Dean&Francis
ISSN 2959-6157

Quantifying Linguistic Divergence: Methodological Evolutions and Case Studies in Lexicostatistics

Yiyang He*

Beijing National Day School, Beijing, China
*Corresponding author.
Email: heyiyang.danny@gmail.com

Abstract:

Lexicostatistics acts as a strong interdisciplinary tool, bringing together the application of statistics and mathematics to language evolution. This paper critically assesses the traditional Swadesh method, highlighting its major steps in the lexicostatistical procedure. It is apparent that, although each step in the traditional method is subject to methodological pitfalls, which greatly impede its accuracy, recent improvements in the method have greatly enhanced its accuracy in glottochronology. From the assessment of the application of lexicostatistics in history, it is apparent that most studies focus on Indo-European languages. This is in line with most linguistic theories, although some aspects in the studies are unique and divergent from conventional theories. This paper, therefore, concludes that lexicostatistics is an emerging field with immense potential for growth, especially through the application of modern technologies like computer intelligence. By extending its scope to other languages, lexicostatistics is likely to give greater and accurate insights into language evolution. This work should be a good guide in the area of Lexicostatistics.

Keywords: Lexicostatistics, Historical linguistics, Statistics, Central limit theorem, Swadesh list.

1. Introduction

Lexicostatistics or glottochronology is a technique that integrates statistical methods into linguistics which attempts to provide dates for language separation, similar to the carbon-14 dating technique used in archaeology [1]. Previous linguistic methods to reconstruct to some extent the history of language have been unable to provide dates apart from written historical records, which this method aims to solve [2]. It promises a measure of time depth for language families without documented history, and thus has gained popularity recently.

The method is first developed by Morris Swadesh in the 50s, and numerous techniques have been proposed to carry out the analysis. The earliest method is restricted to chronology, i.e. date the chronology of their split from the proto-language, using statistical

methods. The method relies on comparing two languages using a given word list. By investigating the correspondence (cognate pairs) between the two languages, one can determine the time of separation between the two languages [1]. However, the computational approaches to historical linguistics stretch beyond chronology, and since then, statistical methods to pertain language classification, cognate detection, comparative reconstruction, simulation of phonological and analogical change have been developed and applied [2]. Although the use of quantitative methods is not new in contemporary linguistics, the approach is relatively novel and mostly regarded as problematic by traditional historical linguists, who consider the method to be unreliable and controversial. Many find the assumptions in the method too idealistic and thus can cause large errors. Nevertheless, the author has seen that the computational methods can be useful in linguistics, due to their systematic approach and their ability to quickly analyse masses of data with little errors, and many techniques for addressing the criticisms raised by traditional linguists have been proposed [2].

In this paper, the authors present the statistical methods of glottochronology with a particular focus on the mathematical components. In addition, the author discusses some recent studies on each step of the method and present some techniques developed to address some criticisms. Last but not least, this paper also curated some applications of glottochronology in different language families and their findings. The authors believe that lexicostatistics is a subject in linguistics that has a lot of potential, especially in the AI era.

2. Methodologies of Glottochronology

2.1 Basics of Glottochronology

The general methodology of glottochronology is established by Morris Swadesh in the 50s, where he took inspiration from the carbon-14 dating technique used in geology. [1] First, a list of universal concepts is selected and their corresponding words in different languages are considered. The technique is to consider the replacement of words as language evolves. Given the M words corresponding to the M concepts in the list, Swadesh hypothesized that the number of unreplaced words after time T can be modeled as

$$M(T) \approx Me^{-\lambda T} \tag{1}$$

where λ is the replacement rate of a word.

Therefore, if one has determined that one language is the ancestor language of another language, equation 1 can then be used to determine the estimated separation of the two languages. Assume that a list of M concepts are chosen and $\mathcal{M}$ words are found to be cognates (equal words that differ by spelling or pronunciation due to geographical, political and historical factors), then the estimated time of separation is given by

$$T \approx -\frac{1}{\lambda} ln\left(\frac{\mathcal{M}}{M}\right) \tag{2}$$

More commonly, the vocabulary of the ancestor language is not given, and this goal is to determine the separation time of two contemporary languages. In this case, the author can fix the vocabulary of one language and think of the words of the other language as being replaced at rate 2λ. Therefore, equation 1 can be rewritten as

$$\Omega(T) \approx Me^{-2\lambda T} \tag{3}$$

where Ω is the number of cognate pairs after time T.

Similarly, if people are given a list of M concepts in two contemporary languages and observed Ω cognate pairs, one can estimate the time of separation to be

$$T \approx -\frac{1}{2\lambda} ln(\omega) \tag{4}$$

where $\omega = \frac{\Omega}{M}$ is the cognate overlap, ranging from 0 to 1 [3].

2.2 Swadesh List and Vocabulary Lists

The list of concepts chosen can greatly influence the accuracy of the lexicostatistical analysis. Swadesh himself first proposed a list with 200 terms in 1952 [4]. However, he reduced the list to 100 words in 1955 based on universality, interlingual ambiguity and prevented words that are non-cultural, potential duplications, sound imitations, form words or those that share identical roots [1]. Nowadays, the most commonly used list is still the Swadesh list adapted from 1952 [4], containing 207 words.

Recently, other shorter lists of concepts are also proposed, such as the 35-word Swadesh-Yakhontov list published in 1991 [4]. This raises the question of the most effective number of concepts in the list, meaning, for what number of vocabularies will the lexicostatistical analysis be the most accurate. In a 2016 paper, Menghan Zhang and Tao Gong investigated this question using statistical approaches [4].

Consider a language with the total vocabulary set V, which contains N words, and the task is to construct a subset X of V with n words, each chosen from V. The author then wants to determine n such that the distribution of sound correspondences in X approximately matches that of V, reaching a predefined significance level α with error ϵ. That is

ISSN 2959-6157

$$P\left(\left|\overline{X}-\mu\right|<\epsilon\right)=1-\alpha,\text{where}\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i,\mu=E\left(V\right) \quad (5)$$

Since the mappings between meanings and phonetic structures in word forms are largely arbitrary across languages, the paper modeled the process of sampling potential cognates as a random sampling process, with the exemplars in X being modeled as i.i.d sampled, due to the enormous set of vocabularies in languages. Therefore, by the central limit theorem, the normalized sound correspondences in X approximates the standard normal deviation. Therefore,

$$P\left(Z<U_{\alpha/2}\right)=1-\alpha,\text{where}Z=\frac{\left|\overline{X}-\mu\right|}{\overline{\sigma}},\sigma^2=Var\left(\overline{X}\right) \quad (6)$$

In addition, the variance of $\bar{X}$ can be expressed by the variance of the distribution of sound correspondences in V as follows

$$\bar{\sigma}^2=Var\left(\bar{X}\right)=\frac{\sigma^2}{n}\frac{N-n}{N-1}\approx\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right),\text{where }\sigma^2=Var\left(V\right) \quad (7)$$

From equation (6) and (7), it is obtained that

$$P(\left|\bar{X}-\mu\right|<\bar{\sigma}U_{\alpha/2})=P(\left|\bar{X}-\mu\right|<\frac{\sigma}{\sqrt{n}}U_{\alpha/2}\left(\sqrt{1-\frac{n}{N}}\right)=1-\alpha \quad (8)$$

which combined with equation 5 yields

$$\epsilon=\frac{\sigma}{\sqrt{n}}U_{\alpha/2}\left(\sqrt{1-\frac{n}{N}}\right)=n(\frac{\epsilon}{\sigma U_{\alpha/2}})^2=1-\frac{n}{N} \quad (9)$$

Solving for n finally gives

$$n=\frac{N(\sigma U_{\alpha/2})^2}{N\epsilon^2+(\sigma U_{\alpha/2})^2} \quad (10)$$

The 2016 paper goes on to assume that the probability of having a sound correspondence in a pair of exemplars with semantic equivalence is p. Then the variance of V can be approximated using the variance of X, which gives

$$\sigma^2=\text{Var}\left(V\right)\approx\overline{\sigma^2}=p\left(1-p\right) \quad (11)$$

Linking it with equation 11 yields

$$n=\frac{NU_{\alpha/2}^2p\left(1-p\right)}{N\epsilon^2+U_{\alpha/2}^2p\left(1-p\right)} \quad (12)$$

The authors of the paper considered the strictest condition ($\epsilon=0.05$, $\alpha=0.05$, $N=5000$), giving $n\approx357$, and the most relaxed condition ($\epsilon=0.01$, $\alpha=0.1$, $N=4000$), giving $n\approx67$. Thus, one can see that the two Swadesh lists introduced above all lie in the interval. In addition, the analysis also suggests that including much bigger list of concepts of over 400 words doesn't bring additional advantage to glottochronology, which is consistent with other discussions. The paper also implemented the Ansari-Bradley test and Spearmn's rho, where they concluded that in most situations, the 100-word Swadesh lists are not statistically special, meaning that other randomly sampled 100 words from the 200-word Swadesh list will have the same sound correspondence distribution as the Swadesh lists.

However, the task of sampling concepts and words still remains a complicated task. This is largely due to the variance between languages that makes a word list hard to be universal. For instance, in his critique of Swadesh's method, Hoijer proposed 28 points of difficulty which he encountered in filling out the 100-item list for Navaho. Two points of difficulty are comprised by the pairs this, that and who, what. In each pair, one item has single Navaho equivalent (this, who), while the other items (that, what) have multiple equivalents. Hymes suggested that a way of solving the issue is by following the two principles: (a) If a specific form could represent multiple items, but is the only option for one particular item, assign it to that unique match; (b) if two forms are potential equivalents for a single item, prioritize the one that does not semantically overlap with other items in the test set [5]. In addition, the different environment where the languages originate also contributes to the difficulty of coming up with a universal word list. For instance, some linguists advocate for using over 300 concepts for Sino-Tibetian languages. The reason behind this is because some words in the Swadesh list, such as "to bark", doesn't have corresponding word forms in some Tibeto-Burman, Miao-Yao, or Zhuang-Dong languages, and other words such as "bamboo" are more stable and resistant to borrowing.

Overall, former statistical analysis of the sampling process suggests that the optimum number of words sampled ranges from 67 to 357. However, in different languages, other consideration, such as the environmental factors and linguistical variations, can cause the optimal selection of words to deviate from statistical analysis. The author hypothesize that it might be necessary to design a word list for every family of languages in order to maintain the accuracy of Glottochronology analysis.

2.3 Recognizing Cognative Correspondence

After compiling a proper word list, the next step is to identify the potential cognates between the two languages, i.e. words that are developed from the same word in a common parent language. The most accurate identification method is by the careful use of the comparative method in reconstructing the proto-language. However, most of the times, detailed comparative studies are not available, and the probable cognates need to be estimated by the inspection method [1].

The inspection method consists of the following procedures: (1) Register borrowed word, either from each other or from a common source, as probable noncognates; (2) Isolate the equivalent morphemes in each pair of words; (3) Test the pairs of equivalent morphemes isolated to determine whether or not they can be considered probable cognates. The probable cognates are determined using the following criterion: (a) Identical phonemes occupying corresponding positions within the morpheme pair; (b) Phonetically similar phonemes (such as t and d, or k and k^w) situated in corresponding positions; (c) Phonetically dissimilar phonemes that are considered matches because their differences are driven by specific environmental factors; (d) Phonemes that exhibit a consistent, systematic pattern of correspondence across comparable positions in the dataset.

This method relies on linguistical techniques and may be sometimes inaccurate. Therefore, Zhang and Gong proposed a statistical approach to calculate the threshold for identifying recurrent sound correspondences among potential correspondences detected in assembled words. The authors considered the number of times a sound in a word matches that of another word such that the correspondence cannot be considered a coincidence. By modeling the probability of a sound occurring k times by accident as binomial distribution, the k that makes the probability to drop below a significant level, e.g. 5%, can be calculated. In addition, the threshold for correspondence is dynamic, meaning that the value of k varies with the frequency of the sound, calculated by multiplying the frequency of that sound in the two languages. For rare sounds, a threshold of 2 is enough to verify correspondence. However, for common sounds (e.g. $/b/-\varnothing$), the threshold rises to 3 or 4 [4].

2.4 Lexicon Evolution

Another key assumption and component of glottochronology is that the rate of replacement remains perfectly constant. However, due to many political and social changes, the rate of replacement may vary, distorting the accuracy of the analysis. In addition, Serva also investigated the effect of the statstical fluctuations with $\mathcal{M}$ and Ω, and concluded the stochastic movement of the values may also lead to erroneous results [3].

The author defines M independent stochastic variables $\sigma_i(t)$ such that $\sigma_i(t)=1$ if the two words for the concept i are cognates, otherwise $\sigma_i(t)=0$. Then, one has

$$\Omega(T)=\sum_{i=1}^{M}\sigma_i(T) \tag{13}$$

In addition, if $\sigma_i(t)=1$, it may become zero at time $t+dt$ with probability $2\lambda dt$, and if $\sigma_i(t)=0$, it remains unchanged at time $t+dt$ with probability 1. Therefore, the author obtains the following differential solution.

$$\frac{d}{dt}E\left[\sigma_i(t)\right]=-2\lambda E\left[\sigma_i(t)\right] \tag{14}$$

which yields the solution $E[\sigma_i(T)]=e^{-2\lambda T}$, given that $\sigma_i(0)$ equals 1. Given that σ_i is a Bernoulli variable,

$$\sigma_i T=\begin{cases}1 & \text{withprobability}e^{-2\lambda T}\\ 0 & \text{withprobability}1-e^{-2\lambda T}\end{cases} \tag{15}$$

with variance

$$\mathrm{Var}[\sigma_i(T)]=E[\sigma_i^2(T)]-E[\sigma_i(T)]^2=e^{-2\lambda T}(1-e^{-}2\lambda T) \tag{16}$$

The author then defined the cognate distance between the two languages to be $\frac{1}{M}\sum_i(1-\sigma_i)=1-\omega(T)$, which takes 0 if the two languages are identical and 1 if when the two languages are completely different, i.e. not having a common ancestor language. Then,

$$\begin{aligned}E\left[\omega(T)\right]&=E\left[\sigma_i(T)\right]=e^{-2\lambda T}\\ \mathrm{Var}\left[\omega(T)\right]&=\frac{1}{M}\mathrm{Var}\left[\sigma_i(T)\right]=\frac{1}{M}e^{-2\lambda T}\left(1-e^{-2\lambda T}\right)\end{aligned} \tag{17}$$

Since $\omega(T)$ is approximately Gaussian distributed, therefore, with probability 95%,

$$E[\omega]-2\sqrt{\mathrm{Var}[\omega]}\leqslant\omega\leqslant E[\omega]+2\sqrt{\mathrm{Var}[\omega]} \tag{18}$$

From Eq. (17), it is found that $T=-\frac{1}{2\lambda}ln(E[\omega])$, while the observed stochastic separation time $\mathcal{T}_\omega$ is $\mathcal{T}_\omega=-\frac{1}{2\lambda}ln(\omega)$. According to Eq. (18), one has a 95% that $T_-\leqslant\mathcal{T}_\omega\leqslant T_+$, where

$$T_\pm=-\frac{1}{2\lambda}ln\left(E[\omega]\mp2\sqrt{\mathrm{Var}[\omega]}\right) \tag{19}$$

Thus, the relative error on separation time can be calculated as

$$R_\omega=\frac{T_+-T_-}{2T}=\frac{1}{4\lambda T}ln\left(\frac{1+2\sqrt{\frac{e^{2\lambda T}-1}{M}}}{1-2\sqrt{\frac{e^{2\lambda T}-1}{M}}}\right) \tag{20}$$

The author plotted the relative error with respect to time and found that R_ω ranges from 49% for short time (0.3 millennia) to 18 % for long time (6 millennia), an error

ISSN 2959-6157

significant enough in reconstructing he genealogicial tree of a family of languages, especially when the ancestor language is very close to the present language. The author also concluded that other methods, such as using Hamming distance, are better in terms of accuracy of the analysis [3].

3. Experiments and Results

The method of glottochronology has been experimented with different families of languages in various occasions with various degrees of success.

3.1 Indoeuropean Classification

Perhaps the most experimented languages is the Indoeuropean family. Dyen et. al. classified the Indoeuropean languages using the lexicostatistics method in 1992. In their paper, they also introduced the new technique of box diagrams. Using the 200-word Swadesh lists, the authors found that the results of the lexicostatistics analysis strongly coincides with the mainstream traditional classifications of the language. However, certain discrepencies between traditional method and glottochronology are also discovered [5]: (1) Failure to identify the Indoiranian group; (2) Identification of Slovenian group and the tripartite division of Slavic; (3) Identification of Proto-Blatoslavic group; (4) Identification of Mesoeuropeic group; (5) Evidence against the Ingveonic hypothesis.

Apart from this study, another research used 22 phonological, 13 morphological and 259 lexical features as coded characters and with combination of computer program, they were able to produce a tree with a "perfect phylogeny" algorithm that tracked the branching of twenty-four ancient and medieval Indo-European languages. However, this classification is not perfect as well, as the position of Germanic could not be determined. A subsequent work in 2005 pointed out that this was due to the fact that Germanic was apparently in contact with the other branches and therefore did not fit the "perfect phylogeny" [2].

Another study employed the algorithms for estimating the divergence time of DNA and applied to linguistics, and confirmed Colin Renfrew's theory on the Anatolian origin of Indo-European languages. Nevertheless, this result is challenged by Chang et. al., where they added ancestry costraints. Their results are roughly in accordance with the dating of the so-called steppe hypothesis in archaeology as regards the homeland of the speakers of Proto-Indo-European [2].

One of the most recent studies concerning the Indoeuropean language family is done by Kassian et. al., where they used three stages of dataset, root cognacy, derivational drift-free, and homoplasy-optimized data to produce a family tree of Indoeuropean languages. Their main finding is the multifurcation of the Inner Indoeuropean clade into four branches ca. 3357–2162 bc: (1) Greek-Armenian, (2) Albanian, (3) Italic-Germanic-Celtic, (4) Balto-Slavic–Indo-Iranian. In another study, mBERT model and K-means machine learning algorithm are used to classify the Indoeuropean languages into 9 groups, which show generally good correspondence with traditional methods, with small inaccuracies (classifying Romanian into Germanic languages) [6].

Some studies also suggest some potential flaws in traditional Glottochronology. For instance, in a 2019 study on Romance languages, the authors found that the word replacement rates associated with the meanings i are not all equal but they may differ by a factor two or more. Also, they concluded that the stability ranking is definitively not the same for different families of languages [7].

3.2 Other Language Families

Lexicostatistical analysis on other language families is significantly less than that on Indoeuropean languages. Some instances of such studies are presented in this subsection.

A 2021 study on the Mandailing and Angkola languages from Sumatra, using the traditional Swadesh list and techniques, concluded that the two languages are in the same sub-family, with estimated separation time at 1,235 AD. The level of kinship of regional languages in Indonesia has not been studied much regarding kinship and also family relations in the linguistic sector between one language and another. Especially for languages that are in the area and also adjacent areas. Therefore, the study has significance in the study of Indonesian languages [8].

Another study on the Dene-Caucasian Macrofamily (DCM) uses rooted network analysis and other techniques to understand the classification of DCM languages. The authors concluded that DCM is a complicated tree of languages, largely influenced by language interactions and language divergence. However, they found evidence that the early spread of DCM languages is the Na-Dene migration [9,10].

1. Conclusion

Lexicocstatistics is a powerful that integrates statistics and mathematics into linguistical studies. The main steps of the traditional Swadesh method include coming up with a word list, identifying cognate pairs, and apply calculations. Each step has potential problems when carrying out the analysis and real life, leading to inaccurate results. Nonetheless, this paper also presented certain improvements on the original method that increases the reliability

of glottochronology. Some past studies on different families of languages are also explored in the paper.

The majority of those studies are focused on Indoeuropean languages, which produces similar results to the mainstream results, but some aspects deviate from the conventional thoughts on those languages. Analysis of other languages using glottochronology also offered new insights in those languages. Since the overwhelming majority of the studies using lexicostatistics are done on Indoeuropean languages, the validity of the method on other languages are underexplored, which could be a potential area of focus for readers. Overall, lexicostatistics is a novel research area that has a lot of potential, especially with the help of computers and AI.

References

[1] Swadesh M. Towards greater accuracy in lexicostatistic dating. International Journal of American Linguistics, 1955.

[2] Piwowarczyk D. Computational approaches to linguistic chronology and subgrouping//Olander T, ed. The Indo-European Language Family. Cambridge: Cambridge University Press, 2022: 33-51.

[3] Serva M. Evolution of the lexicon: A probabilistic point of view. Journal of Statistical Mechanics: Theory and Experiment, 2025(11): 113404.

[4] Zhang M, Gong T. How many is enough?—Statistical principles for lexicostatistics. Frontiers in Psychology, 2016, 7: 1916.

[5] Dyen I, Kruskal J B, Black P. An Indoeuropean classification: A lexicostatistical experiment. Transactions of the American Philosophical Society, 1992.

[6] Kassian A S, Zhivlov M, Starostin G, et al. Rapid radiation of the inner Indo-European languages: An advanced approach to Indo-European lexicostatistics. Linguistics, 2021, 59(4): 949-979.

[7] Pasquini M, Serva M. Stability of meanings versus rate of replacement of words: An experimental test. Journal of Quantitative Linguistics, 2019.

[8] Ii H I H, Zulfitri Z, Amin T S. Lexicostatistics study of Mandailing and Angkola languages. Jurnal Educatio FKIP UNMA, 2021, 7(1): 265-275.

[9] Peter the Great Museum of Anthropology and Ethnography of the Russian Academy of Sciences, Kassian A. The Dene-Caucasian macrofamily: Lexicostatistical classification and homeland. Etnografia, 2023(3).

[10] Hymes D H. Lexicostatistics so far. Current Anthropology, 1960.