

Numerical Methods in Linguistics

An Introduction to Glottochronology

Raamesh Gowri Raghavan

Historical linguistics deals with the evolutionary relationships of languages to one another. Of the many methods of analysis used in this field, the numerical method of *glottochronology* has been useful (as well as controversial) among linguists for its simplistic approach of comparing lists of supposedly 'core' (or basic) words between languages, and finding relationships by determining the percentages of cognates between them. Here I discuss the method with its limitations, and the recent application of sophisticated techniques to overcome them.

I was taught in school that the Indus Valley Civilisation that flourished 5000 years ago declined after the so-called 'Aryan Invasion' from the Northwest. The origin of Sanskrit-speaking civilisation in India is dated to that invasion, and this theory has held some credence due to archaeological and linguistic evidence. For example, pottery of similar design has been found in post-Indus sites in both India and Afghanistan. Avesta, the original language of the Zoroastrians [1], is known to have replaced Elamite in Iran at just around the same time as the first Sanskrit texts were recorded in India. Moreover, writings in these two languages can often be matched word for word, for example

Avesta: *təm amavantəm yazatəm surəm damohu səvistəm mi?rəm
yazai zao?rabyo*

Sanskrit: *tam amavantam yajatam suram dhamasu savistham mitram
yajai hotrabhyah*

(The powerful God, the most powerful Mithra in the world of the creatures, I will adore him with libations; Zoroastrian chant: Avesta: Yasna 10.6).



Raamesh Gowri Raghavan is presently project assistant at the Division of Biochemical Sciences, National Chemical Laboratory, Pune. He works on genetic diversity and origins of the Indian population using mitochondrial DNA markers, and also studies mitochondrial disorders. He is interested in natural history and evolution, linguistics and the origin of humans.

Keywords

Indo-European, glottochronology, Anatolian expansion, migration, Sanskrit, Proto-Indo-European.



Box 1. Hindi and Farsi – Close Cousins

Although not many native Hindi speakers realize it, Farsi is a close cousin of Hindi and undoubtedly one of the easiest foreign languages for a Hindi speaker to learn. This is most obvious when considering basic words of day-to-day usage.

Hindi	Farsi	English
ek	yek	one
do	do	two
teen	seh	three
chaar	chahaar	four
paanch	panj	five
das	deh	ten
bees	beest	twenty
dil	dil	heart
pita	pidar	father
maata	maadar	mother

Indeed, despite a wave of ‘Arabization’ following the Islamic conquest of Iran, modern Farsi still retains many similarities with Sanskrit, as can be seen in the Farsi words ast (is), asb (horse), tishna (thirst), sheer (milk), musht (fist), kaar (work), musha (mouse), and dars (philosophy), to name just a few.

This similarity is because Avesta and Sanskrit both belong to the Indo-European family of languages, and indeed to a common branch of the family (*Box 1*). This fact was first discovered by the British judge Sir William Jones who founded the prestigious Asiatic Society in Kolkata in 1784, and noticed similarities between Sanskrit, Latin and Greek. Most of the languages of India, Iran and Europe belong to one or the other of the eleven surviving branches of this family, and are believed to have had a common ancestor called *Proto-Indo-European* (see *Box 2*).

However, for years together a dispute has raged over the relationships of these branches to each other, and the date and place of origin of Proto-Indo-European. Most linguists and historians advance two alternative hypotheses – The Anatolian Expansion and the Kurgan Expansion. The protagonists of the former believe that Proto-Indo-European was spoken in the Anatolian plateau (present day Turkey) and spread out with the development of agriculture nearly 9000 years ago. The Kurgan school,



Box 2. A Proto-Indo-European Fable: Owis Ek'wooskwe ([The] Sheep and [the] Horses)

Gwrreei owis, kwesyo wlhnaa ne eest, ek'woons espekēt, oinom ghe gwrrum woghom wegħontm, oinomkwe megam bhorom, oinomkwe ghmmenm ooku bherontm. Owis nu ek'womos ewewkwet:

"Keer aghnutoi moi ek'woons agontm nerm widntei". Ek'woos tu ewewkwont: "Kludhi, owei, keer ghe aghnutoi nsmei widntmos: neer, potis, owioom r wlhnaam sebhi gwħermom westrom kwrnneuti. Negħi owioom wlhnaa esti". Tod kekluwoos owis agrom ebħuget.

On [a] hill [a] sheep, that had no wool, saw horses, one [of them] pulling [a] heavy wagon, one carrying [a] big load, and one carrying [a] man quickly. [The] sheep said to [the] horses: "[My] heart pains me, seeing [a] man driving horses". [The] horses said: "Listen, sheep, our hearts pain us when we see [this]: [a] man, [the] master, makes [the] wool of [the] sheep into [a] warm garment for himself. And [the] sheep has no wool". Having heard this, [the] sheep fled into [the] plain.

This fable (Reproduced from [2]) was written by the 19th century German linguist August Schleicher using the hypothetical Proto-Indo-European language which he painstakingly constructed from existing Indo-European languages using detailed comparisons of vocabularies, sounds and grammars. *No tangible proof of the existence of this putative language has, however, been found yet.*

on the other hand, argues for a military expansion out of the steppes of southern Russia (site of the Kurgan culture) not more than 6000 years ago, after the horse was domesticated [2]. The concept of the 'Aryan invasion' of Iran and the Indian sub-continent is based on the Kurgan Expansion hypothesis.

Investigating language relationships is analogous to analysis of molecular phylogeny – comparing many words from different languages just like we compare genes from different organisms. Nevertheless, deciding whether two words in two languages have the same ancestor (i.e. they are orthologues; *cognates* being the linguistic term) is not straightforward and subject to the expert's judgement. For example, *door* (English), *dver* (Russian), *der* (Farsi) and *dvar* (Sanskrit) are obvious cognates. However, with *sona* (Hindi) and *zar* (Farsi), it is difficult to immediately realise that their common ancestor was *swarna* (Sanskrit-Avesta). Languages evolve faster than genes, and undergo changes not only in vocabulary, but also in grammar and sounds, e.g. the two 'sha's of Sanskrit are now indistinguishable, but they must have been differently pronounced since there are two characters in



Box 3. Grimm's Law

A more or less regular change in sounds from one Indo-European language to another. For example, *k, t, p* in Latin typically become *h, th, f* in English and *h, d, v/f* in German – Pater/Father/Vater or Frater/Brother/Bruder. Similarly 's' in Sanskrit typically becomes 'h' in Avesta. Such changes occur in all language groups, and are often instrumental in deciding cognates (see *Box 2*). This law was formulated by Jakob Grimm, who is better known as the co-author of the *Fairy Tales* by the Brothers Grimm.

the Sanskrit alphabet. Similarly, four distinct sounds, each with their own letter, in Arabic have become condensed into the sound of 'za' in Urdu, leaving Urdu with four letters for one sound. Hence, inferring language relationships is quite a task. However, some regular patterns of change in languages do exist, for example Grimm's law (see *Box 3*).

A method once popular with historical linguists was *glottochronology*, or the use of the tongue clock. One makes lists of words in a language, and searches for its cognates in other languages (see *Box 4*). The percentages of cognates between any two languages are then scored on a similarity matrix. This is called *lexico-statistics*. Finally, using linkage algorithms, dendrograms are drawn (see *Box 5*) to represent the relationships of the languages under study. Yet, glottochronology has the drawbacks of implicitly assuming that

1. the rate of change of words is the same for all languages; and
2. this rate is constant over time.

Box 4. The Swadesh List

Of the word-lists used in glottochronology, the Swadesh list is most widely used. This was devised by Morris Swadesh (1909-1967), an American linguist based in Mexico, who, incidentally, had nothing to do with our freedom movement in the 1950s. He chose two hundred and seven words which he considered 'basic' to the vocabulary of all languages (and hence most conserved), and collected their cognates from various Indo-European languages of Europe. The list has remained controversial from the start. It works very well for northern languages, but is not so applicable for studies of tropical languages, since it contains words such as ice and snow. However it is the list with the most exhaustive data, and is therefore very popular. Further information can be found at the following website: <http://www.ntu.edu.au/education/langs/ielex/IE-DATA1>.



Box 5. The Glottochronological Procedure

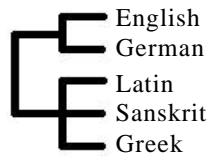
1. Consider the following table:

English	Latin	Sanskrit	Greek	German
Is	Est	Asti	Esti	Ist
Father	Pater	Pita	Patros	Vater
Mother	Mater	Mata	Matros	Mutter
God	Deus	Deva	Theos	Gott

2. Now construct a matrix of similarity. For example, there is a German cognate (word descended from a common ancestor) for every English word, whereas only three Latin cognates (Deus ? God) , and so on. Thus we get:

	English	Latin	Sanskrit	Greek	German
English	1+1+1+1=4				
Latin	1+1+1+0=3	4			
Sanskrit	3		4	4	
Greek	3	4	4	4	
German	4	3	3	3	4

3. We can see that English and German are closer, so we put them in one cluster, and the others in another cluster, and put all in a bigger set. So we get the tree:



This tree can be dated as mentioned in the text, but suffers from the glottochronological defects discussed. For example, it fails to capture the minor difference between Greek and Latin/Sanskrit (Theos vs. Deus/Deva), which is linguistically quite a significant consonant shift. *This is only an example tree, and is not factually correct.*

This might not always be true e.g. Greek has undergone relatively little change over thousands of years, while Avesta and Sanskrit have evolved into Farsi and Hindi, and many other languages. Moreover, by using only percentages of cognates, glottochronology misses out on subtle but important phonetic



details (see *Box 5* again). For these reasons, glottochronology has become unpopular with linguists, who instead use the complex though subjective Comparative Method, involving grammar and sounds, in addition to vocabulary.

With advanced computing methods such as Bayesian inference [3], Maximum Likelihood and Markov Chain Monte Carlo [4], the assumptions inherent in glottochronology can be dealt with. These quantitative methods are extensively used to analyse and calibrate molecular phylogenies. In a recent paper published in the journal *Nature*, Russell Gray and Quentin Atkinson applied these methods to linguistics for the first time to infer the phylogeny of 87 Indo-European languages and dialects [5]. The authors present their solution to two problems:

1. the inter-relationships of the branches of the Indo-European language family; and,
2. the likely time and place of its origin.

Cognates were coded in a binary fashion and the phylogeny was constructed using the Maximum Likelihood method. This method seems best for drawing trees when the different branches evolve at different rates, wherein apparent phenotypic similarity might obscure evolutionary links. For example, in *Box 5* Latin clusters with Sanskrit, but is actually more closely related to English (see *Figure 1*); this is because English has diverged more from the proto-language than either Sanskrit or Latin have, so it appears more distant from Latin than it really is. Further, Markov Chain Monte Carlo (MCMC) methods were used to test whether the given topology and branch-lengths could arise from random combinations of the data alone. Constraints based on historical information about the first writings in each language were used to filter sample trees from the MCMC, in order to attribute absolute ages of divergence of the language subgroups. *Figure 1* demonstrates the inter-relationships of the major families so derived. The tree includes Tocharian, an extinct language of Central Asia, and Hittite, the now-extinct language of the once-powerful city of Hattusas (modern Bogazköy in

Indo-European languages have spread in multiple migratory waves supplanting previous languages, so the date of origin of a language tells nothing about the date it came to be where it is today.



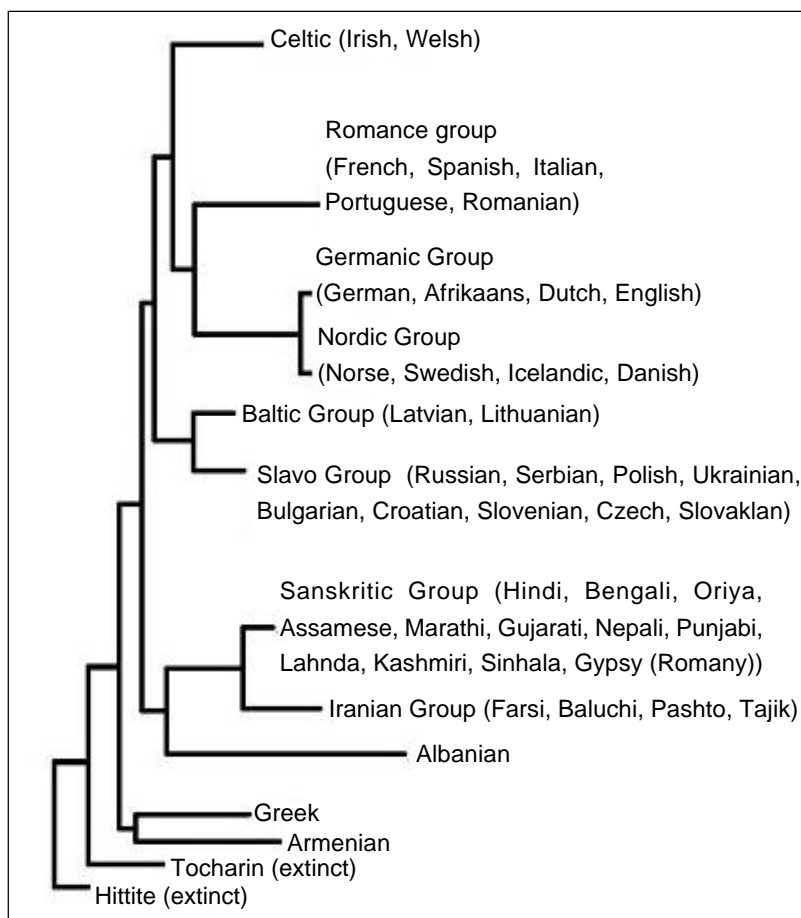


Figure 1. A simplified tree of Indo-European languages (Modified from [5]).

north-central Turkey), more than 9000 years ago. Hittite occurs at the root of the tree, suggesting that the Indo-European languages might have originated in Anatolia.

Thus, the more recent results tend to favour the Anatolian expansion hypothesis over the Kurgan Expansion one. The branch for the Sanskritic group diverges at nearly 7000 years ago. This might mean that Sanskrit may have come to India earlier than the designated date (about 1500 BC [6]) for the 'Aryan invasion' (if it ever happened). Unfortunately, we are restrained from drawing such a simple conclusion for two reasons. Firstly, the Indo-European languages have spread in multiple migratory waves supplanting previous languages, so the date of origin of a language tells nothing about the date it came

New languages arise when populations of speakers of the old language become isolated from each other.

to be where it is today. The Kurgan expansion generated a second wavefront of Indo-European languages, bringing the Baltic and later Slavic tongues to East Europe. Secondly, really large historic migrations have more often been agricultural and pastoral, rather than military in nature, beginning with marginalised groups venturing into new areas to avoid competition in their overpopulated homelands. These migrations are slow processes occurring over hundreds of years, determined by geographical and climatic factors such as mountain barriers and rainfall. New languages arise when populations of speakers of the old language become isolated from each other. Conflicts arise when these speakers of now different languages come into contact again and compete for the same resources. Often the more technologically advanced group gains political dominance and the whole population may end up speaking the language of the elite [7]. In this manner, India and Europe both have received multiple migrations leading to their present linguistic and cultural diversity.

One hopes that with application of the new and sophisticated computing methods, glottochronology will enjoy a revival and provide exciting solutions to many more linguistic puzzles.

Suggested Reading

- [1] www.ancientscripts.com provides an introductory study of linguistics for the layman.
- [2] Jared Diamond, *The Rise and Fall of the Third Chimpanzee*, Vintage, London, 1991. A book about Indo-European migrations.
- [3] Mohan Delampady and T Krishnan, Bayesian Statistics, *Resonance*, Vol. 7, No.4, pp. 27-38, 2002.
- [4] K B Athreya, Mohan Delampady and T Krishnan, Markov Chain Monte Carlo Methods, *Resonance*, Vol. 8, No.4, pp. 17-26, 2003.
- [5] R D Gray & Q D Atkinson, Language-tree divergence times support the Anatolian theory of Indo-European origin, *Nature*, Vol. 426, pp.435-439, 2003.
- [6] Jawaharlal Nehru, *The Discovery of India*, Oxford University Press, 1946.
- [7] Jared Diamond, *Guns, Germs and Steel - A Short History of Everybody for the Last 13,000 Years*, Vintage, United Kingdom, 1998. A useful book for an introduction to human migrations.

Address for Correspondence
Raamesh Gowri Raghavan
Division of Biochemical
Sciences, National Chemical
Laboratory
Pune 411008, India.
Email:azhvan@yahoo.co.in.

