GS
Garland Science

HUMAN EVOLUTIONARY GENETICS

second edition

Jobling · Hollox · Hurles · Kivisild · Tyler-Smith

Just how closely is English related to Danish, or to Gujarati? The idea of putting a number on the distance between languages was first explored in 1831 by Samuel Rafinesque, who invented **lexicostatistics** in order to win a competition in Paris of the Société de Géographie, to determine the origin of Asiatic **negritos**. Since there were no other contenders, he was awarded a *médaille d'encouragement* worth only 100 francs rather than the advertised gold medal of 1000 francs. Rafinesque's negative finding showed the languages of disparate negrito peoples to be unrelated. It was never published, but his method was popularized by Jules Dumont d'Urville (**Figure 1**), who headed the jury.

Dumont d'Urville describes Rafinesque's technique thus: Between two terms expressing the same idea in two different languages we assign six degrees of relationship: 0 for completely unrelated terms, 1/5, 2/5, 3/5, or 4/5 for partially related terms, and 5/5 (or 1) for terms that are identical or nearly so. The total number of correspondences is then divided by the number of words compared. A score of 135/5 or 27 in a list of 45 words gives us a relationship of 0.60, whereas a score of 35/5 or 7, divided by 45, would give 0.15.

In the later nineteenth century, the idea that lexicostatistics could yield a separation date for branches of a language family gave rise to **glottochronology**. Following extensive criticism, lexicostatistics and glottochronology are anathema to most historical linguists (apart from a maverick subset) today.

Here are the criticisms in a nutshell. The distinction between vocabulary that is basic and words that are not is vague. In fact replacement of so-called basic vocabulary happens quite frequently. Rendering terms from different languages into English often masks substantive differences of meaning. Lexicostatistics often fails to distinguish between borrowed words (**identity by state**), and inherited words (**identity by descent**).

Figure 1: Jules Dumont d'Urville (1790–1842)

Recognizing **cognates** (words sharing a common origin) through systematic sound correspondences and clear similarities in form and meaning presumes detailed knowledge of language history. Many apparent cognates represent mere chance look-alikes. Languages change at highly variable rates. Back in 1850, Schleicher observed that languages spoken in tranquil backwaters (for example, Lithuanian, Georgian) change slowly, whereas languages in a constant maelstrom of social upheaval (for example, English, Mandarin) change quickly. Finally, laws about how speech sounds (**phonemes**) change regularly through time are vital, and correspondences between morphological systems and **grammatical** markers have far greater weight than mere lexical correspondences.

The math has undergone refinement, but Gray and Atkinson's[4] revolutionary use of **Bayesian** glottochronology to assess hypotheses for the origins of Indo-European (see **Section 12.5**) was greeted with indignation in conservative linguistic circles. In addition to objections against the misrepresentation and misinterpretation of language data, the methodological criticisms outlined above were reiterated, especially the issue of borrowings, chance resemblances, and false cognates.

Bayesian analysis of word correspondences is less problematic in Austronesian, where dispersal of the language family largely involved the colonization of previously uninhabited islands and therefore less contact. The claim[6] that the misidentification of borrowed vocabulary versus inherited word-forms does not compromise the validity of lexicostatistical findings is generally rejected by linguists. The other criticisms listed above also remain to be addressed.

In 1848, August Schleicher introduced the family tree model for language phyla by analogy to the **phylogeny** of biological species. In conventional models, whole integer values could be assigned to the nodes in the phylogeny, based on the intuitions of knowledgeable historical linguists. Yet linguists generally doubt that numbers derived from readily manipulable though complex mathematical models, based on simple and sometimes false lexical comparisons, can be very meaningful.

Nonetheless, the mathematization of linguistic phylogeny seems inevitable. The way forward is to accommodate the criticisms identified by linguists and to tweak the model, as Dunn et al. have done.[3] Mathematical approaches to linguistic phenomena are producing ever more intriguing results. A Bayesian analysis of lexical change may have shed some new light on the shape and rate of language evolution.[5] However, sometimes a headline-grabbing finding is just a foregone conclusion that could have been foretold by anyone familiar with, say, the phoneme inventories of **Khoisan** languages.[1] Numbers can be useful, but they actually ought to reflect realities accurately if they are to tell us something new and meaningful.

George van Driem, Department of Linguistics, University of Berne