to Indo-European, and discover that the methods used were not as some have claimed, and the methods that were employed provide valuable insights. In Chapter 6 we see how prominent language families were established, and we look at work in linguistic classification in other areas of the world. We inspect the methods that were employed to establish the Finno-Ugric (and Uralic), Semitic, Dravidian, Austronesian, Sino-Tibetan, and several Native American language families, and also those utilized in work on the language families of Africa and Australia. We find that the methods which were utilized to establish the accepted language families have generally been consistent with those of the Indo-Europeanists, and the work on non-Indo-European families frequently further contributed to historical linguistic thinking in general. We survey principles, criteria, and methodology in general in Chapter 7. In Chapter 8, the "other" tradition in historical linguistics is considered, the attempts to get at language history and classification through a combination of psychological, typological, and evolutionary notions. This was a highly influential tradition in the history of linguistics which ultimately was supplanted by the approach now associated with the Neogrammarians and their legacy. In Chapter 9 we employ the methods and criteria surveyed in Chapter 7 to evaluate the better known long-range proposals, Altaic, Ural-Altaic, Nostratic, Eurasiatic, Amerind, and Indo-Pacific – well-known but controversial "macro-family" hypotheses. We show why most historical linguists do not accept these proposals. Chapter 10 deals with various recent attempts to see beyond the comparative method and with proposed frameworks which depart significantly from traditional methods. In particular, Chapter 10 looks critically at proposals argued to be able to recover aspects of linguistic prehistory beyond the limitations of the traditional comparative method. Chapter 11 presents a critique of recent proposals concerning how and why linguistic diversification and language dispersal take

place, followed by our recommendations for how these are best approached. Chapter 12 evaluates critically the "Proto-World" hypothesis, the idea that it is possible to show that all the languages of the world are related and descend from a single original language. In Chapter 13, the conclusions, we reiterate the main findings of the book and attempt to point the way to productive future research. Finally, in the Appendix we list, but do not discuss or evaluate, most of the proposals of distant genetic relationship that have been made.

1.2   Evidence and criteria

It will not spoil any surprises to come if we disclose here at the outset that throughout the history of linguistics the criteria employed in both pronouncements about method and in actual practice for establishing language families consistently included evidence from three sources: **basic vocabulary**,[3] **grammatical evidence** (especially morphological), and **sound correspondences**. Hoenigswald's summary of the points upon which seventeenth- and eighteenth-century linguistic scholars agreed is telling. Hoenigswald (1990a:119–20) quotes from Metcalf's (1974:251) similar summary:

First, . . . there was "the concept of a no longer spoken parent language which in turn produced the major linguistic groups of Asia and Europe." Then there was . . . "a concept of the development of languages into dialects and of dialects into new independent languages." Third came "certain minimum standards for determining what words are borrowed and what words are ancestral in a language," and, fourth, "an insistence that not a few random items, but a large number of words from the basic vocabulary should form the basis of comparison" . . . fifth, the doctrine that "grammar" is even more important than words; sixth, the idea that for an etymology to be valid the differences in sound – or in "letters" – must recur, under a principle sometimes referred to as "analogia."[4]

[3] For a general discussion of the importance which basic vocabulary as a criterion for comparing languages has had throughout the history of linguistics, see Muller 1984.

[4] That is to say, there was a great deal of individual variation in the criteria used and the kinds of evidence sought in early attempts to establish language families and not every scholar used all three of these sources of evidence, but many did use all three and on the whole, these three were widely recognized and accepted.

    We note also, as Hoenigswald shows here, that many have mistakenly claimed Jones was the first to imagine the ancestor language "perhaps no longer" existed (cf. Pedersen 1983[1916]:34–5). The view was quite common, repeated from at least Ludolf in 1692 onward (Waterman 1978:61). Hoenigswald's citation also shows the inaccuracy of Greenberg's belief that "none of these families [Indo-European, Uralic, or Austronesian] was discovered by finding regular sound correspondences" (1996a:133).

Thanks to past scholarship, we have the evidence that demonstrated the family relationships at our disposal. However, today we are confronted not just by the question of how accepted language families were established, but also with the issue of how to tackle possible genetic relationships among languages where the answer is not at all obvious. That is, on all sides we encounter far-flung proposals of very remote relationships among languages not at present known to be related. Janhunen and Kho's (1982:179) characterization of attempts to find distant relatives of Korean applies equally well to many other enthusiastically proposed but poorly supported long-range relationships:

The various hypotheses concerning genetic affinities . . . offer an instructive record of what the use and misuse of well established scientific methods can yield . . . it is a record of unchecked attempts at proving preconceived ideas, which often from the very beginning stand in sharp contradiction with the suggestions of common sense. The characteristic feature of this type of proposals is the axiomatic belief in the existence of discoverable genetic connections, while the possibility that a language might actually not have any living (or historically recorded) relatives is left without due considerations.

We do not mean to say that there is no reliable work that attempts to establish new language relationships – quite the contrary – our purpose in this book is to encourage such work. Nevertheless, many unrestrained and indefensible hypotheses plague the literature at present. We hope to put the more influential ones in perspective in this book, emphasizing methodological implications for all such work.[5]

1.3   Setting the stage

A ubiquitous mistaken belief is that comparative linguistics and the discovery of Indo-European started with Sir William Jones' famous declaration of 1786, often called the "philologer" passage (see Chapter 3). We set the stage for the historical survey of methods for establishing language relationships by considering briefly some precursors and contemporaries of Jones who reveal a very different history. This is seen by contrasting Jones' famous statement with those of some of his predecessors. Jones' philologer passage, so often cited, is:

The *Sanscrit* 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

[5] We are, of course, not the only nor first ever to be concerned with methodology for establishing language relationships. See, for example, Aalto 1965; Adam 1881; Allen 1953; Austerlitz 1983, 1991; Bright 1970; Callaghan and Miller 1962; Collinder 1946–8, 1964; Cowan 1962; Doerfer 1967, 1973; Fähnrich 1971; Gatschet 1879–80, 1886; Grolier 1990; Joki 1963, 1973; Kroeber 1913; Lehmann 1994; Pisani 1971; some papers in Shevoroshkin and Sidwell 1999; Starostin 1999a, 1999b; Thomason 1993; Voegelin and Voegelin 1985; and many others.

no philologer could examine them all three without believing them to have sprung from *some common source*, which, perhaps, no longer exists. There is a similar reason, though not quite so forcible, for supposing that both the *Gothic* and *Celtick*, though blended with a very different idiom, had the same origin with the *Sanscrit*; 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*. (Jones 1798[1786]:422–3)

Although highly celebrated, this quotation is, nevertheless, remarkably similar to Andreas Jäger's statement of 1686, a hundred years earlier:

An ancient language, once spoken in the distant past in the area of the Caucasus mountains and spreading by waves of migration throughout Europe and Asia, had itself ceased to be spoken and had left no linguistic monuments behind, but had as a "mother" generated a host of "daughter languages," many of which in turn had become "mothers" to further "daughters." (For a language tends to develop dialects, and these dialects in the course of time become independent, mutually unintelligible languages.) Descendants of the ancestral languages include Persian, Greek, Italic (whence Latin and in time the modern Romance tongues), the Slavonic languages, Celtic, and finally Gothic and the other Germanic tongues. (Quoted in Metcalf 1974:233)[6]

   Statements such as this one from Jäger and various others (Chapter 2) require adjustments in the traditional thinking about the role of Jones and the origin and development of the comparative method and Indo-European linguistics. For example, before Jones' famous "philologer" passage was published (delivered in 1786, published 1798), Jonathan Edwards, Jr. (1788[1787]) had discovered and reported the family relationship among the Algonquian languages with methods not unusual for his day but superior to Jones' (see Chapter 6):

This language [family] is spoken by all the Indians throughout New England. Every tribe . . . has a different dialect [language]; but the language is radically [genetically] the same [i.e. members of the same family]. Mr. Eliot's [1663] translation of the Bible is in a particular dialect [i.e. the Natick or Massachusett language] of this language. The dialect followed in these observations is that of Stockbridge [i.e. the Mohegan language]. This language [i.e. the Algonquian family] appears to be much more extensive than any other language in North America. The languages of the Delawares in Pennsylvania, of the Penobscots bordering on Nova Scotia, of the Indians of St Francis in Canada [Abnaki?], of the Shawanese [Shawnee] on the Ohio, and of the Chippewaus [Ojibwa] at the westward end of Lake Huron, are all radically the same with the Mohegan [are of the same language family]. The same is said concerning the languages of the Ottowaus [Ottawa], Nanticooks [Nanticoke], Munsees, Menomonees, Messisaugas, Saukies [Sauk], Ottagaumies [Fox], Killistinoes [Cree], Nipegons [Winnebago], Algonkins, . . . &c. That the languages of the several tribes in New England, of the Delawares, and of Mr. Eliot's

---

[6] Note also that Jäger includes Slavic, which Jones thought was connected with non-Indo-European Central Asian languages, and unlike Jones, for whom Celtic and "Gothic" (Germanic) were "mixed" with non-Indo-European languages, Jäger accurately classifies them as members of the same family without the assumed admixture.

Bible, are radically the same [belong to the same family] with the Mohegan, I assert from my own knowledge. (Edwards 1823[1788]:8)

It is not to be supposed, that the like coincidence is extended to all the words of those languages. Very many words are totally different. Still the analogy is such as is sufficient to show, that they are mere dialects [i.e. sisters] of the same original language [family]. (Edwards 1823[1788]:11)

To show the genetic relationship, that is, "to illustrate the analogy between the *Mohegan*, the *Shawanee* [Shawnee], and the *Chippewau* [Ojibwa] languages" (Edwards 1823[1788]:9), Edwards presented a list of "some 60 vocabulary items, phrases, and grammatical features" (Koerner 1986:ii). Jones, in contrast, presented no actual linguistic evidence at all. In Jones' words:

I am sensible that you must give me credit for many assertions which, on this occasion, it is impossible to prove; for I should ill deserve your indulgent attention, if I were to abuse it by repeating a dry list of detached words, and presenting you with a vocabulary instead of a dissertation; but, since I have no system to maintain, and have not suffered imagination to delude my judgement; since I have habituated myself to form opinions of men and things from *evidence* . . . I will assert nothing positively, which I am not able satisfactorily to demonstrate. (Jones 1799a:49)

Edwards concluded from the evidence he presented that these languages are "radically the same" (i.e. related, from the same family), though he was fully aware also of differences among them and of their differences from Iroquoian languages, members of an unrelated family (with which he also had first-hand acquaintance).

Edwards' case is striking; it shows that neither Jones nor Indo-European occupy the privileged position so often attributed to them in the history of linguistics. The discovery of the Algonquian family relationship has a pedigree comparable to Indo-European, but, as we will see later, Edwards' methods were superior to Jones', and Edwards was not alone (see Chapters 3 and 6).

As this shows, and as will be seen in more detail in subsequent chapters of this book, Jones' role in the history of linguistics has been misread; there is much in this history which will prove methodologically interesting for attempts to establish family relationships among languages as yet not known to be related.

1.4   Classification, but of what?

Until relatively recently, many classifications of languages were not intended to be purely linguistic in nature. They were usually seen as classifying nations or "races," as well as languages, and quite often they were not based on linguistic evidence alone. As we shall see (Chapter 3), Sir William Jones was interested in the history of the "human races" rather than in language per se. This was a common theme which persisted in scholarship until the beginning of the twentieth

century. Jones' plan was to write a history of peoples of Asia – language ("and letters") was only one source of information which Jones believed could be used towards this end, to be utilized in conjunction with information from philosophy and religion, remains of sculpture and architecture, and documents of sciences and arts (Jones 1798[1786]:421). In this regard, again, Jones was not unusual among early linguistic scholars. The view was shared by Kraus, Leibniz, Hervás y Panduro, Monboddo, Vater, Schlegel, Grimm, Humboldt, Schleicher, and others (see Chapter 2).[7] As Humboldt (1836–9:220) put it, "The relatedness of the nations is very often confused with that of the languages, and the demands of historical and ethnographic research with those of linguistics."[8]

For these scholars, their linguistic comparisons were part of a broader history and anthropology (also philosophy and psychology) of the nations and "races" of the world. Hervás y Panduro (1800:1), in his broad-scaled classification of languages of the world, confirmed how common the use of these non-linguistic sources of evidence was for historical interests:

In the histories of nations, all authors claim to arrive at their primitive state with the telling of their deeds. Taking the various directions which they are presented by hieroglyphic monuments, mythology, tradition, calendars, alphabets or writing, and historical accounts of the nations, they fly through immense stretches of the obscure times of antiquity.[9]

The theme of language in concert with other sources of evidence to establish origins and to determine the classification of nations and "races," as well as of languages, was to persist into the early twentieth century.

An especially clear instance which tells how intertwined the classification of race, nation, and language were, and how various sources of information could serve these ends, comes from David Cargill's response to questions about Fijian and its relatives (see Chapter 6). Cargill was among the first missionaries to Fiji, a scholar well-trained in Latin, Greek, Hebrew, and Bible translation, and he knew Tongan before being sent to Fiji. When asked in 1839 by the Secretaries of the Wesleyan Methodist Society in London to discuss the relationship of Fijian to "the other Polynesian dialects," he responded:

[7] For example, Leibniz wrote in 1692, "Languages are the most ancient monuments of the human race, and they serve best for determining the origin of people" (Waterman 1978:59). Court de Gébelin (1773–82), in his search for the primitive or first language of humans, "undertook an impressive etymological analysis of Greek, Latin and French. Nor did he neglect coats of arms, coins, games, the voyages of the Phoenicians around the world, American Indian languages, medallions, and civil and religious history as manifested in calendars and almanacs" (Eco 1995:94).

[8] "Man verwechselt sehr häufig die Verwandtschaft der Nationen mit der Sprachen, und die Forderungen der geschichtlichen und ethnographischen Forschung mit den sprachwissenschaftlichen" (Humboldt 1836–9:220).

[9] "En las historias de las naciones todos sus escritores pretenden llegar al estado primitivo de ellas con la relación de sus hechos. Ellos tomando los varios rumbos que les presentan los monumentos geroglificos, mitología, tradición, calendarios, alfabetos ó escritura, y las noticias de la historia de las naciones vuelan por inmensos espacios de los obscuros tiempos de la antigüedad."

The difference between the Feejean [Fijian] language, and the other Polynesian dialects forms the second topic which you have suggested. The points of contrast will be more clearly seen, by first mentioning the points of resemblance between that language and the other members of the Polynesian family. If the Feejeean's consanguinity with the other South Sea Islanders were clearly established, the affinity between his language and theirs would be an inference from that fact. But as it is not the design of this statement to attempt to prove that the Feejeean is the offspring of that branch of the human family to which all the other Polynesian tribes belong, I shall merely enumerate some of the arguments which appear to favour that hypothesis. The principal arguments are his physical conformation, – his political and religious customs, – his unequivocally Polynesian vices, and his language. The last argument relates to the present subject. That the language of Feejee and all the dialects of Polynesia with which I have any acquaintance are not different members of different families but members of one and the same family, and of consequence of one common origin appears demonstrable on the following grounds. 1st. There are words which are common to all the dialects; 2ly. There are other words which have evidently sprung from one stock, but have assumed a different appearance in adaptation to the genius of the people; 3ly. All the dialects are characterized by the same peculiarities, and 4ly. the same idiom [overall grammatical structure and typology] prevails among all the languages. (Quoted in Schütz 1972:4)[10]

In our attempts to sort through methods and classifications of languages, it will be important to keep in mind exactly what evidence was presented in each case, for what goals it was utilized, and which methods were employed. It will be necessary to concentrate on linguistic considerations while separating out, so far as possible, non-linguistic evidence and interests.

As is evident already from some of the quotations cited above, the terminology encountered in earlier works both varies much and often does not match that employed today. It is important not to read present-day content that was not in the minds of the original authors into past work, nor to misread seemingly similar terms when they do not in fact refer to the same thing now as then. For example, we have seen in these quotes reference to "analogy" in the sense of sound correspondence (though not precisely equivalent to today's sense of that term), to "radical" for "genetic relationship," and reference to "dialect" where today's sense would be "independent language," and to "language" with today's sense

---

[10] See Chapter 7 for more on Cargill. Debates at around the close of the nineteenth century concerning the basis for classifying the tribes of North America for the Smithsonian Institution's Museum exhibits reflect a more recent instance of the failure to separate language and "race," and of linguistic classification serving as a primary basis for the historical and anthropological understanding of peoples in general. Smithsonian's Otis T. Mason leaned towards basing the Museum's displays on a classification along the unilinear lines assumed in social Darwinism, that is, assumed "progress" in social evolution from "savagery" to "barbarism" to "civilization" (cf. Hinsley 1981:99). However, at that time Major John Wesley Powell, founder of the Bureau of American Ethnology, produced the first comprehensive classification of American Indian languages of North America (Powell 1891), after which Mason was stimulated to adopt for the museum the principle of displaying the peoples of Native America organized along language family lines (Hinsley 1981:110). In fact, the primary value of Powell's linguistic classification was persistently seen to be its utility for ethnological classification.

of "language family." Johann Reinhold Forster (1778), naturalist on Captain
Cook's second voyage (writing slightly before Jones' "philologer" statement
was published), uses such terminology in a way that makes its different sense
clear (see Chapter 6); he also reflects the assumed race–nation–language cor-
relation of the times:

It has always been customary among the more critical and chaste historians, to reckon
all such nations as speak the *same general language* [members of the same language
family], to be of the same tribe or race . . . By the SAME GENERAL LANGUAGE, I
understand all the various subordinate dialects [languages] of one language [family]. No
one for instance, acquainted with the subject, will deny, that the Dutch, Low-German,
Danish, Swedish, Norwegian, Icelandic, and the English (in respect to such words as owe
their origin to the Anglo-Saxon) are dialects [languages] subordinate to the *same general
language* [language family], together with the present High-German, and the remains
of the Gothic in Ulfila's New Testament. But allowing for this, yet we find that these
dialects differ in many respects . . . many words, however, though somewhat modified,
always preserve enough of the original type, to satisfy the critical etymologist, that they
belong to the *same general language*, as subordinate dialects. This short digression will
therefore open a way to prove that the five races [Polynesian groups], which I enumerated
as belonging to the first tribe, are really descended from the same original nation; for
they all speak a language [individual but related languages] that has in the greatest part
of their words, a great and striking affinity. (Forster [1778], in Rensch 1996:184–5)

1.5   Similarity vs. proof of language relationships

In the chapters to follow, we will see that observation of similarities among
languages is the logical and most frequent starting point for attempts to demon-
strate relationships among languages. Unfortunately, we will also see that some
scholars' methods stop at the identification of similarities among compared lan-
guages, where they assume this is sufficient to demonstrate that the languages
under comparison are related to one another by virtue of descent from a common
earlier ancestor. However, assembling similarities among languages is just the
beginning. Similarities can be due to several things other than just inheritance
from a common ancestor (genetic relationship). The following are the major
explanations for similarities among languages (see Chapter 7):
1.   Accident (chance, coincidence).
2.   Borrowing (language contact).
3.   Onomatopoeia, sound symbolism, and nursery forms.
4.   Universals and typologically commonplace traits.
5.   Genetic relationship – inheritance from a common ancestor.
As will be seen in the chapters of this book, in order to establish a plausible
hypothesis of genetic relationship it is necessary to eliminate other possible
explanations (1–4), leaving a genetic relationship (5) the most likely. The core
of the generally accepted methodology for investigating possible relationships