INTERNATIONAL JOURNAL OF LANGUAGE EDUCATION AND APPLIED LINGUISTICS (IJLEAL)
ISSN: 2289-7208 e-ISSN: 2289-9294
VOL. 10, ISSUE 1, 17 – 27
DOI: https://doi.org/10.15282/ijleal.v10.3320

ORIGINAL ARTICLE

Genetic Relationship among Languages: An Overview

Ghayeth Ersheidat[1]* and Hafsa Tahir[2]

[1]Department of Translation, Faculty of Arts, Yarmouk University, Irbid, Jordon.
[2]Department of Biology, Faculty of Scence and Technology, Virtual University of Pakistan, Lahore, Pakistan.

ABSTRACT – This paper reviews the basic concepts of historical linguistics and the comparative techniques used by various linguists who studied Indo-European and American languages to determine a shared ancestry among languages. This paper also evaluates the major concepts of historical linguistics and the well-grounded theories and classifications that have guided and shaped the modern linguistic classification practices. For over one and a half century, historical linguists have been deducing the origins of different languages. Genetic classifications have been proposed for languages from all parts of the world and thus far, 142 language-families have been identified. Although all of these classification schemes are controversial in terms of their validity and reliability but with the progress in the field of bioinformatics, the problems in linguistic reconstruction have been greatly resolved. Therefore, the historical classification schemes that have been proposed earlier are being radically revised as further progress is made. It is suggested that, to develop further understanding of the typical pattern of language diversification and genetic classification of languages, more recent studies based on sophisticated bioinformatics and statistical techniques for linguistic data analysis should be reviewed.

ARTICLE HISTORY
Revised: 4 March 2020
Accepted: 9 March 2020

KEYWORDS
Historical comparative-method
Historical linguistics
Indo-European languages
Language family
Linguistic-tree
Proto-language

INTRODUCTION

For over one and a half century now, linguists have been tracing the course of origin of different languages, the root from which they all stem and also genetic association between these languages. Seeking answers to similar questions and dealing with the same challenges as human geneticists, the historical linguists have adduced classification schemes for the languages from all over the world and grouped them into language families (Atkinson & Gray, 2005). Genetic classification of languages is based on the hypothesis of common origin and the term "genetic" is derived from "genesis" which means "the origin of something" (Ofori, 2014). For such classification, linguists have adopted a tree model similar to that of a family tree or a phylogenetic tree used by geneticists and evolutionary taxonomists.

The tree model represents the history of language families. A "language family" is a term used to describe a group of languages that are thought to be related as having descended through a common ancestor i.e., parental language or "proto-language" (Rowe & Levine, 2014). However, the linguistic ancestry is not as precise as the familial biological ancestry (List, Nelson-Sathi, Geisler & Martin, 2013), and most of the languages have short recorded history, therefore their ancestor is rarely known. Each descendant language is called "daughter language" and daughter languages within a language family are believed to be genetically (a biological analogy) or genealogically related (Rowe & Levine, 2014). They are represented by branches within the linguistic tree and are also referred to as genetically related sister languages. For instance, Spanish, French, Portuguese, Romanian, Italian, and Catalan are all derived from Latin and are regarded as daughter languages.

As claimed by Ethnologue, thus far, 7111 human languages have been identified throughout the world and this number is continuously in flux; as all of these languages are "living-languages" (meaning that they are currently in use as a primary source of communication among specific groups of people). These living human languages have been distributed into 142 different language families. Out of 142 languages, only six stand out as major language families with the largest number of native speakers, namely: Indo-European, Sino-Tibetan, Niger-Congo, Austronesian, Trans-New Guinea, and Afro-Asiatic. Additionally, 12 dead or extinct language families have also been identified, having no descendant language or native speakers left (Pariona, 2019). Moreover, there are a few languages which have not been classified because they were never sufficiently studied or perhaps, they only existed inside their individual speech communities.

This grouping of the languages into families has been established on the basis of historical linguistics (also known as comparative linguistics) research methods, suggesting the fact that members within a language family deriving from a common proto-language retain its features or at least reflexes of these features. August Schleicher, a 19th century linguist, who devised the language tree model in 1861, suggested the method of validating the genetic relationship among languages and reconstruction of their parental proto-language, which is called the historical comparative method. Proto-language, therefore, is a hypothetical language which is reconstructed. Proto-languages have been reconstructed for various language families. Some known proto-languages include: Proto-Indo-European, Proto-Algonquian, Proto-Dravidian, Proto-Athabaskan, and Proto-Oto-Manguean; ancestors of Indo-European, Southern Indian, Native American, Mesoamerican language families. Among these proto-languages, Proto-Indo-European is the most well reconstructed and

*CORRESPONDING AUTHOR | Ghayeth Ersheidat | ✉ gh.ersheidat@gmail.com

established, and is assumed to be the ancestor of 448 different languages belonging to Western Asia, India and Europe (Ethnologue Report for Indo-European, 2019). Most historical linguists shared a common belief that the eventual proof of genetic relatedness lies in the reconstruction of proto-languages (Hock & Joseph, 2009). Thus, in order to establish genetic relationship between languages and reconstruct their proto-language, they all compared the similarities between these languages. The similarities constituted mainly those in semantic, syntactic, morphological and phonological features.

This paper aims to review comparative techniques used by various linguists to determine a shared ancestry among languages and the evidence they provided for the genetic groupings.

VISUAL REPRESENTATION OF GENETIC/GENEALOGICAL RELATIONSHIP (TREE MODEL)

To understand a language tree model, let us consider five modern languages which we label as K, L, M, N, O. These languages are supposed to be genetically related if they adhere to a number of conditions (Campbell & Poser, 2008). These conditions that are necessary for genetic relations (discussed later in this article), are a series of correspondences which cannot be assigned to convergence (chance of sound-meaning similarities in two languages) or borrowing (a result of exchange of words between languages that are in close contact with each other). If we say that languages K, L, M, N, O are genetically related, it would entail that they are derived from a single ancestor, a proto-language. In this case, the proto-language is Proto-KLMNO as seen in Figure 1(a).

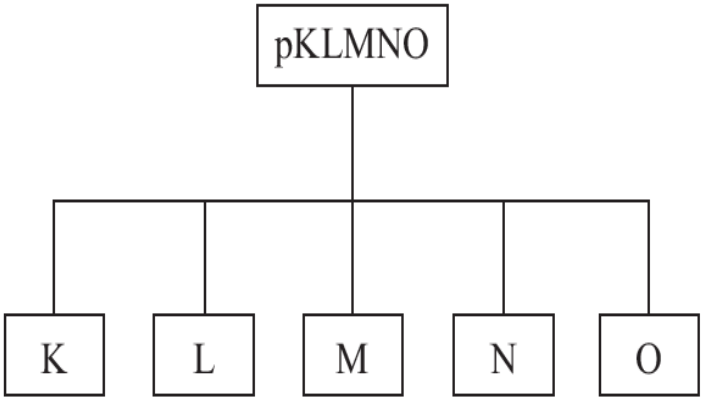

Figure 1(a). An unordered genealogical tree, where K, L, M, N, O are genetically or genealogically related, entailing that they descend from a single common ancestor (proto-language); in this case proto-language is Proto-KLMNO (François, 2014).

Figure 1(a) is a simple representation of an internal structure of a language family, when there is a lack of sufficient data. With more data, the genealogical tree would depict as in Figure 1(b).

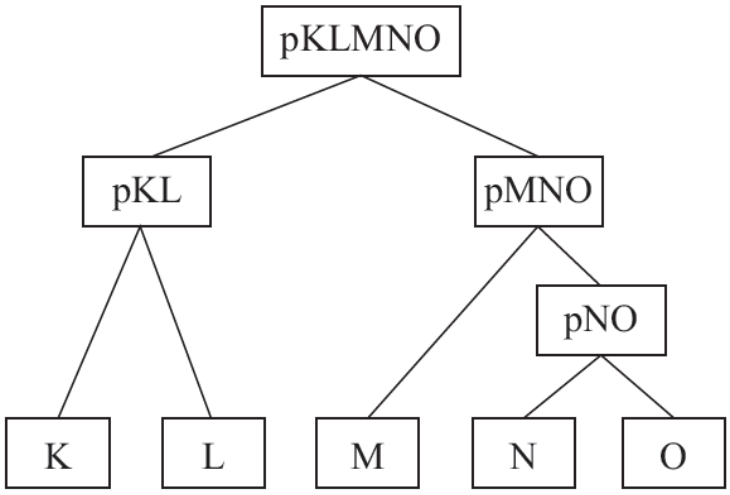

Figure 1(b). A genealogical tree representing internal subgroups. KL, MNO and NO are subgroups (François, 2014).

Figure 1(b) depicts the ideal scenario when there is enough information available to identify the existing sub-groups (KL, MNO, NO) within the language family. These subgroups contain languages which have more recent shared lineages. It is a common practice to interpret the cladistic renditions of language families in terms of history (François, 2014). Cladistic renditions entail the use of linguistic features to reconstruct the phylogeny of languages. Cladistic method is similar to the comparative method but it involves an explicit use of parsimony that allows much faster analysis of large datasets. The outcome of cladistic analysis is a cladogram (a tree shaped diagram). In the structure of the cladogram from top to bottom, the sequence of nodes, is assumed to mirror the chronological order of the historical events. Another conception is that every node in the tree diagram represents an individual language community thus, a split in a tree is equated with the distribution of an earlier unified language community into distinct social groups.

There are several criticisms of the language family tree model. The critics largely focus on the assertion that internal structure of the language tree model which changes with the criteria of classification (Edzard, 1998). Linguists are still debating which languages should be included in a certain language family (Gelderen, 2014). Before discussing the work of various linguists who determined a shared ancestry among different languages, it is necessary to understand the two important concepts: genetic relatedness and language similarity.

GENETIC RELATEDNESS IN CONTRAST TO LANGUAGE SIMILARITY

Genetic relatedness and language similarity are two separate concepts. Similar languages may not always be genetically related. Dissimilar languages, on the other hand, may be traceable to a common ancestor (Georgi et al., 2010). Genetically related languages display shared retentions of the ancestor language such as typological features, vocabulary and grammar which cannot be explained by borrowing, chance resemblance and sound symbolism. On the other hand, the languages with no common origin which appear to be similar present shared innovations. These innovations are acquired as a result of borrowing and other means (e.g., those means which are neither genetic nor to have bearing on the language family concept). Consider an example of three languages: English, Persian and Finnish, whereby, the first two are Indo-European languages and the last belongs to the Uralic language family. Although English and Persian are genetically associated, they exhibit very different typological features. While English and Finnish are not genetically linked, they are typologically very similar as shown in Table 1. This is due to the fact that English has become geographically distant from Persian and closer to Finnish.

Table 1. Comparison of some typological features of English, Persian and Finnish.

Feature Name	English	Persian	Finnish
Order of Verb, Subject, and Object	SVO	SOV	SVO
Order of Noun Phrase and Adposition (a cover term for prepositions and post positions)	Prepositions	Postpositions	Prepositions
Order of Noun and Adjective	Adjective-Noun	Noun-Adjective	Adjective-Noun

METHODS

Literature Review and Selection of Methodology

Google Web, Google Scholar, Microsoft Academic 2.0 and NCBI Databases were mainly used to acquire the data for this review paper. Different MeSH-terms and key words were used to retrieve the historical-linguistics based information and necessary research articles, for instance: “Genetic Linguistics”, “Linguistic Family-Tree’’, “Wave-Theory” and “Historical Comparative Analysis’’. Historical linguistics books that provide information regarding genetic relationship among various languages, and which were published over the last 6 to 7 decades have been consulted for the current review. Moreover, the majority of research papers that have been reviewed fall into the category of papers published between 2000 and 2019.

DISCUSSION

The purpose of this section is to review the work of various linguists who studied Indo-European languages and American languages. This section consists of a review of the comparative techniques and theories presented by the linguists for the genetic classification of the languages.

Linguists Who Worked on Indo-European Languages

According to historical linguistics, languages can be minimally related to each other or highly related or not related at all. Languages related through descent show similar features. In history, the similarities among languages were first described in the 18th century by Sir William Jones, who compared Sanskrit (an ancient Indian language) with Greek,

Gothic, Persian, Celtic and Latin, as well as provided evidence for their relatedness. He also provided the first formal evidence in 1786 that, Proto-Indo-European language was the ancestor of all of these languages (Poser & Campbell, 1992; Spadafora & Cannon, 1992). Jones observed that many words in these sister languages were having same meaning and were also phonemically identical. He named these sets of words as "cognates". Adducing that similarity between the cognates was due to their common origin, Jones thereby provided the main premises of 'The Relatedness Hypothesis'. He believed that the words with similar meanings would have similar sounds in all languages, if sound and meaning were just casually related to each other. Since that was not the case consequently, resemblance in sound and meaning should be the result of a common origin.

The contemporary Danish philologist and linguist, Rasmus Rask worked on Jones's conclusion and made contributions to comparative linguistics including laying the foundation of what was later known as Grimm's Law (Winge, 2009). Rask was the first person to describe symmetries in sound differences in some specific languages and today he is known as one of the main discoverers of Indo-European sound laws and the founding father of several linguistic disciplines (Hufnagel, 2016). For instance, he found certain sound correspondence between Greek and Germanic languages, e.g., the Greek sound *ph*, such as in *phrater* (in English *brother*), consistently changes to *b* in German *brüder*.

In 1816, Franz Bopp elucidated the conjugational system of the Sanskrit in comparison with the conjugational systems of Persian, Greek, Latin and Germanic (Bopp & Windischmann, 2010), and wrote a Comparative Grammar. After Bopp's work, Indo-European studies attained the status of academic discipline, leading to August Schleicher's *Compendium* (1861).

Historical Comparative Method

The historical comparative method involves observing similarities in languages to rule out the degree of relationship among these languages and to reconstruct the ancestors (proto-languages). The historical comparative method can be summarised as a following set of instructions:

1. Establish a genetic family (i.e., a group of languages are genetically relevant) based on the strength of diagnostic evidence.
2. Gather presumed cognate-sets for the family including both lexical features and morphological paradigms.
3. Contrive the sound resemblances from the cognate-sets, ignoring the irregular cognate-sets.
4. Then use the following strategies to reconstruct the proto-language of the family.
   a. Use the sound resemblances established in Step 3 and reconstruct the proto-phonology, by practicing conventional wisdom with regards to the directions of sound variations.
   b. Use the cognate-set gathered in Step 2, to reconstruct proto-morphemes (both lexical features and morphological paradigms), utilising the proto-phonology that is reconstructed in Step 4a (Ross & Durie, 1996), additional steps include subgrouping, generating the family tree and finally building an etymological dictionary (See Figure 2).

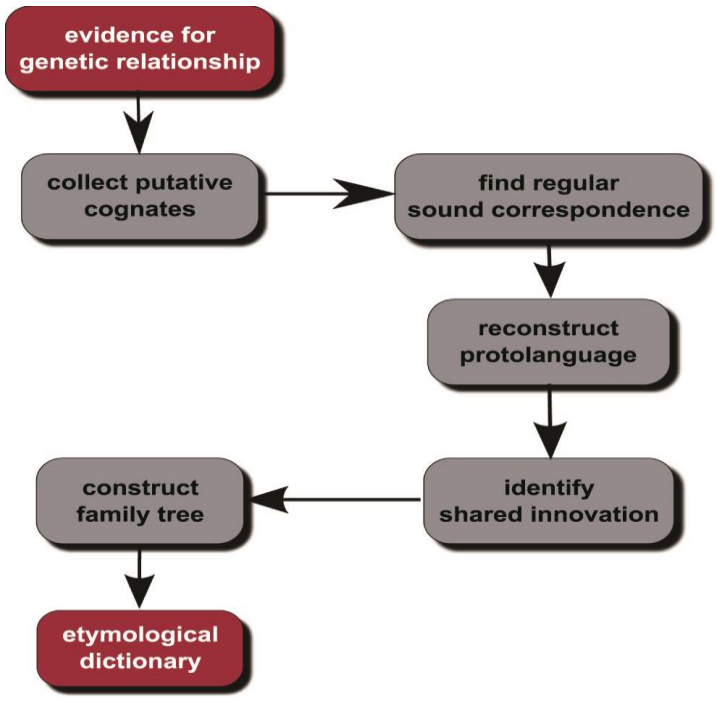

Figure 2. Graphical depiction of Historical Comparative Method (Jäger, 2019).

More recently, some discussions seemed to assume that the historical comparative method is a technique for scientifically verifying that languages are linked by developing phonological similarities between them. Although sound laws are undoubtedly a significant tool in the historical linguist's kit, but the comparative method cannot be considered as synonymous with the application of sound laws.

In 1882, German linguist and philologist Jakob Grimm followed up Rask's suggestion of symmetry in sound differences. Taking it one step further, he made an explanation which systematically satisfied the correspondences between particular consonants in the Germanic languages and those found in Greek, Latin and Sanskrit. He concluded his work in the form of four volumes of *Deutsche Grammatik* (Germanic Grammar), which included what is called Grimm's law or the first Germanic sound – the first formulated explanation of sound shift from a parent language to a daughter language. This law was actually discovered in 1806 by Friedrich von Schlege, later on by Rask then Grimm extended it to add standard German. The law indicates a shift from proto-Indo-European to the daughter Germanic languages in the form of, a set of correspondences between Germanic fricatives and stops, and the stop consonants of Latin, Greek and other Indo-European languages. It deals with the alterations in three natural classes of sounds and can be summarised in three steps, which must be considered as three successive phases in a chain shift.

a) Proto-Indo-European voiced [+asp] stop [-continuant] consonants changed into [-asp] plain voiced consonants; where [+asp] and [-asp] indicate presence and absence of aspiration; the sounds [bh], [dh], and [gh] or *voiced aspirated stops* belonging to one of the three natural classes of sounds, became unaspirated.
b) Proto-Indo-European voiced stop [-continuant] consonants such as [b], [d], and [g] became voiceless [-continuant] consonants like stops [p], [t], and [k].
c) Proto-Indo-European voiceless stop [-continuant] consonants [p], [t], and [k] in turn became [+continuant] or fricatives.

Table 2. The three phases of Grimm's Law, a shift from Proto-Indo-European to Germanic languages. PIE = Proto-Indo-European and Gmc = Germanic.

PIE → Gmc Phase I	PIE → Gmc Phase II	PIE → Gmc Phase III
$b^h \rightarrow b$	b → p	p → f
$d^h \rightarrow d$	d → t	t → þ [θ]
$g^h \rightarrow g$	g → k	k → x/h

In the formula form, Grimm's law can be written as follows:

C >[-asp] C >[-voice] C >[+cont]

[+voice] [+voice] [-voice]

[-cont] [-cont] [-cont]

[+asp]

Since these shifts did not occur in any other Indo-European languages, they helped in defining the Germanic languages. Grimm then found another sound shift (known as second sound shift) relating merely to a form of German known as High German. Grimm also introduced a precise methodology for comparative studies which tremendously influenced the evolution of historical linguistics. In 1875, Danish linguist Karl Verner expanded on Grimm's law, which is referred to as Verner's Law. Verner's Law states that the placement of stress affected Indo-European consonant shift. This would explain why, voiceless inter-vocal stops and fricatives f, þ [θ], x/h turned into voiced fricatives β, ð, ɣ respectively, when there is no original Indo-European stress on the immediately preceding vowel within the same word (reversal of Grimm's law in few cases). Verner's law can also explain the development of the word *father*. Proto-Indo-European *t* developed into *ð* rather than the expected *θ*, Proto-Germanic *faðēr* rather than expected *faθēr* (Glottopedia, 2019).

German linguist Karl Brugmann (1849-1919), evaluated the field of historical linguistics and provided description of phonetic laws and their operation, morphology and word formation in the five volumes of *Grundriss der vergleichenden Grammatik der indogermanischen* (Outline of the Comparative Grammar of the Indo-Germanic Languages), published from late 1880s to early 1890s. Ferdinand de Saussure (1857-1913), a Swiss linguist and semiotician, developed a laryngeal theory. His ideas laid the foundation of modern Indo-European studies. The modern Indo-European scholars namely: Jochem Schindler, Helmut Rix, and Calvert Watkins, have contributed in establishing better understanding of morphological aspect of linguistics and its role in determining genetic connections among languages in the last few decades of the 20th century.

Macrofamilies

A *macrofamily* or *superfamily* is a group of two or more proto-languages. Various linguists have proposed different macrofamilies for example, in 1903, a Danish linguist, Hogler Pedersen suggested a macrofamily called Nostratic, by grouping the Afro-Asiatic, Indo-European, Eskimo-Aleut and Uralic language families together. Language families which are often grouped to form macrofamilies are those which cannot be substantiated as phylogenetic units by the valid historical linguistic methods. The biggest problem with the idea of macrofamilies or super-families is the length of time which has been passed since their hypothetical existence. Another example of macrofamily is Indo-Uralic, which is a controversial hypothetical macrofamily which consists of Indo-European and Uralic languages. Indo-Uralic hypothesis (i.e. a hypothesis suggesting a genetic relationship between Indo-European and Uralic languages) is mainly derived from the early publications by Björn Collinder (1934, 1945, 1954). Most of his suggestions were later considered obsolete even by himself (Klein et al., 2018).

The Wave Model

In order to report few of the inadequacies of the linguistic tree model, Johannes Schmidt (1843-1901), a German linguist developed the wave model or *Wellentheorie* (Wave theory) of language relatedness in 1872. In this model, he drew circles around languages which shared one or more particular characteristics; each language within a circle shared the characteristic represented by the circle. For instance, consider a wave model for a sample of Indo-European languages (see Figure 3). The wave model shows linguistic relationships more precisely than the tree model. The circles in the figure describe linguistic features (morphological, syntactic or phonological) regarded as common for the languages placed inside them, which is also representative of the idea that linguistic features diffuse. Languages which lie close to each other are related to each other while those which are not in close vicinity can influence each other through phenomena such as warfare and trade. The circles suggest that, languages are not unified systems but have numerous variations within them. As the new similarities are found among languages, more and more new circles are added to the figure. The language groups within circles, unlike those in the language tree model, can overlap. The tree model does not imply the contact between the languages once derived from their ancestor whereas, the wave model implies relationships between languages which remain in contact (François, 2014).

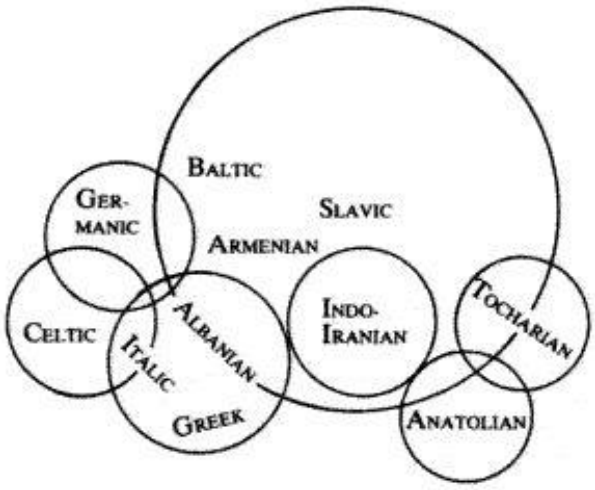

Figure 3. The Wave Model of genetic relatedness among languages: A sample of Indo-European languages (Renfrew, 1989, 1990).

This model, just like the family tree model, fails to depict that the languages which are not genetically related but still appear to be similar due to several reasons such as cultural contact, chance similarities and language universals. Although both the tree and wave model have faults, they have been very useful particularly when used simultaneously to find genetic relationship between languages and track linguistic evolution. However, the connections among languages have much more complexity than either of these models can depict, separately or together. Therefore, more complex models have been developed for example, the punctuated equilibrium model which is inspired by the biological evolutionary model.

Proto-Human Languages

In 1905, Alfredo Trombetti tried to prove genetic relatedness of all languages in the world and his first scientific attempt to determine the reality of monogenesis can be found in his book *L'unità d'origine del linguaggio* (Ruhlen, 1994). Monogenesis is a concept in linguistics that all human languages have descended from a single common ancestor called Proto-Human language or Proto-World. Trombetti conjectured that the Proto-Human language had been spoken between 100,000 and 200,000 years ago, presumably in the Middle Paleolithic period or close to the primary emergence of human

beings (Trombetti, 1923). This concept was however deemed purely hypothetical and was dismissed by several linguists during late 19th and early 20th century, when another concept, polygenesis, which was contrasted with monogenesis became widely popular. Polygenesis is the theory that human languages have evolved as various lineages without any influence from one another (Saussure & Harris, 2016).

In mid-20th century, an American linguist, Morris Swadesh (1909-1967) supported the monogenesis doctrine (Ruhlen, 1994). He classified native American languages, and he was also the pioneer of the two principal methods for investigating extensive relationships among languages, lexicostatistics and glottochronology. Lexicostatistics is a technique in comparative linguistics which involves comparison of the percentage of lexical cognates among different languages to find out their relationship (or quantitative assessment of the genetic relationship of the languages), while the use of lexicostatistics for the dating of language branching is referred to as glottochronology. Swadesh compiled a list of 207 basic concepts which occur in all languages, for the use in comparative linguistics as shown in Table 3. This lexicostatistical list serves in lexicostatistics for defining sub-groupings of languages while, in glottochronology it helps to establish dates for different branching points in the language tree (Embleton, 1992).

Table 3. Some of the terms from Swadesh's final list (1971).

No.	Term (English)	No.	Term (English)	No.	Term (English)
1	I	8	Not	15	Small
2	Thou	9	All	16	Woman
3	We	10	Many	17	Man
4	This	11	One	18	Person
5	That	12	Two	19	Fish
6	Who?	13	Big	20	Bird
7	What?	14	Long	21	Dog

Linguistis Who Worked on North American Indian Languages

Edward Sapir (1884-1939), an American linguistic-anthropologist proposed and codified the distant genetic connections among North American Indian languages which had a very profound impact. He presented this linguistic classification in an article in Encyclopedia Britannica in 1929, which still has the greatest importance among the articles that claim to establish distant relationships between American-Indian languages. Sapir used basically the same comparative method that was used by Indo-European linguists and other historical linguists of that time. His method involved reconstructing linguistic history by comparing particular morphological features and structures of languages which are thought to be genetically related. All the principles of this comparative method were described in Antoine Meillet's La méthode comparative (1925).

Sapir's comparative method was based on an axiomatic assumption that few resemblances among languages are of such a type that they can only be explained by the hypothesis of genetic inheritance from a single common original. He suggested to classify languages mainly by resemblances of morphology, and in this way, he was also able to classify those languages which had no lexical resemblances at all. When there was a conflict between lexical evidence and the morphological evidence of genetic relationship, he explained it as the consequence of lexical borrowing. Sapir found out that lexical borrowing was frequent, often taking place on a big scale whereas morphological borrowing also happened but as a rare phenomenon, occurring only in the presence of extreme lexical borrowing. Regardless of the importance of his work, it has never been critically discussed in a complete manner. However, certain aspects of it have gained repeated attention (Cowan et al., 1986).

Franz Uri Boas, a German-American anthropologist, also known as Father of American Anthropology, suggested that no simple genetic classification of languages was possible, taking into account the fact that large-scale diffusion occurs constantly in every aspect of language and any given language could have multiple roots. Boas assumed that genetic relatedness would be easily recognisable from resemblances in all aspects (lexical, morphological, semantic) of the languages compared. However, his proposal did not receive significant amount of favor from other linguists, especially from those who were familiar with the accomplishments of Indo-European comparative philology.

Alfred Louis Kroeber (1876-1960), an American cultural anthropologist who received his PhD under Franz Uri Boas, on the other hand completely ignored the conflicting morphological evidence and classified languages solely on

the basis of word comparisons (1913). Kroeber's suggestion is more reasonable among others and could be employed to obtain a careful preliminary classification.

In order to systemically reconstruct the historical development of languages, it is necessary to establish valid hypotheses about genetic relationships among them by comparing the languages. Now the question which arises here is, what aspects of language are germane for comparison: morphemes (meaningful morphological units), meaning and order of morpheme classes, phonemes (contained within morphemes, perceptually different units of sound), morphemes that themselves are roots, assigned to the non-roots (affixes) or lexicon with inflectional grammatical function?

Conditions for Proving the Genetic Relationship of the Languages

For establishing an exact proof of genetic relationship among languages, one needs to satisfy a series of conditions. For instance, consider the following quantitative conditions used by American linguists for proving genetic relationships.

1. Sufficient number of comparable roots. Fifty identical roots are hardly sufficient. Languages for which genetic relationships have been proved generally exhibit at least 400 identical roots. Roots must be comparable in respect of semantics; they must have identical meanings. In short, the roots must satisfy all of the requirements for comparison in both quantitative and qualitative respects.
2. Series of affix (grammatical morpheme) correspondences.
3. Structural similarities; same structural system (word order, syntax).
4. Series of phoneme correspondences.
5. Basic word correspondences, general human terms such as parts of human body, numerals, etc. Basic words can be separated into three further categories namely,
    a) Essential Basic Words (e.g., eye)
    b) Marginal Basic Words (e.g., eye-lash)
    c) Intermediary Basic words.

Out of these three categories, only essential basic words can be taken as a source for proving genetic relationship; as the other two categories are very unstable and are easily loaned while the marginal basic words are stable and go back to a common ancestor proto-form (Doerfer, 1981).

Greenberg's Methods

When doing comparison, the resemblance of sound and meaning in roots (morphemes) is referred to as lexical resemblance, and the similarity of sound and meaning in non-roots is considered as grammatical resemblance. Sound-meaning resemblances are most significant in determining historic or genetic relationships among languages; but the fact which inevitably becomes prominent at the outset is that, all of these similarities stem from historic roots. The relationship between sound and meaning is arbitrary (Greenberg, 1972). The causes of sound-meaning similarities between two languages can therefore, be of various types: chance (convergence), symbolism (sound-symbolism), genetic relationship (common origin), and borrowing. Where a and b are non-historic causes while c and d are historic causes. After knowing the causes for similarities, the historical linguists' task is to eliminate chance and symbolism which would lead to the hypotheses of the historic relationships; and segregate borrowing from the instances on which genetic relationship would be hypothesised. Presence of a significant number of sound-meaning similarities or resemblance of twenty percent or more is considered due to historic factors i.e., borrowing and/or genetic relationship.

When the languages show similarities in fundamental vocabulary and grammatical items, it is then a sure indication that they are genetically related. Borrowing on the other hand, results in mass resemblances which appear in cultural vocabulary or in semantic areas reflecting cultural nature of contact i.e., pointing towards 1, 2 or 3 languages as donor. It leads to the comparison of closely related languages to generate language groups and comparison of these groups with similarly generated groups.

In the middle of the 20th century, when scholars were making attempts to classify African languages on the basis of racial and typological traits, Joseph H. Greenberg, one of the most important linguistic-anthropologists of his time gave an early version of African and American language classes. He compiled comparative core vocabularies of all the languages existing in an extended region and examined as many languages as possible of the particular area. Rather than comparing just two languages, Greenberg performed mass comparison, because he believed that statistical reliability of lexical resemblances (or series of lexical cognates which determine genetic relationship) improves as the number of data points are raised. He rejected the idea prevalent among linguists of his time that, historical comparative reconstruction was the only way of finding out genetic relationships among languages. He suggested that genetic classification is preliminary to comparative reconstruction as reconstruction is not possible without knowing which languages to be compared.

In 1966, he classified hundreds of African languages into just four families (Elders, 2003). They are: Afroasiatic, Niger-Kongo, Nilosaharian, and Khoisan. In 1971, Greenberg gave the Indo-Pacific hypothesis, a proposal of Indo-Pacific macrofamily comprising of Papuan languages, Andamanian languages and Tasmanian languages. In 1987, Greenberg suggested three macrofamilies in Americas: Eskimo-Aleut, Na-Dené and Amerind in 2000-2001. He also proposed a macrofamily called the Eurasiatic macrofamily, comprising of subfamilies Indo-European, Uralic, Altaic (Mongolian, Korean, Turkic, Tungusian and, Japanese), Eskimo-Aleut languages, and several isolated languages (for example, Etruskian).

Genetics and Linguistics Affinity

After Greenberg's death in 2001, his colleague which is also his student, Merritt Ruhlen, suggested more radical application of his method of genetic classification. Ruhlen proposed that partial reconstruction of proto-world is possible. Both Greenberg and Ruhlen had to take controversies and criticism from several linguists. Most of the criticism that Joseph Greenberg faced centers on his technique of language classification, mass comparison. Similar criticism was directed at Ruhlen because he defended Joseph Greenberg's mass comparison technique. As stated previously, mass comparison is the comparison of selected elements of the morphology and basic vocabulary of the languages being investigated. These selected elements are examined for similarities in sound and meaning, and on this basis of which a hypothesis of genetic classification of languages is formulated. Greenberg and Ruhlen suggested that such classification is the first step in the historical comparative method and the reconstruction of a proto-language. They argued that the reconstruction of a proto-language can only be carried out after a hypothesis regarding genetic classification of larger groups is formulated whereas other linguists claim that only the application of the comparative method can prove a genetic relationship. The comparative method should be applied to lower-level groups first, and then it should be applied to progressively higher-level (larger) groups. Greenberg is thus criticised for not following steps 1 to 4, particularly steps 3 and 4, in the historical comparative method. The majority of historical linguists consider the successful accomplishment of steps 1 to 4 is essential for the proof of genetic relationship between languages.

Ruhlen along with other scholars namely Robert Sokal and Jiangtian Chen added genetic perspective to a linguistic one and made contribution to a cross-disciplinary pre-history of mankind. For the inference and reconstruction of proto-world, they investigated potential correlation between genetic and linguistic lineages in the largest possible region worldwide, also at the longest possible time depth. They referred to the biological family tree of modern mankind developed in 1988 by Luigi Luca Cavalli-Sforza (a colleague of Greenberg at Stanford) with the help of genetic analysis. Figure 4 represents the Luigi Luca Cavalli-Sforza diagram. In Figure 4, the left side represents inter-population genetic distances of forty-two world populations while the right side shows languages of these populations organised in the form of sixteen high scale phyla in turn converging into a single node.

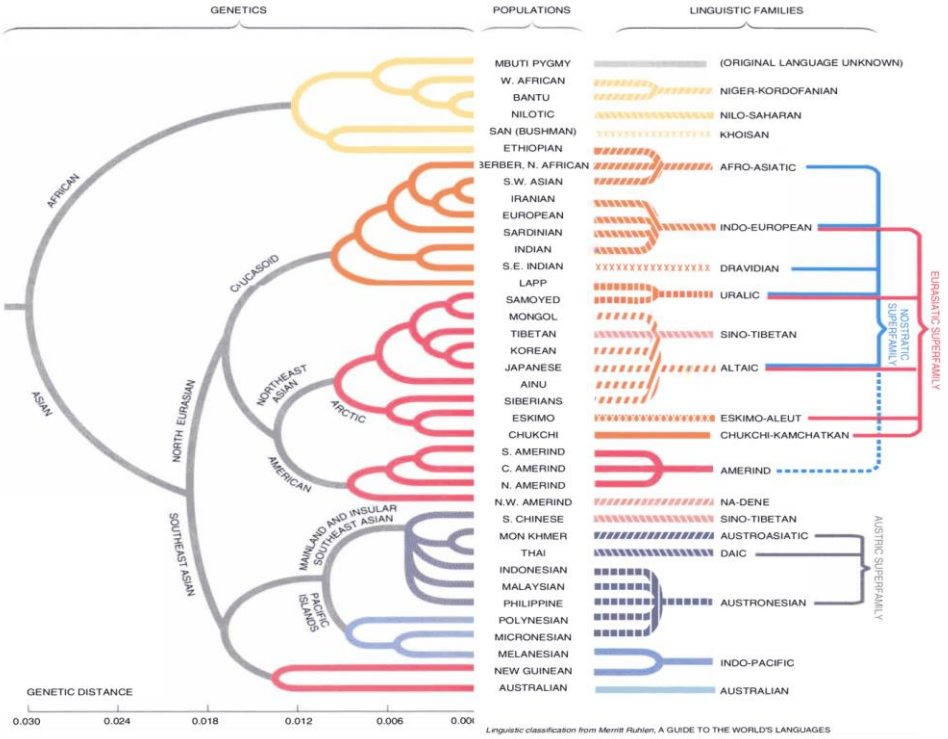

Figure 4. Cavalli-Sforza diagram showing correlation between languages (right side) and genetics (left side) (Cavalli-Sforza, 1991).

Many scholars believed that human linguistic and genetic diversifications go in lockstep (Chen et al., 2012). However, later on with compilation of the data on genetic diversity, it was discovered that these two lineages diverge at different rates. However, a few correlations can be found between genes and languages (Colonna et al., 2010). Africa is considered to be the most genetically diverse region in the world, and genetic diversity decreases with distance from Africa. On the other hand, linguistic diversity is low in Africa and Europe while high in Americas and Oceania. This difference of patterns is probably due to the reason that languages are fast to mutate and slow to diffuse; in comparison to genes, which mutate slowly and are very fast to diffuse (Nettle, 2008).

With the advancements in the field of bioinformatics, the problems in the reconstruction of linguistic trees, similar to those problems in evolutionary biology, have been greatly resolved. One such example is the application of Bayesian methods to lexical and phonetic data that has generated dated linguistic phylogenies for eighteen language families encompassing approximately 3,000 languages (Hamilton & Walker, 2019). Nowadays, bioinformatics statistical techniques for inducing genetic relationships are being increasingly applied to the available linguistic data. These tools have helped to recover evolutionary history and the history of human languages (Jäger, 2015; Jäger et al., 2017). A huge number of collections of the comparative linguistic data have become available in digital form, giving the historical linguistics field another boost (Jäger, 2019). This linguistic data is readily available through several linguist databases such as *WALS* or *World Atlas of Language Structures* (Haspelmath et al. 2008), *Ethnologue* (Lewis et al. 2016), *Glottolog* (Hammarström et al. 2016). These databases are catalogues of linguistic features, particularly WALS database which catalogs linguistic features for over 2,556 languages in 208 language families, using 142 features in 11 categories (Georgi et al., 2010). Some of the computational statistical techniques used for the analysis of linguistic data include parsing methods, clustering methods, syntactic projection methods and morphological induction techniques. These techniques are used to establish genetic relationship between languages that are assumed to have similar morphological, syntactic, semantic or phonological features (Jäger & Sofroniev, 2016). By applying computational statistical techniques, linguists have brought significant advances in broad-coverage genetic classification of languages. The isolated efforts of historical and computational linguists, and bioinformaticians have provided a major impetus to the emerging field of '*Computational Historical Linguistics'* (Jäger, 2019).

CONCLUSION

Languages that are known to have common ancestral origin are said to be genetically related and the reconstruction of the ancestral language is referred to as a proto-language. Languages that share common ancestor are grouped into a language family. Various language-families have been identified and genetic classifications have been proposed for languages from all parts of the world by various linguists, who have classified them into 142 families. A language family tree diagram or a wave diagram is used to display the relationships among languages. Although both the family tree and wave models have faults, they have been very useful particularly when used simultaneously to find genetic relationship between languages and to track linguistic evolution. Two complementary techniques for classifying languages genetically, or reconstructing a proto-language are the historical comparative method and internal reconstruction. As the current review has demonstrated, most historical linguists believe that the eventual proof of genetic relationship lies in the reconstruction. Thus, in order to establish the genetic relationship between languages and reconstruct their proto-language, they all performed direct comparison on the similarities between the languages. These similarities constituted mainly semantic, syntactic, morphological and phonological features. Similarities in vocabulary proposed one classification, whereas similarities in morphology proposed another. Although all of these theories are controversial in terms of their validity and reliability, with the progress in the field of bioinformatics, the problems in linguistic reconstruction have been greatly resolved. Nowadays, bioinformatics statistical techniques are being increasingly applied to the available linguistic data for inferring genetic relationships among languages. For future reviews in the field of historical linguistics, we suggest that scholars should consider more recent research based on modern bioinformatics and statistical techniques to assess the broad range of linguistic databases. It will enable explicit understanding of genetic classifications of languages as well as the typical pattern of language diversification.

REFERENCES

Atkinson, Q. & Gray, R. (2005). Curious Parallels and Curious Connections—Phylogenetic Thinking in Biology and Historical Linguistics. *Systematic Biology*, *54*(4), 513-526.

Bopp, F., & Windischmann, K., J., H. (2010). *Über Das Conjugations system Der Sanskritsprache*. Olms.

Campbell, L., & Poser, W.J. (2008). *Language classification: history and method*. Cambridge University Press.

Cavalli-Sforza, L., L. (1991). Genes, Peoples and Languages. *Scientific American*, *265* (5), 104-110.

Chen, J., Sokal, R., & Ruhlen, M. (2012). Worldwide Analysis of Genetic and Linguistic Relationships of Human Populations. *Human Biology*, *84*(5), 555-572.

Colonna, V., Boattini, A., Guardiano, C., Dall'Ara, I., Pettener, D., Longobardi, G., & Barbujani, G. (2010). Long-Range Comparison between Genes and Languages Based on Syntactic Distances. *Human Heredity*, *70*(4), 245-254.

Cowan, W., Foster, M., & Koerner, E. (1986). *New Perspectives in Language, Culture, and Personality.* John Benjamins Pub. Co.

Doerfer, G. (1981). *The Conditions for Proving the Genetic Relationship of Languages.* Kyoto Sangyo University.

Edzard, L. (1998). *Polygenesis, Convergence, and Entropy*. Harrassowitz.

Elders, S. (2003). *African Languages. An Introduction: Bernd Heine and Derek Nurse*. Cambridge University Press.

Embleton, S. (1992). *The Computer Developed Linguistic Atlas of England*. Oxford University Press.

Etnologue Report for Indo-European. (2019). *Ethnologue.* Retrieved 03 October 2019 from https://www.ethnologue.com/subgroups/indo-european.

Francois, A. (2014). Trees, Waves and Linkages: Models of Language Diversification. In C. Bowern & B. Evans (Eds.) *The Routledge Handbook of Historical Linguistics*, pp. 161-189, Routledge.

Gelderen, E. (2014). *A history of the English language*. John Benjamins Pub.

Georgi, R., Xia, F., & Lewis, W. (2010). *Comparing language similarity across genetic and typologically-based groupings.* Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 385-393.

Greenberg, J., H. (1972). *Essays in Linguistics*. The University of Chicago Press.

Hamilton, M., & Walker, R. (2019). Nonlinear diversification rates of linguistic phylogenies over the Holocene. *PLOS ONE, 14*(7), e0213126.

Hammarström, H., R. Forkel, M. Haspelmath & S. Bank. (2016). *Glottolog 2.7*. Max Planck Institute for the Science of Human History, Jena. Retrieved 04 March 2020, from http://glottolog.org

Haspelmath, M., M. S. Dryer, D. Gil & B. Comrie. (2008). *The World Atlas of Language Structures online*. Max Planck Digital Library. http://wals.info/.

Hock, H. & Joseph, B. (2009). *Language History, Language Change, and Language Relationship.* Mouton de Gruyter.

Hufnagel, S. (2016). *Document Sans Titre*. Tabularia. https://doi.org/10.4000/tabularia.2666.

Jäger, G. (2015). Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of The National Academy of Sciences, 112*(41),12752-12757.

Jäger, G., & Sofroniev, P. (2016). Automatic Cognate Classification with a Support Vector Machine. In Dipper S., Neubarth F. & Zinsmeister H. eds. *Proceedings of the 13th Conference on Natural Language Processing*, vol. 16 of Bochumer Linguistische Arbeitsberichte 128–134 Ruhr Universität Bochum.

Jäger, G., J.-M. List & P. Sofroniev. (2017). Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics.* ACL.

Jäger, G. (2019). Computational historical linguistics. *Theoretical Linguistics, 45*(3-4), 151-182.

Klein, J., Joseph, B., & Fritz, M. (2018). *Handbook of Comparative and Historical Indo-European Linguistics*. De Gruyter Mouton.

Lewis, M. P., G. F. Simons & C. D. Fennig (eds.). (2016). *Ethnologue: Languages of the world*, (9th ed). Dallas, SIL International.

List, J., Nelson-Sathi, S., Geisler, H., & Martin, W. (2013). Networks of Lexical Borrowing and Lateral Gene Transfer in Language and Genome Evolution. *Bioessays, 36*(2), 141-150.

Nettle, D. (2008). *Genetic and linguistic diversity: Global distribution and implications for prehistory*. Retrieved 5 October 2019, from https://pdfs.semanticscholar.org/8e84/ab04bac0a68499647a4d7ef2cdaf73cdde27.pdf

Ofori, S. (2014). *Genetic Classification of Languages. Presentation*, Department of Linguistics, University of Ghana.

Pariona, A. (2019). Language Families of the World. [online] *WorldAtlas*. Retrieved 29 September 2019, from https://www.worldatlas.com/articles/language-families-with-the-highest-number-of-speakers.html

Poser, W., & Campbell, L. (1992). Indo-European Practice and Historical Methodology. *Annual Meeting of The Berkeley Linguistics Society, 18*(1), 214.

Renfrew, C. (1989). The Origins of Indo-European Languages. *Scientific American, 261*(4), 106-114.

Renfrew, C. (1990). *Archaeology and Language*. Cambridge University Press.

Ross, M., & Durie, M. (1996). *The Comparative Method Reviewed.* Oxford University Press, USA.

Rowe, B. & Levine, D. (2014). *A Concise Introduction to Linguistics.* Routledge.

Ruhlen, M. (1994). *On the Origin of Languages.* Stanford University Press.

Saussure, F., & Harris, R. (2016). *Course in General Linguistics*. Bloomsbury.

Spadafora, D, & Cannon, G. (1992). The Life and Mind of Oriental Jones: Sir William Jones, The Father of Modern Linguistics. *The American Historical Review, 97*(5), 1522.

Trombetti, A. (1923). *Elementi Di Glottologia.* N. Zanichelli.

Verner's Law - Glottopedia. (2019). *Glottopedia.Org.* Retrieved 06 October 2019, from http://www.glottopedia.org/index.php/Verner%27s_law

Winge, V. (2009). Rasmus Rask | Gyldendal - Den Store Danske. Retrieved 22 January 2020, from http://denstoredanske.dk/Sprog,_religion_og_filosofi/Sprog/Sprogforskeres_biografier/Rasmus_Kristian_Rask