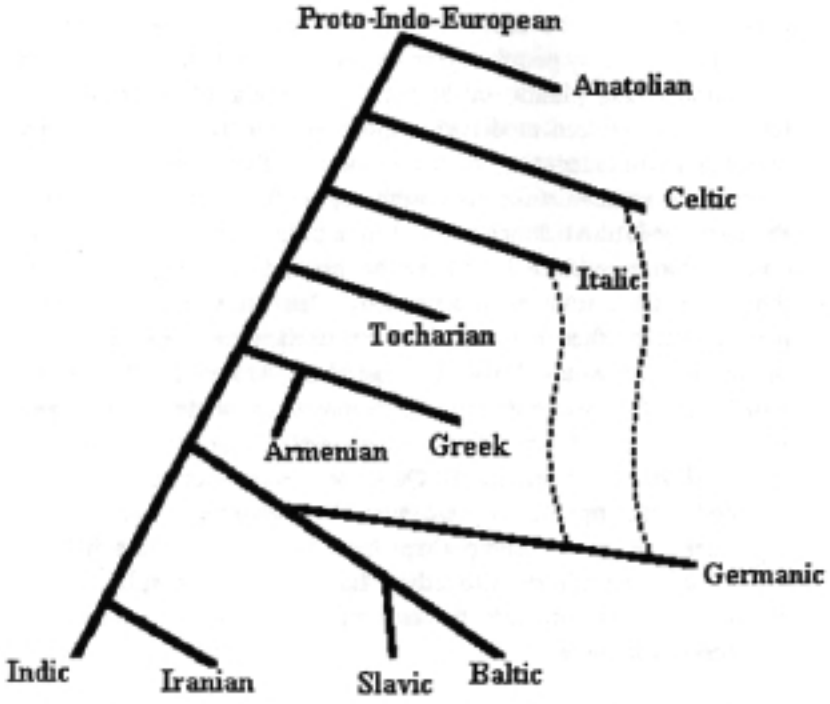
Computers and Linguistics:
The Indo-European Family Tree by R.L. Fowler

Languages, like people, are related to one another, and their relationships can be drawn up in the form of a genealogical chart. Wherever you find two languages with similar vocabularies and grammatical structures, you can be sure that they were descended from some common ancestor, no matter how widely separated they might be in space and time. The greater the separation, of course, the harder it becomes to detect the relationships. When a group of speakers emigrates from the common homeland to a new location, they will for a time continue to speak the same language as their former compatriots; but eventually, little differences in pronunciation and idiom will creep in, which over a long period of time become so great that a wholly new language is born. We can observe in the modern world many differences between English as spoken in the United Kingdom, Canada, the United States, South Africa, and Australia. Modern telecommunications act as a brake preventing these dialects of English from becoming altogether different languages, but the natural process of differentiation is still operating. If we ever suffer the collapse some apocalyptic visionaries predict, future historians will pick over the surviving documents and conclude quite readily that the occupants of the countries I mentioned, at least those who came to them at a crucial stage in their formation as modern nations, were descended from inhabitants of the British Isles.



The existence of the Indo-European family of languages was established as long ago as 1816 by a linguist with the wonderful name of Franz Bopp. At the age of twenty-five years he discovered that Sanskrit, Greek, Latin, ancient Persian and the Germanic languages are all descended from a common ancestor, which English-speaking linguists now call "Indo-European". It was of course clear that some members of the group were quite closely related to each other, while others displayed a greater number of idiosyncrasies and seemed to stand by themselves; it was also obvious that some (such as English) developed considerably later than others. In the end everyone accepted Bopp's insight, and the efforts of linguists were devoted to establishing exactly *how* these languages were to be arranged in the family tree.

The task is beyond the powers of any human, because the amount of data is overwhelming. The problem is to identify the common features that are truly significant out of hundreds of thousands of overlaps between this language and that. An especially complicating factor is that some similarities are independently developed by different

languages and don't descend from a shared ancestor. A similar problem is faced by people trying to draw up family trees for species of animals and plants, or by scholars trying to determine the relationship between medieval manuscripts on the basis of shared mistakes. In the latter case, the presumption is that if two manuscripts display the same mistake, they were copied from the same ancestor; this ancestor in turn belongs to a different branch of the tradition from other manuscripts which don't have the mistake. However, scribes can make the same mistake independently, and if you grouped your manuscripts together on the basis of one of these independently-made mistakes, you would be—well, mistaken. Moreover, the normal situation is that, while manuscript A may share some mistakes with manuscript B, it will share others with manuscript C—which in its turn shares others with manuscript B! Quite often, too, a scribe will consult a second manuscript after copying a text from his main exemplar, thus introducing readings—and perhaps more mistakes—from a different branch of the tradition altogether. Languages can display similar cross-fertilization, making the task of drawing up the family tree hopelessly complicated.

It looks like a job for a computer. Until recently, however, this problem exceeded the enormous calculating power of even the most advanced machines. Fortunately a breakthrough has been made which appears to have solved the problem. According to the Toronto *Globe and Mail* of January 6, 1996, computer scientist Tandy Warnow at the University of Pennsylvania has developed an algorithm which allows the computer to perform this operation with much greater efficiency. Linguists Ann Taylor and Don Ringe used this program to draw up the genealogy of all languages descended from Proto-Indo-European, the putative progenitor of the family, spoken in the third millennium BC and now (of course) totally extinct.

The chart not only shows clearly the familial relationship of the various languages, but also arranges them in chronological sequence, with the oldest ones at the top. First to break away was the Anatolian group, which includes a number of now-extinct languages such as Hittite, Luwian, Lycian, and Lydian. These languages were spoken in the western parts of modern Turkey. Next came the Celtic group, which includes Welsh, Gaulish (extinct), Irish and Scottish Gaelic, among others. That these two groups pull down the top two spots on the chart lends some support to those who place the homeland of the speakers of Proto-Indo-European somewhere in Eastern Europe. Unfortunately, archaeologists have yet to find any trace of these people. The next breakaway bunch are the Italic languages: Latin chief among them, but also Oscan and Umbrian, languages which boasted many speakers in ancient Italy but which have now disappeared. The Romance languages descended from Latin (Italian, French, Spanish and others) must be placed at this point in the tree. Interestingly, the Italic languages stand higher up in the tree than Greek, even though our oldest evidence for the Greek language, the famous Linear B tablets of the fifteenth to the thirteenth centuries BC, antedates considerably our oldest evidence for Latin, and antedates even more our oldest evidence for Celtic languages. The linguists theorize that the German group, who migrated after the Celtic and Italic groups but subsequently came back into contact with them, acted as a catalyst on these two and produced, at a late stage, Latin, Gaelic, and the others mentioned above. The similarity between the Celtic and Germanic languages is readily seen and the contact of the groups a matter of historical record, but the theory is a little harder to accept in the case of Latin. At any rate, this secondary influence (the "cross-fertilization" I spoke of earlier) is indicated by the broken lines.

Next comes another totally extinct group, Tocharian or Tokharian, originally domiciled in West China; we see now the first group of descendants to migrate eastwards. Moving down the chart, we find Armenian and Greek sharing a common ancestor. This is an interesting result, because some linguists had placed Armenian in the Slavic group. Archaeology shows an influx of newcomers who shared a common culture (involving, among other things, horses) arriving in the Greek peninsula and in Troy towards the end of the third millennium BC; they are commonly identified as Indo-Europeans (and realizing this fact helps to explain why Homer makes the Trojans so much like the Greeks; apart from poetic convenience, it may have reflected the historical reality of a shared ancestry). Were these the ancestors of the Armenians?

All but last comes a group that further splintered into two subgroups, one of which subdivided again. In the Germanic group we find the various kinds of German, Old Norse (now extinct) and all the Scandinavian languages except Finnish, Dutch, Flemish, Afrikaans,

Frisian, and our own English. This group, as explained above, exerted a secondary influence on the Celtic and Italic languages. The Baltic group includes Latvian and Lithuanian, and in the Slavic group, which is large, are to be found Russian, Ukrainian, Bulgarian, Serbian, Czech, Polish, Macedonian, Albanian, Illyrian and Thracian (both extinct), Slovenian, and others.

Last to split were Iranian (modern Iranian, the extinct Avestan, Persian) and Indic, which comprises many of the languages of the modern sub-continent; the main ancient representative is Sanskrit.

The results are fascinating, all the more so in that one of the principal investigators, Dr. Ringe, had for years resisted the hypothesis that the Anatolian group was the first to break away. When his own program again and again said that this was the case, even he had to concede the argument. Computers, after all, have no personal stake in the outcome of academic debate, even one that's nearly two hundred years old.