0 words (limit of 3,000)

Brett Kessler
Washington University in St. Louis
Psychology Department
Campus Box 1125
One Brookings Drive
St. Louis MO 63130-4899 USA
1-314-935-8839
bkessler@wustl.edu

LANGUAGE FAMILIES

A language family is a set of languages that developed from the same ancestral language. The best-known example is the Indo-European family, which comprises over a hundred languages that, even in premodern times, extended from the Indian subcontinent to northwestern Europe. This family is well known not only because it contains many of the world's most widely spoken languages, such as Bengali, English, French, German, Hindi, Portuguese, and Russian, but also because it was the main focus of research in the nineteenth century, when linguistics was established as a modern science. However, many other language families are subjects of intense research today, including:

- Afro-Asiatic, which includes Arabic, Hebrew, and several languages of northern Africa, including Ancient Egyptian, Hausa, and Somali
- Algic, which includes several native languages of North America, such as Wiyot, Yurok, Cheyenne, Ojibwa, and Shawnee
- Austronesian, which includes over a thousand languages spoken from Madagascar to Polynesia, such as Bahasa Indonesia, Fijian, Tagalog, and Tahitian
- Dravidian, which includes most of the non-Indo-European languages of India, such as Malayalam, Tamil, and Telugu

- Niger-Congo, which includes most of the languages of sub-Saharan Africa, including Igbo, Swahili, and Zulu
- Sino-Tibetan, which includes Burmese, Chinese, and Tibetan
- Tupi, comprising several languages of South America, including Guarani
- Uralic, which includes most of the non-Indo-European languages of Europe, such as Estonian, Finnish, Hungarian, Nenets, and Sami (Lapp)

This list is only a sample of the hundreds of known language families and of the languages included in each. For a comprehensive listing, see Gordon (2005).

Familial metaphors are the standard terms of art. Languages in the same family are said to be genetically related to each other. A language from which other languages developed is called an ancestor, or parent, of those languages. Words that descend from the same form in an ancestral language are related, or cognate. This homey terminology is undergoing some competition with that of modern cladistics as used in biology, but certain linguistic concepts do not translate well. In biology, it is understood that all biological taxa are related to each other, and family is but a mid-level taxon. Linguists assume that relationships between languages must be proved, and a language family is a maximal taxon. In principle, even isolates – languages that do not group with other languages – can trivially be considered families by reinterpreting their *DIALECT*s as separate languages.

The common ancestor of an entire language family is assigned a name by prefixing *Proto-* to the name of the family, as in Proto-Indo-European and Proto-Afro-Asiatic. The place where the protolanguage was spoken is called the homeland of the language family.

The study of language families is part of *HISTORICAL LINGUISTICS* and is contextualized within a particular model of *LANGUAGE CHANGE*: divergence. When innovations in one part of a language community fail to spread to other parts, differences accumulate until the community can be said to speak different languages. It is this historical process that language-family theory is meant to model. But, perhaps because language families are commonly illustrated by showing similarities between languages (e.g., English *mouse* is cognate with Latin *mus*), the idea arises that relatedness is about similarity between languages. In fact, there is no requirement that cognates be similar at all (e.g., English *two* is related to Armenian *yerku*), and many sources of similarity are disavowed as being irrelevant to the model. These include borrowing (see *CONTACT, LANGUAGE*), onomatopoeia, universals (*ABSOLUTE AND STATISTICAL UNIVERSALS*), and chance similarities.

The study of language families typically involves one or more of the following enterprises:

- Demonstrating that languages are related
- Reconstructing the common protolanguage
- Subgrouping the languages by hypothesizing intermediate ancestors
- Associating linguistic data with historical and archaeological data

The following sections first describe the traditional and still dominant methods for pursuing these tasks and then sketch and evaluate some new methodologies.

Traditional Methods

The traditional technique is the *COMPARATIVE METHOD*. The linguist studies characteristics that rarely recur across languages, such as grammatical paradigms and the

associations between sound and meaning in morphemes. Efforts are made to discard loans and onomatopoeia, although the former is a difficult and often intractable problem. Matching morphemes across languages by meaning, one looks for recurrent sound correspondences. For example, English *f* corresponds to Latin *p* in *father=pater*, *feels=palpat*, *few=pauca*, and many other words. If a large number of recurrent correspondences are found, the languages are related. The recurrences are also used to reconstruct the protolanguage (see *HISTORICAL RECONSTRUCTION*).

After a language family is identified, the next step is subgrouping, identifying the branches or groups within the family. Subgrouping seeks to uncover the history of the divergence (cladogenesis) of a language family. If the family contains three or more languages, the linguist looks for evidence that some proper subset of those languages may have descended from an intermediate common ancestor. This is done by looking for shared innovations (synapomorphies) – sound changes or new words or grammatical constructions that were not in the ancestor language but are found in two or more of the descendant languages. For example, the fact that English, German, Swedish, and several other languages have *f* where Proto-Indo-European had *p* is a shared innovation that indicates that those languages may have a shared intermediate ancestor that underwent this change; otherwise we would have to assume that each of those languages separately innovated the change of *p* to *f* or borrowed the innovation from another language. In fact, the preponderance of evidence supports such an intermediate language and a branch (clade) of languages descending from it: the Germanic languages. Other branches of Indo-European include the Balto-Slavic (including Bosnian, Lithuanian, Polish, and Russian), Celtic (including Breton, Irish, and Welsh), Italic (including Latin and the

Romance languages), Indo-Iranian (including Bengali, Farsi, Pashto, and Urdu), and the extinct Anatolian branch, which included Hittite, Luvian, and Lycian. Several other languages, including Greek, Albanian, Armenian, and half a dozen extinct languages, do not share an agreed intermediate branch at all.

Associating language history with external facts entails pinning the protolanguage to a particular time and place – its homeland – and demonstrating how it spread from there. The time depths under consideration mean that written records are rarely if ever available. The primary linguistic tool is to look for words found in multiple branches of a language family and exhibiting all the regular sound correspondences; they are assumed to date back to the protolanguage and therefore to name objects found in its environment. For example, a pan-Indo-European word for 'wheel' suggests the protolanguage split up no earlier than the invention of the wheel, some six thousand years ago (Mallory 1989). Another technique is to look for areas of greatest linguistic diversity. The fact that the Austronesian languages are much more diverse in Taiwan than anywhere else supports the theory that they developed there longest; that is, that Taiwan was the homeland for the Austronesian family (Blust 1999). A third technique is to seek archaeological evidence of population movements that may have disseminated a language family. In the case of Austronesian, knowledge of how people spread through the Pacific and Indian Oceans is consistent with the theory of a Taiwan homeland. In the case of Indo-European, it has often been noted that early adopters of horse-drawn wheeled chariots would be in an ideal position to spread their languages through much of Europe and Asia. A well-received theory points to the chariot users who lived in the Pontic–Caspian region about six thousand years ago (Mallory).

Challenges to the Traditional Method

The traditional comparative method is still the basic framework within which language families are researched, but it is not perfect. It is a complicated process that demands a great deal of knowledge about all the relevant languages. It can be misled by loanwords, and it offers little guidance in distinguishing true shared innovations (synapomorphies) from independent identical innovations (homoplasies). The linguist must constantly decide whether multiple languages could have undergone a particular change independently and how likely they would be to have borrowed it. In reality, of course, anything that happens once can happen twice, and there is nothing that is not subject to borrowing (Thomason and Kaufman 1988). The true solution is probabilistic, but hard numbers are lacking, and the investigator is often left with unlikely cladograms like the fifteen-way branching tree of Indo-European.

Another disappointment is that little progress has been made in the past century in pinning down the Indo-European homeland or proving that additional languages are related to English – topics of recurrent interest among linguists, archaeologists, and enthusiasts alike. More disappointing is that when linguists have claimed that language families such as Uralic are related to Indo-European – such groupings often being given the Eurocentric name Nostratic (‘our’ family) – the methodology has given no firm guidance as to how significant the evidence is, with the result that many linguists find themselves uncomfortably agnostic on whether Nostratic has been proved or not. Unlike in modern experimental sciences, there are no statistical techniques for estimating the probability that the number of correspondences found is due to a real relationship between languages rather than to chance. Rules of thumb were developed to provide

some guidance; a typical piece of advice is to treat words as potentially cognate only if at least three of their consonants are found in recurrent correspondence sets. But such rules are very approximate, not tailored to the specific structures of the languages at hand, and they discouraged linguists from applying the method to languages with short morphemes.

Greenberg addressed several of these concerns with a technique called multilateral comparison. Tables are constructed listing the translation equivalents for many concepts in many different languages. It is claimed that the tabular layout itself makes the relationships between the languages, even their correct subgrouping, patent. Using this technique, Greenberg presented an analysis of the languages of Africa (1963), which is now considered standard, then went on to hypothesize language families that lumped established families together into much larger families – what became known as deep linguistic relationships. The dozens of families and isolates of the Americas were reduced to three families (1987); Indo-European, Uralic, Japanese, and several other families were lumped into a family called Eurasiatic (2002). Multilateral comparison has proved popular among enthusiasts, in part because it requires no special language expertise, in part because it appears to reveal many new, deep, relationships. Unfortunately, there is no way to evaluate a methodology that simply calls for contemplating raw data until patterns emerge.

Several researchers have shown, however, that some of Greenberg's key ideas can be transformed into algorithmic (reproducible) methodologies that introduce to language family research the benefit of statistical significance testing. Oswalt's procedure (1998) minimized experimenter bias by requiring that a specific concept list be used and that one specify in advance specific criteria for measuring degree of similarity between two

languages. Baxter and Manaster Ramer (2000) added reliable significance testing procedures based on randomization tests. Kessler and Lehtonen (2006) adapted the technique to handle multiple languages in a single test, informally confirming Greenberg's claim that such large-scale comparisons are inherently more powerful than two-language comparisons. Ringe (1992; see Kessler 2001 for extensive discussion and methodological refinements) measured not similarity but the number of recurrent sound correspondences. This has the advantages both of being closer to the traditional comparative method and of generating correspondences useful for subgrouping and reconstruction. Disappointingly, however, none of these neo-Greenbergian techniques found evidence for the deep relations that were advertised for the original, impressionistic, method.

Other new techniques have concentrated on subgrouping. Lexicostatistics (Swadesh 1955) was an early attempt to facilitate subgrouping and also assign dates to protolanguages. The idea was that if languages replace a constant number of words per century with new words, then by measuring what percentage of a list of words is cognate between languages, one could calculate when the languages diverged and even construct a family tree. Although these assumptions were mostly wrong and were therefore rejected by most linguists, many people still use lexicostatistical techniques as a rough indication of a language's history in the absence of more compelling data.

Arranging many shared innovations into a binary tree is an extremely laborious undertaking, especially given the possibility that some identical innovations are independent (homoplastic). The recent development of computational cladistic methods similar to those used in biology (e.g., Ringe, Warnow, and Taylor 2002) is a tremendous

advance in helping the linguist find optimal trees. In addition, several solutions to the problem of borrowing have emerged in the form of programs that construct networks instead of trees. Shared innovations that cannot be cleanly attributed to a shared ancestor are taken as evidence of contact, obviating somewhat the need to make a priori judgments about whether borrowing was involved (e.g., Bryant, Philimon, and Gray 2005; Nakhleh, Ringe, and Warnow 2005).

The problems of finding homelands and tracing the spread of languages still requires resorting to data that are often suggestive but not definitive. Renfrew (1987) added a new perspective when he theorized that languages may be spread by the movement of culture rather than by the movement of people. He suggested that Indo-European languages were spread from Anatolia along with the adoption of agriculture. Most linguists have not accepted this theory, in part because it is incompatible with such linguistic data as an Indo-European word for the wheel, which postdates the spread of agriculture by millennia. Recently, further data are afforded by genetic analyses of populations (*GENETICS AND LANGUAGE*). The presence of a Pontic genetic component in Europe is compatible with the idea that invaders from the Pontic–Caspian region brought Indo-European languages into Europe (Cavalli-Sforza, Menozzi, and Piazza 1994).

Prospects

Recent computer techniques add simplicity, reproducibility, and quantitative rigor to methodologies for proving relationships between languages, but so far there has been no noticeable increase in power over what experts are able to do by hand. Failure to corroborate the sort of deep relationships conceived by Greenberg may mean that better

techniques need to be developed; or that the languages are not in fact related; or that the answer is unknowable. Because languages are always changing, they constantly lose information that links them to their relatives; there must come a point at which any remaining commonalities between languages are indistinguishable from chance levels. But even if the more pessimistic predictions are true and new methods are unlikely to greatly expand intensively studied families like Indo-European, they may greatly ease new analyses of lesser known languages.

New computerized cladistic methods are likewise already aiding the analysis of complex language families and are providing Indo-Europeanists food for thought. However, the development and application of such algorithms could benefit from compiling and deploying data about the probability of various types of linguistic innovations and borrowings.

To date, the new methodologies have not been adopted by most practitioners. While it is easy to fault established researchers for conservatism, it is also true that quantitative methods typically cannot take into account the diverse types of information linguists are accustomed to reasoning with. Fortunately, the emerging partnerships between linguists and cladists should help bridge the gap between old and new approaches and lead to the widespread adoption of hybrid methodologies.

—Brett Kessler

Works Cited and Suggestions for Further Reading

Baxter, William H. and Alexis Manaster Ramer, 2000. “Beyond Lumping and Splitting: Probabilistic Issues in Historical Linguistics.” In *Time Depth in Historical*

Linguistics, ed. Colin Renfrew, April McMahon and Larry Trask, 167–188. Cambridge, England: McDonald Institute for Archaeological Research.

Blust, Robert. 1999. "Subgrouping, Circularity and Extinction: Some Issues in Austronesian Comparative Linguistics." In *Selected Papers from the Eighth International Conference on Austronesian Linguistics*, ed. E. Zeitoun and P. J. K Li, 31–94. Taipei: Academia Sinica.

Bryant, David, Flavia Filimon, and Russell D. Gray. 2005. "Untangling Our Past: Languages, Trees, Splits and Networks." In *The Evolution of Cultural Diversity: A Phylogenetic Approach*, ed. Ruth Mace, Clare J. Holden, and Stephen Shennan, 69–85. London: UCL Press.

Cavalli-Sforza, Luigi Luca, Paolo Menozzi, and Alberto Piazza, 1994. *The History and Geography of Human Genes*. Princeton University Press

Gordon, Raymond G., Jr. (ed.). 2005. *Ethnologue: Languages of the World*. 15th ed. Dallas, TX: SIL International. Content also available online at http://www.ethnologue.com/

Greenberg, Joseph H. 1963. "The Languages of Africa." *International Journal of American Linguistics*, supplement 29(1), pt. 2.

Greenberg, Joseph H., 1987. *Language in the Americas*. Stanford (CA): Stanford University Press.

Greenberg, Joseph H., 2002. *Indo-European and its Closest Relatives: the Eurasiatic Language Family: Lexicon*. Stanford, CA: Stanford University Press.

Kessler, Brett, 2001. *The Significance of Word Lists.* Stanford, CA: Center for the Study of Language and Information.

Kessler, Brett and Annukka Lehtonen. 2006. “Multilateral Comparison and Significance Testing of the Indo-Uralic Question.” In *Phylogenetic Methods and the Prehistory of Languages*, ed. P. Forster and C. Renfrew, 33–42. Cambridge, England: McDonald Institute for Archaeological Research.

Mallory, J. P. 1989. *In Search of the Indo-Europeans: Language, Archaeology and Myth.* London: Thames & Hudson.

Nakhleh, Luay, Don Ringe, and Tandy Warnow. 2005. “Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages.” *Language* 81: 382–420.

Oswalt, Robert L., 1998. “A Probabilistic Evaluation of North Eurasiatic Nostratic.” In *Nostratic: Sifting the Evidence*, ed. J. C. Salmons and B. D. Joseph, 199–216. Amsterdam: Benjamins.

Renfrew, Colin. 1987. *Archaeology and Language: The Puzzle of Indo-European Origins.* London: Pimlico.

Ringe, Don A., Jr., 1992. *On Calculating the Factor of Chance in Language Comparison.* Philadelphia, PA: American Philosophical Society.

Ringe, Don, Tandy Warnow, and A. Taylor. 2002. “Indo-European and Computational Cladistics.” *Transactions of the Philological Society* 100: 59–129.

Swadesh, Morris, 1955. “Towards Greater Accuracy in Lexicostatistic Dating.” *International Journal of American Linguistics* 21: 121–37.

Thomason, Sarah Grey and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley, CA: University of California Press.