

PAPER • OPEN ACCESS

# Evolution of the lexicon: a probabilistic point of view

To cite this article: Maurizio Serva *J. Stat. Mech.* (2025) 113404

View the [article online](#) for updates and enhancements.

## You may also like

- [Gaussian quantum metrology for mode-encoded parameters](#)  
Giacomo Sorelli, Manuel Gessner, Nicolas Treps et al.
- [Malagasy dialects in Mayotte](#)  
Maurizio Serva and Michele Pasquini
- [L183, a Quiescent Core?](#)  
Jian-Jun Zhou, Xing-Wu Zheng and Yu-Xi Chen

PAPER: Interdisciplinary statistical mechanics

## Evolution of the lexicon: a probabilistic point of view

Maurizio Serva 

Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica,  
Università dell'Aquila, L'Aquila, Italy  
E-mail: [serva@univaq.it](mailto:serva@univaq.it)

Received 1 October 2025

Accepted for publication 23 October 2025

Published 17 November 2025



Online at [stacks.iop.org/JSTAT/2025/113404](https://stacks.iop.org/JSTAT/2025/113404)  
<https://doi.org/10.1088/1742-5468/ae195d>

**Abstract.** The Swadesh approach for determining the temporal separation between two languages relies on the stochastic process of word replacement (when a completely new word emerges to represent a given concept). It is well known that the basic assumptions of the Swadesh approach are often unrealistic due to various contamination phenomena and misjudgments (horizontal transfers, variations over time and space of the replacement rate, incorrect assessments of cognacy relationships, presence of synonyms, and so on). All of this means that the results cannot be completely correct. More importantly, even in the unrealistic case that all basic assumptions are satisfied, simple mathematics places limits on the accuracy of estimating the temporal separation between two languages. These limits, which are purely probabilistic in nature and which are often neglected in lexicostatistical studies, are analyzed in detail in this article. Furthermore, in this work, we highlight that the evolution of a language's lexicon is also driven by another stochastic process: gradual lexical modification of words. We show that this process also represents a major contribution to the reshaping of the vocabulary of languages over the centuries, and we also show, from a purely probabilistic perspective, that taking into account this second



Original Content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

random process significantly increases the precision in determining the temporal separation between two languages.

**Keywords:** language evolution, lexicostatistics, morris swadesh, stochastic replacement of words, stochastic modification of words, overlap and distance between languages

## Contents

1. An unavoidable preamble.....	2
2. Cognate overlap: the classical glottochronology .....	6
3. Normalized Hamming distance and overlap .....	8
4. Blind use of normalized Hamming overlap .....	10
5. Normalized Hamming overlap limited to cognate pairs .....	13
6. Conclusions.....	15
Acknowledgment .....	16
Appendix A .....	16
Appendix B .....	19
References.....	22

### 1. An unavoidable preamble

Glottochronology is a lexicostatistical tool that allows us to establish when a mother tongue gave rise to two daughter languages, or, what is the same thing, how many years have passed since the most recent common ancestor of two related languages. This methodology, in principle, also allows us to reconstruct the family trees of language families, with dates assigned to each branching point.

The theory, conceived by Morris Swadesh in the 1950s, was inspired by the success of the carbon-14 dating technique developed at the time [1–5].

Swadesh’s founding hypothesis was that a word associated with a given concept in a language could be replaced in time by a totally new one at a constant probability rate, exactly as a carbon-14 atom can be replaced at a constant rate by a more stable carbon atom.

Actually, the replacement of a word by a totally new synonym in a language is a rare but observable event whose probability rate seems to be more or less constant over the centuries. Therefore, if one considers the  $M$  words corresponding to  $M$  concepts in a given language (a Swadesh list), the number of un-replaced words after a time  $T$  will

be approximately

$$\mathcal{M}(T) \simeq M e^{-\lambda T}, \quad (1)$$

where  $\lambda$  is the replacement rate of a single word. Obviously,  $\mathcal{M}(0) = M$  (there are no changed words at the initial time) and  $\lim_{T \rightarrow \infty} \mathcal{M}(T) = 0$  (all words are changed at very large times). The replacement rate was estimated by Swadesh to be  $\lambda = 1.4 \times 10^{-4}$  per year, in appendix B we give an independent estimate of  $\lambda$  which confirms this value.

The consequence of this assumption can be easily understood by an example: consider the same Swadesh list of  $M$  concepts in Spanish and Latin and suppose you find by a direct check a number  $\mathcal{M}$  of pairs of words in the two languages which correspond to the same concept and which are cognates (equal words in this context since eventual difference is only due to random modification in spelling and pronunciation due to geographical, political and historical factors), then your estimate of the separation time  $T$  between the two languages can be obtained by means of (1) as

$$T \simeq -\frac{1}{\lambda} \ln\left(\frac{\mathcal{M}}{M}\right). \quad (2)$$

More often it happens that you do not know the vocabulary of the ancestor language; nevertheless you would like to know the time separation between two contemporary languages, which means the time from their most recent common ancestor. In this case, the Swadesh assumption tells you that the probability rate that in the pair of words corresponding to the same concept in the two languages at least one of them replaced is  $2\lambda$ . Therefore, given two Swadesh lists of  $M$  words, the number of cognate pairs in the two languages after a time  $T$  will be approximately

$$\Omega(T) \simeq M e^{-2\lambda T}. \quad (3)$$

The main consequence of this hypothesis can be easily understood with an example: consider the same Swadesh list of  $M$  concepts in Spanish and Italian and suppose to find by direct observation a number  $\Omega$  of pairs of words in the two languages that correspond to the same concept and that are cognate (equal words in this context since any eventual difference is only due to small repeated random changes in phonology and/or orthography), then, the estimated separation time  $T$  between the two languages can be obtained by means of (1) as

$$T \simeq -\frac{1}{2\lambda} \ln\left(\frac{\Omega}{M}\right). \quad (4)$$

Hereafter, we will name ‘cognate overlap’ the intensive variable  $\omega = \frac{\Omega}{M}$  which goes from 0 (languages totally different) to 1 (coinciding languages). This is the glottochronological formula, as first proposed by Swadesh in [1] (as an aside, in a footnote, he gives credit to the physicist Gordon Gould for the mathematical aspects of the work). This technique is the lexicostatistical equivalent of the molecular clock, used in evolutionary biology to estimate the time elapsed since two species diverged from their common ancestor. The molecular clock, similarly to glottochronology, is based on the idea that DNA

and protein sequences evolve at a constant rate over time and it is therefore based on counting the number of genetic coincidences between two species.

There are a number of reasons why the equalities (2) and (4) do not exactly hold, the main of which are:

- The hypothesis the rate of replacement remains perfectly constant in time may be not respected. Evidence is that the rate depends on historical events, for example at the end of Roman Empire, there were rapid lexical evolution due to the political and social changes caused by the entry of new populations in regions previously under Roman rule, on the contrary, the later standardization of national Romance languages was a factor in slowing down their lexical evolution;
- The creation of a Swadesh list implies a choice between various synonyms for any concept. This choice is not univocally determined by a codified rule and, in consequence, the cognate overlap may change according to different approaches. For example, in Italian there are two different words to say cheese, one is *formaggio*, which is the same as the French *fromage*, the other is *cacio*, the same as the Spanish *queso*;
- Languages swap words, thus increasing their cognate overlap (if loans are not detected). For example, Italian borrowed the following words from French: toilette, bricolage and bidet and both languages borrowed some common words from English. This phenomenon, named horizontal transfer, is typically more relevant between those languages that are geographically closer;
- There is a big problem concerning the individuation of cognates (i.e. equal words). It is easy to see that the Italian word *uomo* and the French word *homme* are the same, it is less evident that *leite* in Portuguese and *γάλα* (*gála*) in Greek are the same word, we can only infer this through historical phonological reconstruction. Moreover, the detection of cognates is a matter of individual choice and, therefore, it may be different for different scholars.

The point of this paper is that there is a more fundamental problem with the equalities (1) and (3):  $\mathcal{M}(T)$  and  $\Omega(T)$  are random quantities so that statistical fluctuations from the average may lead to an erroneous evaluation of the separation time by (2). In fact, the standard deviation of  $\frac{\mathcal{M}(T)}{M}$  and  $\omega = \frac{\Omega(T)}{M}$  are proportional to  $\frac{1}{\sqrt{M}}$ . While fluctuations are not a problem concerning the carbon-14 decay since  $M$  (the number of atoms) is of the order of the Avogadro number ( $M \simeq 6 \times 10^{23}$ ), the largest Swadesh list only contains  $M = 207$  concepts. Therefore, the equality (1) only holds on average while the observed values of the stochastic variable  $\mathcal{M}(T)$  can be well different from its average, as we will show. Exactly the same problem arises with the equality (3) for the stochastic variable  $\Omega(T)$ .

This is a point that is widely underestimated and/or misunderstood in lexicostatistics: mathematics alone sets a limit to the precision with which one can determine the separation time between two languages. In this paper, we only focus on these probabilistic limits, neglecting the other relevant sources of possible incorrect estimation listed above. Therefore, we assume that the scholar dealing with the Swadesh approach is in the ideal, although highly unrealistic, situation in which:

- The rate of replacement is strictly constant in time and it is universal (the same for all languages);
- He is able to univocally draw up the Swadesh list (absence of synonyms);
- Languages do not swap words (no horizontal transfers are allowed);
- He is able to determine cognacy with absolute precision (no false cognate pairs, no missing cognate pairs).

Nevertheless, as we will show, even if these unrealistic ideal assumptions are satisfied, the estimated and the real separation times can be catastrophically different only because of probabilistic fluctuations. This is really a relevant bug which is to be attributed exclusively to the fact that the number of concepts is small (207) at variance with carbon-14 dating, where  $M$ , the number of atoms, is thermodynamically large ( $6 \times 10^{23}$ ). We will perform in section 2 a precise analysis of the mathematical limits of validity of the Swadesh method, assuming that all the ideal conditions listed above are satisfied.

A second, but not less relevant, point of this paper is a mathematical analysis concerning the advantage of using a different and, ultimately, complementary strategy, to evaluate the time separation between languages. Swadesh only used lexical replacements, while he did not consider the gradual independent modification of cognate words as a possible source of information. The idea to take advantage of this second source arises from the simple observation that two initially identical terms employed in different geographical locations (say Madrid and Florence) become gradually different because of small cumulative modifications, even in the absence of replacements (such as *homo*  $\rightarrow$  *hombre/uomo*).

Although a single random change in one of two cognate words (such as the modification of a single character) has a smaller impact concerning their differentiation compared to a replacement, it is much more frequent. We will show, indeed, that the processes of replacement and gradual modification have comparable cumulative random effects on lexicon evolution, but the latter with smaller fluctuations. In sum, replacements and gradual modifications equally contribute to the reshaping of the vocabulary of languages over the centuries, but taking into account the second random process significantly increases the precision in determining the time separation between two languages.

As for replacements, we have here a purely probabilistic perspective described in section 3, which assumes that random gradual modifications take place at a constant and universal rate  $\mu$ . Moreover, we will make the simplifying assumptions that all words are composed of the same number  $L$  of characters (the length of words is  $L$ ) and that each character may assume  $N$  possible values (the effective number of letters of the alphabet is  $N$ ). Both the effective quantities  $L$  and  $N$  are experimentally estimated in appendix B.

In section 4 we will describe a strategy where only the ‘Normalized Hamming Distance’ of words is used to compare the difference between languages. In our probabilistic approach we maintain the simplification that all words have equal length. The method applies blindly to all concepts, independently of the fact that the two associates words in the two languages are cognate or not. Therefore, with respect to the Swadesh approach, the advantage is that no prior identification of cognates is requested, thus avoiding all possible human errors associated with this procedure. On experimental

ground, the method was proposed in [6, 7] and impressively later applied to the huge Automated Similarity Judgment Program dataset, which includes nearly all recognized language families in the world and many subfamilies [8]. In these researches, given that real words may have different length, ‘normalized Levenshtein distance’ was used instead of normalized Hamming distance. In fact, while Levenshtein [9] and Hamming distances are both edit distances the first also allows for insertions and deletions and for this reason it has to be used in case of words of different length.

More recently a combined strategy has been used where word distances have been calculated only for cognate pairs [10]. In section 5 we analyze on a mathematical ground this strategy, which, for very large separation times, works better than the blind strategy of section 4, nevertheless, the price to pay is the possible errors associated with the identification of cognate terms, as in the Swadesh approach.

## 2. Cognate overlap: the classical glottochronology

The number of concepts entering in a Swadesh list is  $M$ , each of them is individuated by an index  $i = 1, \dots, M$ . Given a pair of contemporary languages one has  $M$  words for the  $M$  concepts for each of them. Thus, we define  $M$  independent stochastic variables  $\sigma_i(t)$  such that  $\sigma_i(t) = 1$  if the two words for the concept  $i$  are cognates, otherwise  $\sigma_i(t) = 0$ . Therefore, the number of cognate pairs at time  $T$  (time from the last common ancestor) is

$$\Omega(T) = \sum_{i=1}^M \sigma_i(T). \quad (5)$$

The cognate overlap, which we have defined as

$$\omega(T) = \frac{\Omega(T)}{M}, \quad (6)$$

goes from 0 (totally different languages) to 1 (coinciding languages).

According to Swadesh the probability that at least one of the two words in a cognate pair is replaced in a time  $dt$  is  $2\lambda dt$ . When one of the two words is replaced a pair of cognates becomes a pair of non-cognates while a pair of non-cognates remains a pair of non-cognates. Therefore, if  $\sigma_i(t) = 1$  it may vanish at time  $t + dt$  with probability  $2\lambda dt$  while if  $\sigma_i(t) = 0$  it remains unchanged at time  $t + dt$  with probability 1. One immediately has the differential equation for the expected value at the left below

$$\frac{d}{dt} E[\sigma_i(t)] = -2\lambda E[\sigma_i(t)] \quad \rightarrow \quad E[\sigma_i(T)] = e^{-2\lambda T}, \quad (7)$$

where the solution at the right above is obtained by assuming that all  $\sigma_i(0)$  equal 1 (at time  $T = 0$  the two languages start to differentiate). Given that  $\sigma_i(T)$  is a Bernoulli variable which takes the values 0 and 1, the probability that  $\sigma_i(T) = 1$  equals the expected

value  $E[\sigma_i(T)]$ , while the probability that  $\sigma_i(T) = 0$  equals  $1 - E[\sigma_i(T)]$ , therefore,

$$\sigma_i(T) = \begin{cases} 1 & \text{with probability } e^{-2\lambda T} \\ 0 & \text{with probability } 1 - e^{-2\lambda T}. \end{cases} \quad (8)$$

The variable  $\sigma_i(T)$  can be interpreted as the cognate overlap between the two words corresponding to the same concept  $i$  while  $1 - \sigma_i(T)$  is their cognate distance (if  $\sigma_i(T) = 1$  the distance equals zero, if  $\sigma_i(T) = 0$  the distance equals one). The cognate distance between the two languages can be simply defined as  $\frac{1}{M} \sum_i (1 - \sigma_i) = 1 - \omega(T)$ , which ranges from 0 to 1 and equals zero for two identical languages (all pairs are composed by cognate words) and equals one when the two languages are completely different (all pairs are composed by non-cognate words). In the last case one concludes that the two languages have no common ancestor.

Given that the variables  $\sigma_i(T)$  only take the values 0 and 1, one has  $\sigma_i^2(T) = \sigma_i(T)$  and given the expectation at the right in (7) it follows  $E[\sigma_i^2(T)] = E[\sigma_i(T)] = e^{-2\lambda T}$ , which implies the variance

$$\text{Var}[\sigma_i(T)] = E[\sigma_i^2(T)] - E[\sigma_i(T)]^2 = e^{-2\lambda T} (1 - e^{-2\lambda T}). \quad (9)$$

According to (5) and (8), the variable  $\Omega(T)$  follows a binomial distribution of parameters  $e^{-2\lambda T}$  and  $M$ . We prefer to compute here directly its expected value and its variance since we will later use similar calculations for variables slightly more complicated. Given that  $\Omega(T)$  is the sum of  $M$  independent and identically distributed variables and given (6), not only one has that

$$E[\omega(T)] = E[\sigma_i(T)] = e^{-2\lambda T}, \quad \text{Var}[\omega(T)] = \frac{1}{M} \text{Var}[\sigma_i(T)] = \frac{1}{M} e^{-2\lambda T} (1 - e^{-2\lambda T}), \quad (10)$$

but also that  $\omega(T)$  is approximately Gaussian distributed, therefore, with probability 95%:

$$E[\omega] - 2\sqrt{\text{Var}[\omega]} \leq \omega \leq E[\omega] + 2\sqrt{\text{Var}[\omega]} \quad (11)$$

where here and hereafter the argument  $T$  is dropped.

From the first equality in (10) one trivially derives

$$T = -\frac{1}{2\lambda} \ln(E[\omega]), \quad (12)$$

nevertheless, the observed stochastic separation time  $\mathcal{T}_\omega$  is

$$\mathcal{T}_\omega = -\frac{1}{2\lambda} \ln(\omega). \quad (13)$$

It is the second quantity which forcefully enters in Swadesh method (a single realization of the stochastic processes) as well in its subsequent versions (see, for example [11–13],)

and in all applications (see, for example, [14–17]). So the point is how different it is  $\mathcal{T}_\omega$  from  $T$ ? According to (11) one has with probability 95%

$$T_- \leq \mathcal{T}_\omega \leq T_+ \quad (14)$$

where

$$T_\pm = -\frac{1}{2\lambda} \ln \left( E[\omega] \mp 2\sqrt{\text{Var}[\omega]} \right). \quad (15)$$

A reasonable measure of the relative error on separation time is therefore

$$R_\omega = \frac{T_+ - T_-}{2T} = \frac{1}{4\lambda T} \ln \left( \frac{E[\omega] + 2\sqrt{\text{Var}[\omega]}}{E[\omega] - 2\sqrt{\text{Var}[\omega]}} \right) = \frac{1}{4\lambda T} \ln \left( \frac{1 + 2\sqrt{\frac{e^{2\lambda T} - 1}{M}}}{1 - 2\sqrt{\frac{e^{2\lambda T} - 1}{M}}} \right). \quad (16)$$

It should be noticed that when  $M$  refers to the number of atoms in a decay process this error disappears since  $M$  is of the order of the Avogadro number ( $M \simeq 6 \times 10^{23}$ ), but it is relevant for glottochronology where  $M = 207$  as it can be seen in figure (1) where  $R_\omega$  is plotted as a function of  $T$  for  $3 \times 10^2 \leq T \leq 6 \times 10^3$ . In fact  $R_\omega$  ranges from 49% for very short times (0.3 millennia) to 18% for large times (6 millennia). This error also implies that one can hardly reconstruct the exact topology of the genealogical tree of a family of languages, especially if the ancestor language, is too close to the present (as it is for the Romance family) or there is more than a single branching event close to the root.

### 3. Normalized Hamming distance and overlap

It is a matter of fact that words, even in the absence of a replacement, are subject to modifications over time. This explains why the Latin word *caseum* transformed into the cognate words *queso* in Castilian and *cacio* in Italian. We assume that these random gradual modifications take place at a constant and universal rate  $\mu$ , which is the mirror hypothesis of a constant  $\lambda$ . Moreover, we make the simplifying assumptions that all words are composed of the same number  $L$  of characters (the length of words is  $L$ ) and that each character may assume  $N$  possible values (the effective number of letters of the alphabet is  $N$ ). Both the effective quantities  $L$  and  $N$  will be experimentally estimated in appendix B comparing pairs of words that certainly have no common ancestor.

These values may significantly differ from the nominal values (the number of letters in the English alphabet is 26 but  $N$  typically has a much smaller value). The reasons for this discrepancy are fundamentally two; the different frequency of different letters and the fact that letters are not monads (in English *h* appears often after a *t*). Nevertheless, if we replace our estimates for  $N$  and  $L$  respectively by 26 (letters in the English alphabet) and by the average length of English words, our final argument would be even stronger. Indeed, the real values of  $N$  and  $L$  are not truly relevant for our argument.

The normalized Hamming overlap of two words is simply the number of pairs of equal characters (in the same position) in the two words, divided by the length  $L$  of the words. This number ranges from 0 (totally different words) to 1 (identical words). The

normalized Hamming distance of two words, as defined in a standard way, is simply the number of pairs of different characters (in the same position) in the two words, divided by  $L$ .

It is very important to remark that from an operative point of view there is no need of deciding if the two words corresponding to the same concept  $i$  are cognate or not, nevertheless, the hidden probabilistic structure makes a difference between pairs of cognate words (whose normalized Hamming overlap is  $\eta_i(T)$ ) and pairs of non-cognate words (whose overlap is  $\xi_i(T)$ ). In fact, two non cognate words have no common origin and, therefore, two characters in the same position may be equal at any time  $T$  only by mere chance. On the contrary, for two cognate words, the two characters are necessarily equal at the initial time  $T = 0$  (words are initially identical).

Let us now consider the case of a pair of non-cognate words corresponding to the same concept  $i$ , the letter in position  $k$  is the same for the two words only by chance (at any time  $T$ , replacements of letters or words do not alter this fact!) therefore having defined  $\xi_{i,k}(T)$  as the variable which takes the value 1 if the two letters are equal and 0 otherwise, one has

$$\xi_{i,k}(T) = \begin{cases} 1 & \text{with probability } \frac{1}{N} \\ 0 & \text{with probability } \frac{N-1}{N}. \end{cases} \quad (17)$$

Then one can define the two variables

$$\xi_i(T) = \frac{1}{L} \sum_{k=1}^L \xi_{i,k}(T), \quad 1 - \xi_i(T) = \frac{1}{L} \sum_{k=1}^L (1 - \xi_{i,k}(T)), \quad (18)$$

which are respectively the normalized Hamming overlap and the normalized Hamming distance for two non-cognate words corresponding to the concept  $i$ .

Let us now consider the case of a pair of cognate words corresponding to the same concept  $i$ , the two letters in position  $k$  are the same at the initial time but later this fact may change because one of the two letters can be randomly substituted by a new one. We assume that in a time  $dt$  a letter value can be substituted with probability  $\mu dt$  by one of the  $N$  possible values (thus, including the departure value), therefore having defined  $\eta_{i,k}(t)$  as the variable which takes the value 1 if the two letters are equal and 0 otherwise, one has that if  $\eta_{i,k}(t) = 1$  then  $\eta_{i,k}(t + dt)$  may vanish at time  $t + dt$  with probability  $2\mu \frac{N-1}{N} dt$ , while if  $\eta_{i,k}(t) = 0$  then  $\eta_{i,k}(t + dt)$  may turn to the value 1 at time  $t + dt$  with probability  $\frac{2\mu}{N} dt$ . In a compact form:

$$\eta_{i,k}(t + dt) = \begin{cases} 1 - \eta_{i,k}(t) & \text{with probability } 2\mu \frac{1+(N-2)\eta_{i,k}(t)}{N} dt \\ \eta_{i,k}(t) & \text{with probability } 1 - 2\mu \frac{1+(N-2)\eta_{i,k}(t)}{N} dt. \end{cases} \quad (19)$$

After some biking, taking into account that  $\eta_{i,k}^2(t) = \eta_{i,k}(t)$ , one obtains the differential equation for the expected value at the left below

$$\frac{d}{dt} E[\eta_{i,k}(t)] = -2\mu \left( E[\eta_{i,k}(t)] - \frac{1}{N} \right) \rightarrow E[\eta_{i,k}(T)] = \left[ \frac{N-1}{N} e^{-2\mu T} + \frac{1}{N} \right], \quad (20)$$

where the solution at the right above is obtained by assuming that all  $\eta_{i,k}(0)$  equal 1.

Given that  $\eta_{i,k}(T)$  is a Bernoulli variable which takes the values 0 and 1, the probability that  $\eta_{i,k}(T) = 1$  equals the expected value  $E[\eta_{i,k}(T)]$ , therefore,

$$\eta_{i,k}(T) = \begin{cases} 1 & \text{with probability } \left[ \frac{N-1}{N} e^{-2\mu T} + \frac{1}{N} \right], \\ 0 & \text{with probability } \frac{N-1}{N} [1 - e^{-2\mu T}]. \end{cases} \quad (21)$$

Notice that the variables  $\xi_{i,k}(T)$  also are subject to the stochastic equations (19) but their probabilities does not vary in time since they are in the steady state distribution already at  $T = 0$ . Notice, in fact, that the probabilities (17) can be obtained from the (21) in the limit  $T \rightarrow \infty$  (when two languages lose memory of their common ancestor).

Let us stress that  $\mu$  is the probability rate that a dice is rolled independently of the output value which may eventually equal the departure one with probability  $\frac{1}{N}$ , therefore

$$\hat{\mu} = \frac{N-1}{N} \mu \quad (22)$$

is the probability rate that a letter in a word is substituted by a different one. Our estimate for  $\hat{\mu}$  is  $1.3 \times 10^{-4}$  (see appendix B). Since the probability rate of changing a given letter as a consequence of a replacement is  $\lambda = 1.4 \times 10^{-4}$ , we conclude that the effect of replacements and gradual modifications has a comparable effect over lexical evolution.

Finally, one can define the two variables

$$\eta_i(T) = \frac{1}{L} \sum_{k=1}^L \eta_{i,k}(T), \quad 1 - \eta_i(T) = \frac{1}{L} \sum_{k=1}^L (1 - \eta_{i,k}(T)), \quad (23)$$

which are respectively the normalized Hamming overlap and the normalized Hamming distance for two cognate words corresponding to the concept  $i$ .

Expected values and variances of  $\xi_i(T)$  and  $\eta_i(T)$  are computed in appendix A, while  $N$ ,  $L$  and  $\mu$  (the rate of a dice roll) are experimentally estimated in appendix B.

#### 4. Blind use of normalized Hamming overlap

Suppose you are unable to decide if two words corresponding to the same concept in two different languages are cognate, or you simply want to avoid errors that may arise from this decision, in this case you simply and blindly measure the normalized Hamming distance or overlap. For what we have exposed in the two previous sections, It is immediate to find out that this overlap is

$$\omega_i^H(T) = \sigma_i(T) \eta_i(T) + (1 - \sigma_i(T)) \xi_i(T), \quad (24)$$

because  $\sigma_i(T)$  equals 1 if the two words are cognate and equals 0 otherwise, moreover  $\eta_i(T)$  apply in the first case and  $\xi_i(T)$  in the second. The average and variance of the

above overlap are computed in appendix A. Moreover, the normalized word distance according to Hamming is

$$d_i^H(T) = 1 - \omega_i^H(T) = \sigma_i(T)(1 - \eta_i(T)) + (1 - \sigma_i(T))(1 - \xi_i(T)), \quad (25)$$

whose average and variance are also computed in appendix A.

Let us stress once more that both overlap and distance, when operatively measured for a pair of words of a dataset, do not require that one knows if the two words are cognate or not, but only a count of the number of pairs of equal (different) letters for each of the  $L$  positions.

It is also useful to define for future use

$$\phi_i(T) = \frac{N}{N-1} \left[ \omega_i^H(T) - \frac{1}{N} \right] = 1 - \frac{N}{N-1} d_i^H(T), \quad (26)$$

whose average and variance are derived in appendix A.

The normalized Hamming overlap between two languages can be obtained by an average over all the concepts included in the Swadesh lists, therefore

$$\omega^H(T) = \frac{1}{M} \sum_{i=1}^M \omega_i^H(T), \quad (27)$$

which is a thermodynamically intensive variable. Analogously, normalized Hamming distance between two languages is

$$d^H(T) = \frac{1}{M} \sum_{i=1}^M d_i^H(T) = 1 - \omega^H(T). \quad (28)$$

It is worth to mention that on a experimental ground the method normalized Levenshtein distance replaces normalized Hamming distance [6–8, 18–27], since words may have different lengths, nevertheless Hamming and Levenshtein are two different versions of edit distance and, therefore, they are basically the same thing. The reason we use here Hamming distance is simply because the related mathematics is much easier.

It is also finally useful to define a third intensive variable as

$$\phi(T) = \frac{1}{M} \sum_{i=1}^M \phi_i(T) = \frac{N}{N-1} \left[ \omega^H(T) - \frac{1}{N} \right] = 1 - \frac{N}{N-1} d^H(T). \quad (29)$$

Given that  $\phi(T)$  is the sum of  $M$  independent and identically distributed variables divided by  $M$ , one immediately has

$$E[\phi(T)] = E[\phi_i(T)], \quad \text{Var}[\phi(T)] = \frac{1}{M} \text{Var}[\phi_i(T)], \quad (30)$$

where  $E[\phi_i(T)]$  and  $\text{Var}[\phi_i(T)]$ , as already mentioned, are computed in appendix A. It turns out

$$E[\phi(T)] = e^{-2(\lambda+\mu)T} \quad (31)$$

and

$$\begin{aligned} \text{Var}[\phi(T)] &= \frac{1}{M} e^{-2\lambda T} (1 - e^{-2\lambda T}) e^{-4\mu T} + \frac{1}{ML} e^{-2\lambda T} (1 - e^{-2\mu T}) \left[ e^{-2\mu T} + \frac{1}{N-1} \right] \\ &\quad + \frac{1}{ML(N-1)} (1 - e^{-2\lambda T}). \end{aligned} \tag{32}$$

We can now follow the same line of reasoning of section 2. Since  $\phi(T)$  is approximately Gaussian distributed, with probability 95% we have that

$$E[\phi] - 2\sqrt{\text{Var}[\phi]} \leq \phi \leq E[\phi] + 2\sqrt{\text{Var}[\phi]} \tag{33}$$

where here and hereafter the argument  $T$  is dropped. From the first equation in (30) one trivially derives

$$T = -\frac{1}{2(\lambda + \mu)} \ln(E[\phi]), \tag{34}$$

nevertheless, the observed stochastic separation time  $\mathcal{T}_\phi$  is

$$\mathcal{T}_\phi = -\frac{1}{2(\lambda + \mu)} \ln(\phi), \tag{35}$$

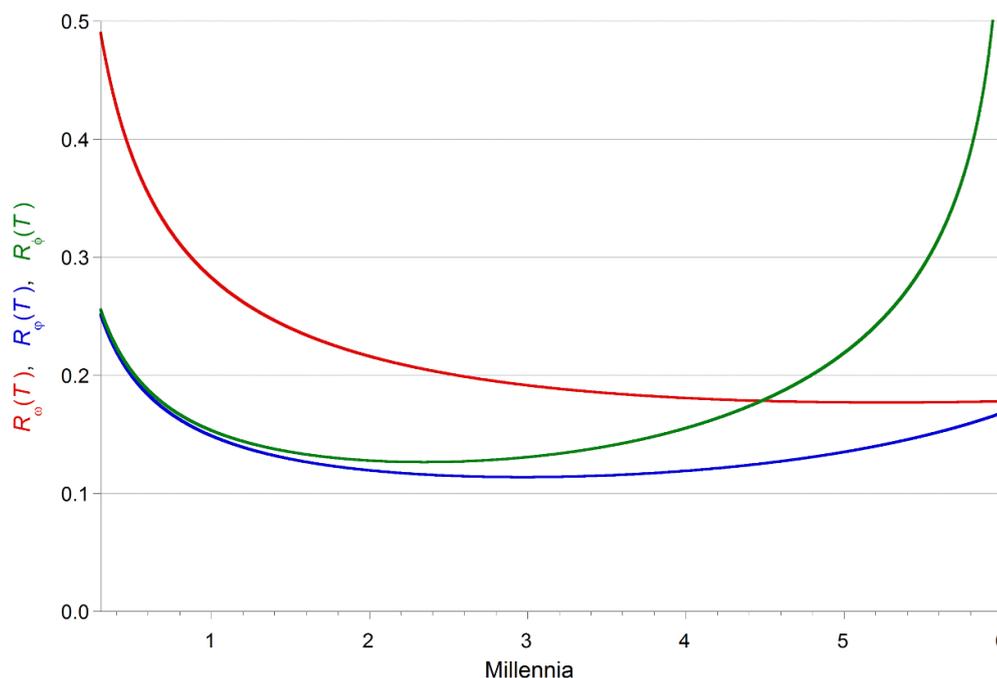
so the point is how different is it from  $T$ ? According to (33) one has that

$$-\frac{1}{2(\lambda + \mu)} \ln\left(E[\phi] + 2\sqrt{\text{Var}[\phi]}\right) \leq \mathcal{T}_\phi \leq -\frac{1}{2(\lambda + \mu)} \ln\left(E[\phi] - 2\sqrt{\text{Var}[\phi]}\right) \tag{36}$$

with probability 95%. A reasonable measure of the relative error concerning separation time can be found as in section 2:

$$R_\phi = \frac{1}{4(\lambda + \mu)T} \ln\left(\frac{E[\phi] + 2\sqrt{\text{Var}[\phi]}}{E[\phi] - 2\sqrt{\text{Var}[\phi]}}\right). \tag{37}$$

As can be observed in figure 1, the relative error  $R_\phi$  is smaller than  $R_\omega$  in the first four millennia, this means that the edit distance approach should be considered as a valid alternative to the Swadesh approach in this interval of time. This is even more true if we consider that there is no error caused by incorrect attributions of cognacy. The improvement is particularly relevant when the last common ancestor language is situated less than two millennia in the past (the relative error, in this case, is one half). On the contrary, for large times the relative error  $R_\phi$  explodes due to the increase of variance due to the contribution of non-cognate pairs, which, however, do not contain any useful information regarding the time  $T$ .



**Figure 1.** The relative errors  $R_\omega$  (red) as defined by (16), concerning the classical Swadesh approach,  $R_\phi$  (green) as defined by (30), (37) and (62), concerning the blind use of normalized edit distance and  $R_\varphi$  (blue) as defined by (40), (45) and (58), concerning the combined use of edit distance and cognate identification. We plot the three errors in function of  $T$  in the interval  $300 \leq T \leq 6,000$ . We use the Swadesh proposal for the replacement rate  $\lambda = 1.4 \times 10^{-4}$  which is confirmed by our estimate. We assume  $M = 207$ , again according to Swadesh. Moreover, we use our estimates  $\mu = 1.6 \times 10^{-4}$ ,  $N = 5.18$  and  $L = 7.63$ .

## 5. Normalized Hamming overlap limited to cognate pairs

Once we discovered that measuring the overlap of words in place of the number of cognates leads to useful results, we can try to combine the two strategies as it was hypothesized in [10]. We construct, therefore, a new intensive variable  $\varphi$  which only uses the overlap  $\eta_i(T)$  of cognate words and which takes the place of  $\omega$  (Swadesh approach) or  $\phi$  (blind use of Hamming distance). Let us preliminarily define

$$\varphi_i(T) = \frac{N}{N-1} \sigma_i(T) \left[ \eta_i(T) - \frac{1}{N} \right], \quad (38)$$

which vanishes when the two words corresponding to the concept  $i$  are non-cognates (when  $\sigma_i(T) = 0$ ) and which linearly depends on  $\eta_i(T)$  when they are non-cognate (when  $\sigma_i(T) = 1$ ). The average and variance of this variable are computed in appendix A.

Then, it is straightforward to define the intensive variable

$$\varphi(T) = \frac{1}{M} \sum_{i=1}^M \varphi_i(T). \quad (39)$$

It is worth noting that only cognate pairs contribute to this sum, therefore, from an operational point of view, all concepts that have cognate words must be identified and the Hamming distance must be calculated only for these word pairs.

Given that  $\varphi(T)$  is the sum of  $M$  independent and identically distributed variables divided by  $M$ , one immediately has

$$E[\varphi(T)] = E[\varphi_i(T)], \quad \text{Var}[\varphi(T)] = \frac{1}{M} \text{Var}[\varphi_i(T)], \quad (40)$$

where  $E[\varphi_i(T)]$  and  $\text{Var}[\varphi_i(T)]$ , as already mentioned, are computed in appendix A. It turns out

$$E[\varphi(T)] = e^{-2(\mu+\lambda)T} \quad (41)$$

and

$$\text{Var}[\varphi(T)] = \frac{1}{M} e^{-2\lambda T} (1 - e^{-2\lambda T}) e^{-4\mu T} + \frac{1}{ML} e^{-2\lambda T} (1 - e^{-2\mu T}) \left[ e^{-2\mu T} + \frac{1}{N-1} \right]. \quad (42)$$

As a consequence of equation (41) one has that the effective time from the last common ancestor is

$$T = -\frac{1}{2(\lambda + \mu)} \ln(E[\varphi]), \quad (43)$$

while the observed one is

$$\mathcal{T}_\phi = -\frac{1}{2(\lambda + \mu)} \ln(\phi). \quad (44)$$

We end up with the new measure of the relative error on the separation time

$$R_\varphi = \frac{1}{4(\lambda + \mu)T} \ln \left( \frac{E[\varphi] + 2\sqrt{\text{Var}[\varphi]}}{E[\varphi] - 2\sqrt{\text{Var}[\varphi]}} \right), \quad (45)$$

whose meaning should be clear at this point.

It can be shown that the relative error  $R_\varphi$  is always smaller than the relative error  $R_\phi$ , a fact that can be also visually appreciated in figure 1. First of all one can write

$$\phi_i(T) = \varphi_i(T) + \chi_i(T), \quad \text{where} \quad \chi_i(T) = \frac{N}{N-1} (1 - \sigma_i(T)) \left[ \xi_i(T) - \frac{1}{N} \right] \quad (46)$$

which can be easily verified by comparing (24), (26) with (38). Moreover, as shown in appendix A,  $E[\chi_i(T)] = 0$ , which means  $E[\phi_i(T)] = E[\varphi_i(T)]$ . Moreover, always in appendix A, it is shown that  $\text{Var}[\phi_i(T)] = \text{Var}[\varphi_i(T)] + \text{Var}[\chi_i(T)]$ . These equalities equally hold when we pass to the intensive variables obtained by summing over all concepts and dividing by  $M$ , therefore

$$E[\phi(T)] = E[\varphi(T)], \quad \text{Var}[\phi(T)] = \text{Var}[\varphi(T)] + \text{Var}[\chi(T)] \quad (47)$$

which finally imply  $R_\varphi \leq R_\phi$ . The reason for these inequalities finally relies on the fact that the contribution of the Hamming overlaps for non-cognate pairs, do not add information but only makes the signal more noisy.

At this point one would conclude that one should always use  $\varphi$  instead of  $\phi$ , but this is not true, because the computation of  $\varphi$  requests the individuation of all cognate pairs which is a procedure that can be subject to errors, as explained at the beginning of the paper.

It is true that the relative error  $R_\varphi$  is always smaller than  $R_\phi$ , but in the first three millennia they are almost equal, therefore the blind strategy should be preferred in this interval of time because the error due to the wrong attribution of cognacy relations likely produces an extra-variance larger than  $\text{Var}[\chi(T)]$ .

This is still true when using non-subjective automatic strategies for cognacy attribution [10, 28–32], thus supporting the empirical results in [32], where a comparison was made between the performances of edit-distance dating and glottochronological dating.

We leave the task of comprehensively discussing the advantages and disadvantages of the three approaches to the concluding section that follows.

## 6. Conclusions

If we take into account both the results shown in figure 1 and the considerations at the end of the previous section, it should be quite clear that up to three millennia the most convenient strategy is the one in section 4, which is based only on the use of the normalized edit distance for all pairs, thus without the need to identify cognacy relations between words. Indeed, both  $R_\phi$  and  $R_\varphi$  are much smaller than  $R_\omega$  and they are almost equal within this time window, thus between the two strategies based on the normalized edit distance the blind one should be preferred because there is no risk of increasing the error as a consequence of a subjective misattribution of cognacy relations.

It is true that we found the values of  $N$  and  $L$  rather roughly and that smaller actual values of  $N$  and/or  $L$  would increase the variances (32) and (42) and, consequently, increase the error  $R_\phi$  and  $R_\varphi$ . However, for the first three millennia both  $R_\phi$  and  $R_\varphi$  are almost independent of  $N$  and  $L$ . This is a consequence of the fact that in this time range the leading term of the variances (32) and (42) is  $\frac{1}{M}e^{-2\lambda T}(1 - e^{-2\lambda T})e^{-4\mu T}$ , that does not depend on these two parameters. Indeed, if we consider only this term we find

$$R_\phi \simeq R_\varphi \simeq \frac{\lambda}{\lambda + \mu} R_\omega,$$

which is indeed what can be visually perceived in figure 1 for the first three millennia.

For longer time horizons (over three millennia) it appears that the best choice should be the mixed strategy, i.e. the one that limits the calculation of the normalized edit distance to cognate pairs (section 5). Only for time horizons longer than six millennia could the Swadesh approach become competitive, but six millennia is considered the time when the glottochronological approach tests the limit of its validity for reasons inherent in the difficulty of identifying cognacy relations between words.

It must also be said that the approach based on the blind use of the normalized edit distance is much more manageable, for example, in the case of the Malagasy language [22, 23, 25, 26], 60 vocabularies are considered for as many varieties of the language, each with 207 words. Therefore, cognacy relations would have to be established subjectively for  $(60 \cdot 59 \cdot 207)/2 = 366,390$  word pairs, a huge task, which, on the contrary, can be avoided when using the strategy in section 4. Moreover, measuring normalized edit distances can be done automatically.

Recently, there has been some increase in the use of normalized edit distance as a tool for investigating also in fields not strictly limited to linguistics, as history, economics, migratory phenomena, esthetics and animal communication. For example, in [26] a specific scenario is proposed regarding the modalities of the Austronesian colonization of Madagascar based on the geographical gradient of introgression of Bantu words into the Malagasy varieties. The influence of language dissimilarity on international economic transactions and on economic growth is respectively quantified in [33] and [34], while [35] studied the effect of linguistic barriers in the destination language acquisition of immigrants. More intriguing, the effects of personality traits on phonaesthetic language ratings are investigated for the first time in [36] with the conclusion that to some degree the sound-based beauty of language is ‘in the ears of the beholder’, which it is not as trivial as it appears, considering that the study is strictly quantitative. Finally, there are two different studies carried out by two different research teams that concern the communication between individuals belonging to the whale species *Megaptera Novaeangliae* [37, 38].

The final conclusion we would like to draw is that methods based on normalized edit distance are not only more manageable and easier to use than traditional methods based on cognate counting (a fairly accepted fact), but they are also generally more accurate (a fact which is rarely recognized in the linguistic community). Moreover, they can be applied much more easily in fields other than traditional linguistics, such as animal communication, for which it is difficult to imagine an adaptation of the concept of cognacy.

## Acknowledgment

The research described in this paper has been developed in the framework of the research project National Centre for HPC, Big Data and Quantum Computing—PNRR Project, funded by the European Union—Next Generation EU.

The author thanks Michele Pasquini who performed the automatic calculations in appendix B, created the figures and had the patience to discuss some probabilistic aspects of the article. He also revised the manuscript, but in no way wanted to appear as a co-author.

## Appendix A

Given (17) one has  $E[\xi_{i,k}(T)] = \frac{1}{N}$ , moreover, given that  $\xi_{i,k}^2(T) = \xi_{i,k}(T)$  (the variable only takes the values 0 and 1), one also has  $E[\xi_{i,k}^2(T)] = \frac{1}{N}$  which leads to  $\text{Var}[\xi_{i,k}(T)] =$

$\frac{N-1}{N^2}$ . The variable  $\xi_i(T)$  is the normalized Hamming overlap between two non-cognate words corresponding to the same concept  $i$ . Its expression is given by the first equation in (18) which is an equal weight average of  $L$  *i.i.d.* variables. Therefore, one also has

$$E[\xi_i(T)] = E[\xi_{i,k}(T)] = \frac{1}{N}, \quad \text{Var}[\xi_i(T)] = \frac{1}{L} \text{Var}[\xi_{i,k}(T)] \\ = \frac{N-1}{LN^2}. \tag{48}$$

This result can be formulated as

$$E\left[\xi_i(T) - \frac{1}{N}\right] = 0, \quad E\left[\left(\xi_i(T) - \frac{1}{N}\right)^2\right] = \text{Var}\left[\xi_i(T) - \frac{1}{N}\right] = \text{Var}[\xi_i(T)] = \frac{1}{L} \frac{N-1}{N^2}. \tag{49}$$

Let us now define

$$\chi_i(T) = \frac{N}{N-1} (1 - \sigma_i(T)) \left[\xi_i(T) - \frac{1}{N}\right], \tag{50}$$

given that  $\sigma_i(T)$  and  $\xi_i(T)$  are independent variables and that  $E\left[(1 - \sigma_i(T))^2\right] = E[1 - \sigma_i(T)] = 1 - E[\sigma_i(T)] = 1 - e^{-2\lambda T}$  and also given the equalities (48), one has

$$E[\chi_i(T)] = 0, \quad \text{Var}[\chi_i(T)] = E[\chi_i^2(T)] = \frac{1}{L(N-1)} (1 - e^{-2\lambda T}). \tag{51}$$

We repeat now the above calculations for variables  $\eta_{i,k}(T)$  and related variables. Given the probabilities in (21) and given that  $\eta_{i,k}^2(T) = \eta_{i,k}(T)$ , it follows

$$E[\eta_{i,k}^2(T)] = E[\eta_{i,k}(T)] = \left[\frac{N-1}{N} e^{-2\mu T} + \frac{1}{N}\right] \rightarrow \text{Var}[\eta_{i,k}(T)] \\ = \frac{N-1}{N} (1 - e^{-2\mu T}) \left[\frac{N-1}{N} e^{-2\mu T} + \frac{1}{N}\right]. \tag{52}$$

The variable  $\eta_i(T)$  is the normalized Hamming overlap between two cognate words corresponding to the same concept  $i$ . Its expression is given by the first equation in (23) which is an equal-weight average of  $L$  *i.i.d.* variables. Given that the Bernoulli variables  $\eta_{i,k}(T)$  are all independent the variables  $\eta_i(T)$  have the following average value and variance:

$$E\left[\eta_i(T) - \frac{1}{N}\right] = E\left[\eta_{i,k}(T) - \frac{1}{N}\right] = \frac{N-1}{N} e^{-2\mu T}, \tag{53}$$

$$\text{Var}\left[\eta_i(T) - \frac{1}{N}\right] = \text{Var}[\eta_i(T)] = \frac{1}{L} \text{Var}[\eta_{i,k}(T)] = \frac{1}{L} \left(\frac{N-1}{N}\right)^2 (1 - e^{-2\mu T}) \left[e^{-2\mu T} + \frac{1}{N-1}\right]. \tag{54}$$

Notice that (53) and (54) reduce to (49) when  $T \rightarrow \infty$ , *i.e.*, when the steady state is reached. Let us consider now the variable  $\varphi_i(T)$  defined as

$$\varphi_i(T) = \frac{N}{N-1} \sigma_i(T) \left[ \eta_i(T) - \frac{1}{N} \right], \tag{55}$$

given the independence of  $\sigma_i(T)$  and  $\eta_i(T)$ , we have

$$E[\varphi_i(T)] = e^{-2(\mu+\lambda)T} \tag{56}$$

and, simply by the definition of variance, we also have

$$\text{Var}[\varphi_i(T)] = \left( \frac{N}{N-1} \right)^2 \left( E[\sigma_i^2(T)] \text{Var} \left[ \eta_i(T) - \frac{1}{N} \right] + E \left[ \eta_i(T) - \frac{1}{N} \right]^2 \text{Var}[\sigma_i(T)] \right), \tag{57}$$

Therefore, taking into account that  $E[\sigma_i^2(T)] = E[\sigma_i(T)] = e^{-2\lambda T}$  and  $\text{Var}[\sigma_i(T)] = e^{-2\lambda T}(1 - e^{-2\lambda T})$  we finally obtain

$$\text{Var}[\varphi_i(T)] = \frac{1}{L} e^{-2\lambda T} (1 - e^{-2\mu T}) \left[ e^{-2\mu T} + \frac{1}{N-1} \right] + e^{-2\lambda T} (1 - e^{-2\lambda T}) e^{-4\mu T}, \tag{58}$$

Let us now define

$$\phi_i(T) = \varphi_i(T) + \chi_i(T) = \frac{N}{N-1} \left[ \sigma_i(T) \eta_i(T) + (1 - \sigma_i(T)) \xi_i(T) - \frac{1}{N} \right], \tag{59}$$

given that  $E[\chi_i(T)] = 0$ , one has

$$E[\phi_i(T)] = E[\varphi_i(T)] = e^{-2(\mu+\lambda)T}, \quad \text{Var}[\phi_i(T)] = \text{Var}[\varphi_i(T)] + \text{Var}[\chi_i(T)] \tag{60}$$

where the last equality is a consequence of the fact that  $\varphi_i(T) \chi_i(T) = 0$ , which in turn is a consequence of the fact that  $\sigma_i(T)(1 - \sigma_i(T))$  identically vanishes. In fact,

$$\begin{aligned} \text{Var}[\phi_i(T)] &= E[(\varphi_i(T) + \chi_i(T))^2] - E[\varphi_i(T) + \chi_i(T)]^2 \\ &= E[\varphi_i^2(T)] + E[\chi_i^2(T)] - E[\varphi_i(T)]^2 \end{aligned} \tag{61}$$

where for the second equality we have used  $E[\chi_i(T)] = 0$ . Since  $E[\chi_i(T)] = 0$ , it is also true that the last expression above equals the last expression in (60). In conclusion:

$$\begin{aligned} \text{Var}[\phi_i(T)] &= \frac{1}{L} e^{-2\lambda T} (1 - e^{-2\mu T}) \left[ e^{-2\mu T} + \frac{1}{N-1} \right] + e^{-2\lambda T} (1 - e^{-2\lambda T}) e^{-4\mu T} \\ &\quad + \frac{1}{L(N-1)} (1 - e^{-2\lambda T}). \end{aligned} \tag{62}$$

## Appendix B

The dataset we use in this appendix to estimate the parameters  $\lambda$ ,  $\mu$ ,  $N$  and  $L$  consists in 60 Swadesh lists of 207 items, overall 12 420 terms collected by the author of the present paper during the years 2018 and 2019. Each list corresponds to a different variety of Malagasy, which is not simply identified by the name of the ethnicity but also by the location where the variety was collected. In turn, the location is identified by the name of a town/village and by latitude and longitude. This last information turns out to be relevant for the results in this appendix since we will need geographical distances for our deductions.

The reasons for using this dataset is that the orthographic realizations are identical for all varieties and that the criteria of selection of words is homogeneous and reliable. In fact, each list was furnished and checked at least by three native language speakers which, for each given concept, were asked to furnish the most common word in their variety as spoken in their town/village. The 60 Swadesh lists dataset, together with the information concerning ethnicities, towns/villages and latitudes and longitudes, can be found in the Supporting Information of [25].

Let us start with parameters  $N$  and  $L$ . The equalities in (49) can be rewritten in a slightly different form as follows:

$$\frac{1}{N} = E[\xi_i], \quad \frac{1}{L} = \frac{N^2}{N-1} E\left[\left(\xi_i - \frac{1}{N}\right)^2\right]. \quad (63)$$

were the argument  $T$  of the  $\xi_i$  has been dropped because inessential in this context. The variable  $\xi_i$  is the normalized Hamming overlap between two non-cognate words, while  $d_i = 1 - \xi_i$  is their normalized Hamming distance. The two above equations can be rewritten using the distances as follows:

$$\frac{1}{N} = E[1 - d_i], \quad \frac{1}{L} = (N-1) E\left[\left(1 - \frac{N}{N-1}d_i\right)^2\right]. \quad (64)$$

It is important to underline that these two equalities hold when the distance is between two non-cognate words corresponding to the same concept in two languages, but they must also hold when the two compared words correspond to different concepts in the same language or to different concepts in two different languages (provided they have the same characteristics in terms of effective length and number of letters).

Then, the effective number  $N$  of different letters' values can be estimated by replacing the theoretical normalized Hamming distance with effectively measurable normalized Levenshtein distance of words referring to different concepts

$$\frac{1}{N} = \frac{1}{S} \sum_{\alpha, \beta} \sum_{i \neq j} [1 - \mathcal{D}(\alpha, i | \beta, j)], \quad (65)$$

where  $\mathcal{D}(\alpha, i | \beta, j)$  is the normalized Levenshtein distance between the word  $i$  of language  $\alpha$  and the word  $j$  of language  $\beta$  is computed using the dataset collected by the author. The first sum goes on all possible pairs of languages (included on the same

language, i.e.,  $\alpha = \beta$ ), the second sum only goes on pairs referring to different concepts ( $i \neq j$ ). The normalization is obtained by  $S = \sum_{\alpha, \beta} \sum_{i \neq j} 1$ . The statistical average above replaces the first probabilistic average in (64), which is reasonable provided the statistics is sufficient.

Following the same procedure we find from the second probabilistic average in (64) that the effective number  $L$  of letters in a word can be estimated by

$$\frac{1}{L} = \frac{N-1}{S} \sum_{\alpha, \beta} \sum_{i \neq j} \left[ 1 - \frac{N}{N-1} \mathcal{D}(\alpha, i | \beta, j) \right]^2. \quad (66)$$

Since we consider 60 contemporary varieties of Malagasy, the number of language pairs is  $(60 \cdot 59)/2$  while, given that a list contains 207 concepts, the number of pairs corresponding to different concepts is  $207 \cdot 206$  thus,  $S = 207 \cdot 206 \cdot 60 \cdot 59 / 2 \simeq 75 \cdot 10^6$ , a number which is sufficiently large to justify the fact that the statistical averages (65), (66) reasonably replace the probabilistic ones in (64).

We find  $N = 5.1770$ , and  $L = 7.6334$ . There is a big difference between this value of  $N$  and the number of letters in the Malagasy alphabet which is 21. This is because  $N$  is an effective number that takes into account the fact that the frequency of the various characters is very different, moreover, letters are not independent inside words (syllables, consonant pairs, ...).

We perform now an independent evaluation of the value of  $\lambda$  which will confirm the value estimated by Swadesh. We rewrite the main formula of glottochronology as

$$\lambda = -\frac{1}{2T} \ln E[\omega(T)], \quad (67)$$

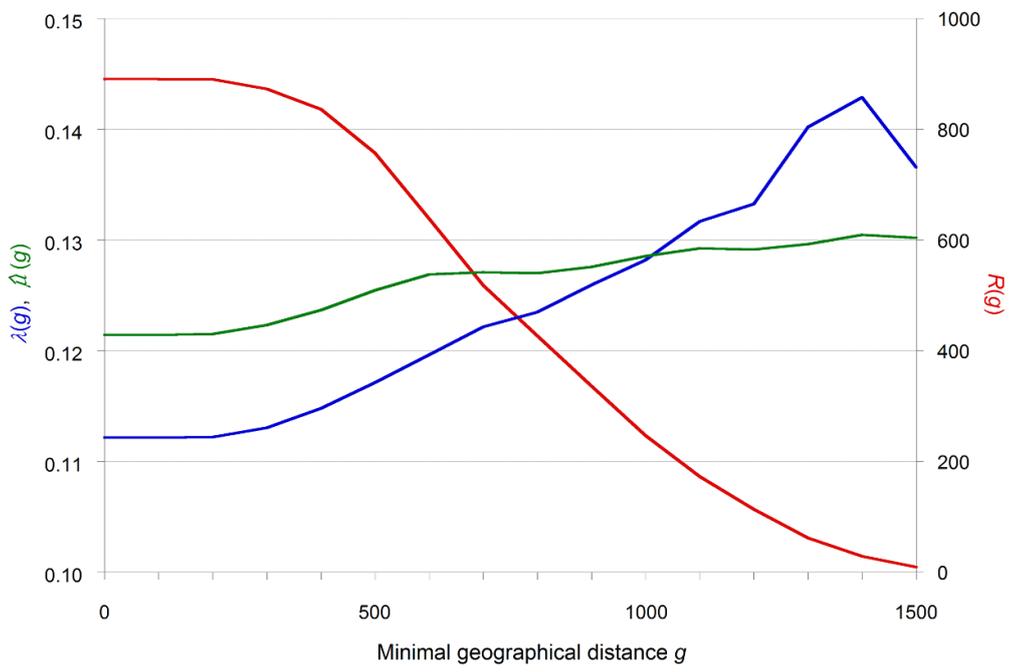
then, we consider the 60 contemporary varieties of Malagasy, and we replace the probabilistic average (67) by the statistical average

$$\lambda(g) = -\frac{1}{2T} \ln \left( \frac{1}{R(g)} \sum_{\alpha, \beta} \omega(\alpha | \beta) \right). \quad (68)$$

where  $\omega(\alpha | \beta)$  is the experimentally determined overlap (number of cognate pairs divided by 207) for varieties  $\alpha$  and  $\beta$ . The cognate pairs are automatically detected following the procedure in [24]. The sum goes on all pairs of varieties that not only match at the root of the tree (see tree in [25], figure 2), but also whose geographical distance is larger than a given threshold  $g$  expressed in Km.

The reason why we ask that all pairs of languages considered in (68) must match at the root is that this choice implies that  $T$  in the formula (68) is the same for all pairs and it is the time distance from the last common ancestor of all Malagasy varieties (whose known value is  $T = 1350$  years, see [22]). Moreover, the reason why we ask that all pairs of languages considered in (68) must also be at a geographical distance larger than  $g$  is what we would like to avoid, at our best, contamination due to horizontal transfers between geographically close varieties. The larger is  $g$  the smaller is contamination, but unfortunately the larger is  $g$ , the smaller is  $R(g)$ , which is the number of elements in the average ( $R(g) = \sum_{\alpha, \beta} 1$ )

Evolution of the lexicon: a probabilistic point of view



**Figure 2.** The parameters  $\lambda(g)$  (blue),  $\hat{\mu}(g) = \frac{N-1}{N}\mu(g)$  (green) and  $R(g)$  (red) as a function of the minimal geographical distance  $g$  (in Km) between two varieties in a pair. The parameter  $\lambda(g)$  is the rate of words replacements,  $\hat{\mu}(g)$  is the rate of effective characters changes and  $R(g)$  is the number of language pairs involved in the average, *i.e.*, the number of language pairs which match at the root of the family tree and whose geographical distance is larger than  $g$ .

The estimated  $\lambda(g)$  is plotted in figure (1) (blue) in function of  $g$  as well as the number of pairs  $R(g) = \sum_{i,j} 1$  considered in the average (red). The larger is  $g$ , the better the result because a larger geographical distance corresponds to a smaller horizontal transfer between languages. This is true until  $R(g)$  becomes too small to have sufficient statistics. In figure 2 the geographical distance threshold  $g$  ranges between 0 and 1500 Km. For all values of  $g$  the value of  $\lambda(g)$  is compatible with the Swadesh estimate  $\lambda = 1.4 \times 10^{-4}$ , nevertheless, as expected, the best result is for  $1200 \leq g \leq 1400$  were the geographical distance is the largest compatible with the constraint of having sufficient statistics.

In a similar way, we can estimate  $\mu$ . Equations (29) and (34) rewrite as

$$\lambda + \mu = -\frac{1}{2T} \ln E \left[ 1 - \frac{N}{N-1} d^H(T) \right], \tag{69}$$

which corresponds to the statistical average

$$\mu(g) = -\frac{1}{2T} \ln \left[ \frac{1}{R(g)} \sum_{\alpha,\beta} \left( 1 - \frac{N}{N-1} \mathcal{D}(\alpha|\beta) \right) \right] - \lambda(g), \tag{70}$$

where  $\mathcal{D}(\alpha|\beta)$  is the ordinary normalized Levenshtein distance between the languages  $\alpha$  and  $\beta$  (of course,  $\alpha \neq \beta$ ) and where, again, the sum goes on all pairs of varieties which not only match at the root of the tree (see the tree in [25], figure 2), but also whose geographical distance is larger than a given threshold  $g$  expressed in Km.

In figure 2, indeed, we plot  $\hat{\mu}(g)$  (green) instead of  $\mu(g)$  because it is the rate of an effective character change. The most reliable result is for  $1200 \leq g \leq 1400$  where the geographical distance is the largest compatible with the constraint of having sufficient statistics. We find  $\hat{\mu} = 1.3 \times 10^{-4}$ , very close to the value  $\lambda = 1.4 \times 10^{-4}$  which means that character changes and word replacements equally contribute to the lexicon evolution.

## ORCID iD

Maurizio Serva  0000-0002-8117-1834

## References

- [1] Swadesh M 1950 Salish internal relationships *Int. J. Am. Linguist.* **16** 57–167
- [2] Swadesh M 1951 Diffusional cumulation and archaic residue as historical explanations *Southwest. J. Anthropol.* **7** 1–21
- [3] Swadesh M 1952 Lexico-statistic dating of prehistoric ethnic contacts *Proc. Am. Phil. Soc.* **96** 452–63
- [4] Swadesh M 1954 Perspectives and problems of Amerindian comparative linguistics *Word* **10** 306–32
- [5] Swadesh M 1955 Towards greater accuracy in lexicostatistic dating *Int. J. Am. Linguist.* **21** 121–37
- [6] Serva M and Petroni F 2008 Indo-European languages tree by Levenshtein distance *Europhys. Lett.* **81** 68005
- [7] Petroni F and Serva M 2008 Languages distance and tree reconstruction *J. Stat. Mech.* P08012
- [8] Holman E W *et al* 2011 Automated dating of the world's language families based on lexical similarity *Curr. Anthropol.* **52** 841–75
- [9] Levenshtein V I 1966 Binary codes capable of correcting deletions, insertions and reversals *Sov. Phys. Dokl.* **10** 707–10
- [10] Pasquini M, Serva M and Vergni D 2023 Gradual modifications and abrupt replacements: two stochastic lexical ingredients of languages evolution *Comput. Linguist.* **49** 301–23
- [11] Dyen I, James A T and Cole J W L 1967 Language divergence and estimated word retention rate *Language* **43** 150–71
- [12] Starostin S 1999 Comparative-historical linguistics and Lexicostatistics *In Historical Linguistics & Lexicostatistics, Melbourne, by Vitaly Shevoroshkin (Author)* vol 3–50 P Sidwell Translated and revised version of the Russian original *Sravnitel'no-istoričeskoe jazykoznanie i leksikostatistika Lingvističeskaja rekonstrukcija i drevnejšaja istorija Vostoka* Moscow 1989
- [13] Starostin G 2010 Preliminary lexicostatistics as a basis for language classification: a new approach *J. Lang. Relationship* **3** 79–116
- [14] Verin P, Kottak C P and Gorlin P 1969 The glottochronology of Malagasy speech communities *Oceanic Linguist.* **8** 26–83
- [15] Gray R D and Jordan F M 2000 Language trees support the express-train sequence of Austronesian expansion *Nature* **405** 1052–5
- [16] Gray R D and Atkinson Q D 2003 Language-tree divergence times support the Anatolian theory of Indo-European origin *Nature* **426** 435–9
- [17] Greenhill S J and Gray R D 2009 *Austronesian Language Phylogenies: Myths and Misconceptions About Bayesian Computational Methods (Austronesian Historical Linguistics and Culture History: a Festschrift for Robert Blust)* ed A Adelaar and A Pawley (Pacific Linguistics) pp 375–97
- [18] Bakker D, Müller A, Velupillai V, Wichmann S, Brown C H, Brown P, Egorov D, Mailhammer R, Grant A and Holman E W 2009 Adding typology to lexicostatistics: A combined approach to language classification *Linguistic Typol.* **13** 167–79
- [19] Petroni F and Serva M 2010 Measures of lexical distance between languages *Physica A* **389** 2280–3
- [20] Wichmann S, Holman E W, Bakker D and Brown C H 2010 Evaluating linguistic distance measures *Physica A* **389** 3632–9

- [21] Wichmann S, Holman E W, Müller A, Velupillai V, List J-M, Belyaev O, Urban M and Bakker D 2010 Glottochronology as a heuristic for genealogical language relationships *J. Quant. Linguist.* **17** 303–16
- [22] Serva M, Petroni F, Volchenkov D and Wichmann S 2012 Malagasy dialects and the peopling of Madagascar *J. R. Soc. Interface* **9** 54–67
- [23] Serva M 2012 The settlement of Madagascar: what dialects and languages can tell us *PLoS One* **7** e30666
- [24] Pasquini M and Serva M 2019 Stability of meanings versus rate of replacement of words: an experimental test *J. Quant. Linguist.* **28** 95–116
- [25] Serva M and Pasquini M 2020 Dialects of Madagascar *PLoS One* **5** e0240170
- [26] Serva M and Pasquini M 2022 Linguistic clues suggest that the Indonesian colonizers directly sailed to Madagascar *Lang. Sci.* **93** 101497
- [27] Wichmann S and Holman E W 2023 Cross-linguistic conditions on word length *PLoS One* **18** e0281041
- [28] Covington M A 1996 An algorithm to align words for historical comparison *Comput. Linguist.* **22** 481–96
- [29] Ciobanu A M and Dinu L P 2014 Automatic detection of cognates using orthographic alignment *Proc. 52nd Annual Meeting of the Association for Computational Linguistics* vol 2 pp 99–105
- [30] List J-M, Greenhill S and Gray R 2017 The potential of automatic word comparison for historical linguistics *PLoS One* **12** e0170046
- [31] Rama T and List J-M 2019 An automated framework for fast cognate detection and Bayesian phylogenetic inference *Proc. 57th Annual Meeting of the Association for Computational Linguistics* pp 6225–35
- [32] Rama T and Wichmann S 2020 A test of generalized bayesian dating: a new linguistic dating method *PLoS One* **15** e0236522
- [33] Ispording I E and Otten S 2013 The costs of Babylon - Linguistic distance in applied economics *Rev. Int. Econ.* **21** 354–69
- [34] Gören E 2018 Consequences of linguistic distance for economic growth *Oxford Bull. Econ. Stat.* **80** 0305–9049
- [35] Ispording I E and Otten S 2014 Linguistic barriers in the destination language acquisition of immigrants *J. Econ. Behav. Organ.* **105** 30–50
- [36] Winkler A, Kogan V V and Reiterer S M 2023 Phonaesthetics and personality - Why we do not only prefer Romance languages *Front. Lang. Sci.* **2** 1043619
- [37] Lilley M S, Garland E C, Rekdahl M, Noad M J, Goldizen A W and Garrigue C 2012 Improved versions of the Levenshtein distance method for comparing sequence information in animals vocalisations: tests using humpback whale song *Behaviour* **149** 1413–41
- [38] Magnúsdóttir E E and Lim R 2019 Subarctic singers: Humpback whale (*Megaptera novaeangliae*) song structure and progression from an Icelandic feeding ground during winter *PLoS One* **14** e0210057