frontiers in Psychology

ORIGINAL RESEARCH
published: 12 December 2016
doi: 10.3389/fpsyg.2016.01916

How Many Is Enough?—Statistical Principles for Lexicostatistics

Menghan Zhang[1]* *and Tao Gong*[2,3]*

[1] *Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center of Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China,* [2] *Haskins Laboratories, New Haven, CT, USA,* [3] *Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangdong, China*

Lexicostatistics has been applied in linguistics to inform phylogenetic relations among languages. There are two important yet not well-studied parameters in this approach: the conventional size of vocabulary list to collect potentially true cognates and the minimum matching instances required to confirm a recurrent sound correspondence. Here, we derive two statistical principles from stochastic theorems to quantify these parameters. These principles validate the practice of using the Swadesh 100- and 200-word lists to indicate degree of relatedness between languages, and enable a frequency-based, dynamic threshold to detect recurrent sound correspondences. Using statistical tests, we further evaluate the generality of the Swadesh 100-word list compared to the Swadesh 200-word list and other 100-word lists sampled randomly from the Swadesh 200-word list. All these provide mathematical support for applying lexicostatistics in historical and comparative linguistics.

Keywords: Swadesh lists, cognates, Bernoulli process, binomial distribution, Ansari-Bradley test, Spearman's rho

OPEN ACCESS

Edited by:
Denise Hsien Wu,
National Central University, Taiwan

Reviewed by:
Christoph Scheepers,
University of Glasgow, UK
Haitao Liu,
School of International Studies, China

***Correspondence:**
Menghan Zhang
hansonmenghan@163.com
Tao Gong
gong@haskins.yale.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *19 July 2016*
Accepted: *22 November 2016*
Published: *12 December 2016*

Citation:
Zhang M and Gong T (2016) How Many Is Enough?—Statistical Principles for Lexicostatistics. Front. Psychol. 7:1916. doi: 10.3389/fpsyg.2016.01916

INTRODUCTION

In linguistics, quantitative approaches such as lexicostatistics and glottochronology have been widely applied to detect hypothetical genetic relations among languages (McMahon and McMahon, 2005; Campbell, 2013). Lexicostatistics refers to the statistical manipulation of lexical materials for historical inferences that abstract away from exact dates (Hymes, 1960). Lexicostatistics compares languages for phylogenetic affinity based on proportion of cognates in a standard basic vocabulary list. Each slot in the list is a concept (meaning), and collected items (words) occupying the same slot are compared cross-linguistically. Some linguists suggest using the term "meaning list" instead of "word list" or "vocabulary list" because the latter two are potentially ambiguous (McMahon and McMahon, 2005). We thus do not make distinction between the terms vocabulary list and meaning list. Unlike lexicostatistics, glottochronology deals in particular with phylogenetic relationships among languages (Campbell, 2013). Strictly speaking, lexicostatistics is a broader approach than glottochronology without specific assumptions such as constant rate of word retention or loss.

Computing lexicostatistics generally proceeds in the following steps (McMahon and McMahon, 2005; Campbell, 2013):

(1) *Assemble a set of word forms from languages being compared based on a list of basic vocabulary.* It would be ideal to collect every word from languages being compared, yet it is infeasible to obtain an exhaustive or very large-scale collection of words, especially for endangered or poorly-documented languages. In practice, linguists usually conduct basic word assembly based on small-scale meaning lists. Two widely-adopted lists for this purpose are the Swadesh lists.

They compile 100 (Swadesh, 1955) or 200 (Swadesh, 1952) concepts. The choice of these concepts is determined mainly by linguistic intuitions and experiences. For example, it has been argued that words encoding these concepts are stable and resistant to borrowing; therefore, the chances that identified cognates are due to borrowing or contact, rather than phylogenetic relation, are reasonably low (note that about 10% of the Swadesh 100-word list are still prone to borrowing). The Swadesh lists have been employed to construct many linguistic datasets of Indo-European and Austronesian languages (e.g., Dyen et al., 1997; Lohr, 2000; Greenhill et al., 2008; Wichmann et al., 2013).

(2) *Identify lexical cognates based on recurrent sound correspondences.* Cognates and recurrent sound correspondences provide strong evidence of a common origin of languages. Recurrent sound correspondences typically occur in vocabulary of languages having phylogenetic relations (or systematically borrowed words in languages having a history of deep contact; Hoijer, 1956; Bergsland and Vogt, 1962). In definition, a "recurrent" sound correspondence must occur in at least two or more matching instances. However, given that there are not many assembled words for comparison, nobody can give a satisfactory answer to questions such as how many instances in the collected words that exhibit a sound matching would allow linguists to classify it as a recurrent sound correspondence, rather than borrowing (Hoijer, 1956; Hock and Joseph, 1996). In other words, there lacks a concrete threshold, in terms of the minimum number of sound matching instances, for determining a recurrent sound correspondence. In practice, linguists often adopt an iterative approach by exhaustively (if possible) listing all matching instances across a given word list to identify a recurrent sound correspondence.

Apart from classical lexicostatistics, there exist a number of additional quantitative approaches in language comparison research, all of which follow roughly the same steps as above (e.g., Oswalt, 1971; Ringe, 1992; Baxter and Ramer, 2000; Lohr, 2000; Kessler, 2001). Given a fixed vocabulary list and a number of lexical cognates exhibiting recurrent sound correspondences, hypothetical phylogenetic relations among languages can be verified.

Despite of its wide applications, there are a number of objections to lexicostatistics (Bergsland and Vogt, 1962; Eska and Ringe, 2004; McMahon and McMahon, 2006). Many of the critics focus on the composition of the vocabulary list, such as what concepts can be utilized for collecting potential cognates and whether it is possible to construct a universal concept list for cognate assembly. Other critics concern the uncertainties inherent in the two steps above. Some of these uncertainties deserve more discussion here (Baxter and Ramer, 2000).

First, it remains unclear whether comparison among 100 or 200 Swadesh words can reasonably demonstrate the relatedness between languages. Apart from the Swadesh lists, some linguists suggest using much smaller vocabulary lists for cognate collection. Some of these lists contain 40 (Holman et al., 2008), 35 (Starostin, 2000), 33 (Baxter and Ramer, 2000), or only 15 (Dolgopolsky, 1986) concepts. By contrast, others advocate using much bigger lists for this purpose, which consist of 300–500 concepts (Greenberg, 1990; Ruhlen, 1994; Li, 1995; Newman, 1995; Huang, 1997; Jiang, 2007). Linguistic intuitions and experiences are still the primary considerations to construct these lists (Heggarty, 2010), and many lists share several concepts with the Swadesh lists.

Second, the threshold of recurrent sound correspondence is subject to not only the size of the vocabulary list but also the occurring frequencies of involved segments in the list. For example, if the vocabulary list is big and two segments appear frequently in the assembled words according to this list, the chance of finding an accidental correspondence or borrowing between them would increase. Hence, it would require more instances of such correspondence to confirm whether it is a recurrent correspondence or not. By contrast, if the vocabulary list is small and two segments are less frequent, a small number (say, two) of matching instances is sufficient to confirm recurrent correspondence between the segments (Ringe, 1992; Kessler, 2001).

Third, Swadesh argued that the 100-word list could reliably reflect the vertical, inheritance relations among languages (Swadesh, 1955). Due to the ease of gathering 100 words compared to 200, many language comparison studies have directly used the Swadesh 100-word list for word assembly. Before taking this simpler approach, one needs to clarify whether the Swadesh 100-word list is quantitatively a special sub-list of the Swadesh 200-word list. This can be clarified by the following two questions:

(1) Whether the distribution of sound correspondences in the words collected by the Swadesh 100-word list can reliably resemble the distribution of the same correspondences in the words collected by the Swadesh 200-word list;

(2) Whether the distribution based on the Swadesh 100-word list can resemble those based on other 100-word sub-lists constructed by randomly sampling from the Swadesh 200-word list. In other words, whether a random split of the Swadesh 200-word list into two 100-word lists yields significantly distinct distributions; if so, the Swadesh 100-word list would not be a special sub-list of the Swadesh 200-word list.

In this paper, we attempt to tackle the above uncertainties from a mathematical perspective. We propose two statistical principles to calculate, respectively, a "conventional" size of the vocabulary list to collect and judge potential cognates and a "reasonable" threshold to identify recurrent sound correspondences. Here, the "conventional" size means the most convenient and informative size of the vocabulary list especially in situations where it is not possible to collect many word forms, or where there is little prior knowledge about the languages being compared. The "reasonable" threshold means the flexible threshold at least accounting for the occurring frequencies of phonetic segments in the collected words. With more information about languages being available, the actual threshold can be updated accordingly. Apart from these principles, we also adopt some standard

statistical tests to evaluate the generality of the Swadesh 100-word list. All the analyses are done in MATLAB (ver. 2015a) and based on some well-known examples. They can be reasonably extended to other complex cases.

In the following sections, we derive the statistical principles from stochastic sampling theorems, and apply them in real cases of assembling cognates and detecting recurrent correspondences. Based on the empirical data and statistical tests, we also evaluate the generality of the Swadesh 100-word list. For the sake of simplicity, we only address two-language comparison. Finally, we discuss the importance of these principles to lexicostatistics.

STATISTICAL PRINCIPLES AND EXAMPLE RESULTS

Conventional Size of the Vocabulary List to Assemble Potential Cognates

We model the task of setting a conventional size of the basic vocabulary list for collecting potentially true cognates as a statistical task of constructing an exemplar set by sampling from a total set. Here, the total set refers to the total vocabulary V of a language, which contains N words. For the sake of simplicity, we assume that V in each language being compared has roughly the same size. The exemplar set X ($\subset V$) contains n words (x_1, x_2, x_3, ... x_n, $n < N$), each chosen from V. Under this setting, the statistical task is to determine n, such that the distribution of sound correspondences in X (obtained by comparing each pair of words aligned by semantic equivalence) approximately matches (reaching a predefined significance level α and within a predefined error rate ε) the distribution of those correspondences in V. In mathematical terms, such matching can be described as in Equation (1):

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

In practice, sampling potential cognates is not random, but guided by the concepts in the adopted vocabulary list. However, in all languages, mappings between meanings and phonetic structures in word forms are largely arbitrary (Hockett, 1960; De Saussure, 1983; Chomsky, 1995; Hock and Joseph, 1996; Hurford, 2012), in the sense that apart from social convention of using word A for meaning B there is no explicit connection between the sound of a word and aspects of its meaning (Dingemanse, 2012). Note that there exist a small proportion of words that show iconicity between the form and meaning aspects (e.g., onomatopoeia, words imitating natural sounds, often in a highly language-specific way; and ideophones, words vividly evoking sensory impressions like sounds, movements, textures, visual patterns, or actions; Dingemanse, 2012; Dingemanse et al., 2015), but these words and their concepts usually do not appear in the vocabulary list for cognate assembly. Such arbitrariness allows us to reasonably simplify the process of sampling potential cognates as a random sampling process.

Strictly speaking, sampling potential cognates is conducted without replacement; after choosing a word from V and putting it to X, remove it from V. However, considering the much bigger size N of V than the size n of X ($N \gg n$) and the arbitrariness in mappings between semantics and phonetics, the sampling process can be viewed as a process with replacement (still keep the word in V after sampling it). In this way, exemplars in X are independent and identically distributed (*i.i.d.*).

The above simplifications enable us to apply stochastic sampling theorems to this task. According to *the central limit theorem* (the probability distribution of the mean of *i.i.d.* variables with finite variance approximately follows a normal distribution), no matter whether the distribution of sound correspondences in V is known or not, the distribution of normalized sound correspondences in X approximates the standard normal distribution with mean 0 and standard deviation 1. In mathematical terms, this distribution can be described by Equation (2) (see Walpole et al., 2011: p. 234 for proof):

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

Here, $U_{\frac{\alpha}{2}}$ is set according to significance level α (see **Table 1**).

According to the statistics of random sampling with replacement, the variance of the distribution of normalized sound correspondences in X can be expressed by the variance of the distribution of sound correspondences in V, as in Equation (3) (see Walpole et al., 2011: p. 154, p. 767 for proofs):

$$\begin{aligned}\bar{\sigma}^2 &= Var\left(\bar{X}\right)=\frac{\sigma^2}{n}\frac{N-n}{N-1}\approx\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right),\ where\ \sigma^2\\ &= Var(V)\end{aligned} \quad (3)$$

Linking Equation (2) with Equation (3), we have:

$$\begin{aligned}P\left(\left|\bar{X}-\mu\right|<\bar{\sigma}\,U_{\frac{\alpha}{2}}\right) &= P\left(\left|\bar{X}-\mu\right|<\frac{\sigma}{\sqrt{n}}U_{\frac{\alpha}{2}}\sqrt{1-\frac{n}{N}}\right)\\ &= 1-\alpha\end{aligned} \quad (4)$$

Linking Equation (4) with Equation (1), we have:

$$\varepsilon=\frac{\sigma}{\sqrt{n}}U_{\frac{\alpha}{2}}\sqrt{1-\frac{n}{N}},\ then,\ n\left(\frac{\varepsilon}{\sigma\,U_{\frac{\alpha}{2}}}\right)^2=1-\frac{n}{N} \quad (5)$$

Solving for n from Equation (5), yields:

$$n=\frac{N\left(\sigma\,U_{\frac{\alpha}{2}}\right)^2}{N\varepsilon^2+\left(\sigma\,U_{\frac{\alpha}{2}}\right)^2} \quad (6)$$

Given a list of collected words with semantic equivalence from languages being compared, comparison between each pair of words having equivalent meanings and, respectively, from the two languages being compared represents a single trial of phonetic matching between the two languages. Such comparison has two possible outcomes: match or mismatch (see **Figure 1**). Note that this is obviously an initial model that

TABLE 1 | $U_{\frac{\alpha}{2}}$ of the standard normal distribution at different significant level α, and the conventional sizes calculated using Equation (8) at error rate ε = 0.05 or 0.1 and total vocabulary size N = 4000 or 5000.

α	0.2	0.1	0.05	0.02	0.01	0.002
$U_{\frac{\alpha}{2}}$	1.282	1.645	1.96	2.326	2.576	3.09
n ($\varepsilon = 0.05$, $N = 4000$)	157.866	253.456	350.498	476.568	569.158	770.815
n ($\varepsilon = 0.1$, $N = 4000$)	40.670	66.526	93.788	130.833	159.288	225.260
n ($\varepsilon = 0.05$, $N = 5000$)	159.122	256.709	356.750	488.202	585.829	801.713
n ($\varepsilon = 0.1$, $N = 5000$)	40.753	66.748	94.230	131.694	160.567	227.826

Gray cells indicate the calculated conventional sizes in the eight conditions of our estimation.

glosses over problems of semantic ambiguity and one-to-many or many-to-one matches. Considering that the exemplars in X are independent and identically distributed (*i.i.d.*), the process of detecting sound correspondence among the exemplars can be conceived as a Bernoulli process having two outcomes (exhibiting a sound correspondence or not). Accordingly, the probability distribution of sound correspondences in X follows a Binomial Distribution, in which the probability of having a sound correspondence in a pair of exemplars with semantic equivalence is p (then, the probability of not having a sound correspondence is $1-p$).

Following the binomial distribution, the variance of the distribution in V approximates the variance of the distribution in X, the latter of which is calculated as in Equation (7) (see Walpole et al., 2011: p. 130 for proof):

$$\sigma^2 = Var\,(V) \approx \bar{\sigma}^2 = p(1\ -\ p) \tag{7}$$

Link Equation (7) with Equation (6) and we have:

$$n = \frac{N\left(U_{\frac{\alpha}{2}}\right)^2 p(1\ -\ p)}{N\varepsilon^2 + \left(U_{\frac{\alpha}{2}}\right)^2 p(1\ -\ p)} \tag{8}$$

Without prior knowledge of potential sound correspondences (this is often the case when linguists face the data of a language for the first time) and considering the binary outcome of detecting a sound correspondence, we naturally set $p = 0.5$. In real language data, sound correspondence could be rare, so p will be much smaller. However, mathematically speaking, $p(1-p)$ reaches its maximum value at $p = 0.5$, and the maximum value of $p(1-p)$ leads to the maximum value of n, which makes Equation (8) equally applicable to cases having either many or few sound correspondences. In other words, the value n calculated at $p = 0.5$ is conventional, independent of potential sound correspondences in the exemplar set.

In Equation (8), the conventional size n relies on the total vocabulary size N of a language, and the significance level α and error rate ε of the sampling process. Here, we confine α and ε in a statistically acceptable range between 0.05 and 0.1, and set N within 4000–5000. Corpus linguists have estimated that this amount of words could cover more than 95% of the texts of a language (Laufer and Ravenhorst-Kalovski, 2010) and arguably suffice for basic comprehension (Nation and Warning, 1997).

We use Equation (8) to calculate the conventional size n in eight conditions formed by combinations of different values of ε (0.05 and 0.1), α (0.05 and 0.1), and N (4000 and 5000) (see the gray cells in **Table 1**).

In the strictest condition having the smaller error rate and significance level and the bigger total vocabulary size ($\varepsilon = 0.05$, $\alpha = 0.05$ ($U_{\frac{\alpha}{2}} = 1.96$), $N = 5000$), we have:

$$n = \frac{N\left(U_{\frac{\alpha}{2}}\right)^2 p(1\ -\ p)}{N\varepsilon^2 + \left(U_{\frac{\alpha}{2}}\right)^2 p(1\ -\ p)} = \frac{5000 \times 1.96^2 \times 0.5 \times 0.5}{5000 \times 0.05^2 + 1.96^2 \times 0.5 \times 0.5} = 356.750 \approx 357 \tag{9}$$

In the most relaxed condition [$\varepsilon = 0.1$, $\alpha = 0.1$ ($U_{\frac{\alpha}{2}} = 1.645$), $N = 4000$], we have:

$$n = \frac{N\left(U_{\frac{\alpha}{2}}\right)^2 p(1\ -\ p)}{N\varepsilon^2 + \left(U_{\frac{\alpha}{2}}\right)^2 p(1\ -\ p)} = \frac{4000 \times 1.645^2 \times 0.5 \times 0.5}{4000 \times 0.1^2 + 1.645^2 \times 0.5 \times 0.5} = 66.526 \approx 67 \tag{10}$$

Conventional sizes calculated in the other six conditions all lie within the boundary values specified by Equations (9) and (10). Accordingly, the conventional size of the vocabulary list for assembling potential cognates is within the range [67, 357] (see **Figure 2**).

As shown in **Figure 2** and **Table 1**, in a relaxed condition ($\varepsilon = 0.1$), when N is between 4000 and 5000, both sizes of the Swadesh lists (100 and 200) reach a significance level below 0.05 (α is within [0.02, 0.05] for size 100, and within [0.002, 0.01] for size 200). In other words, the distribution of sound correspondences in these amounts of words reliably reflects the distribution of those correspondences in the languages being compared (reaching a confidence level $(1-\alpha)$ above 95%; in statistical terms, in over 95% cases the distribution of sound correspondences in a sample of 100 or 200 words approximates, within the given error rate, the distribution of those correspondences in the total vocabulary set). These calculations indicate that both Swadesh lists are statistically large enough to estimate language relatedness.

Some linguists (e.g., Embleton, 1986: p. 92–93) advocate using the Swadesh 200-word list rather than the Swadesh 100-word

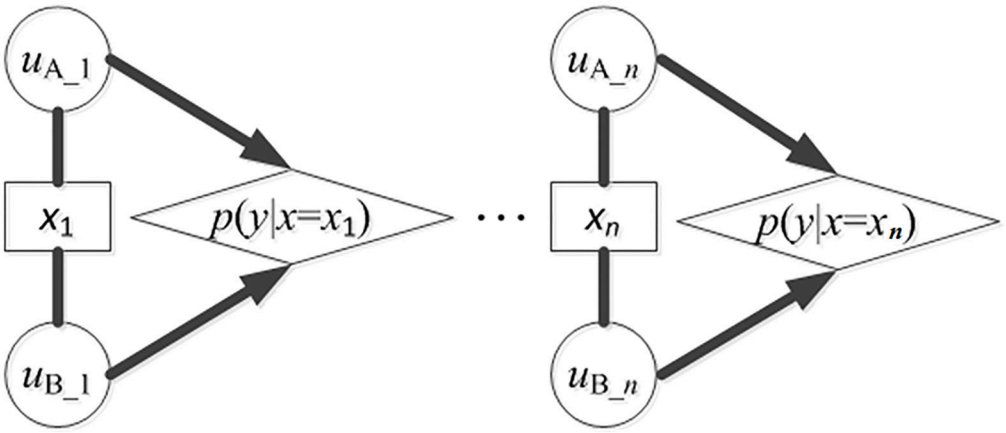

FIGURE 1 | Detection of sound correspondences in assembled words from languages A and B. x_i ($i = 1$ to n) is the concept in the vocabulary list for collecting potential cognates, n is the size of the vocabulary list. u_{A_i} is the word form from A that is semantically equivalent to x_i. u_{B_i} is the word form from B that is semantically equivalent to x_i. $p(y|x = x_i)$ is the probability that some segments in u_{A_i} and u_{B_i} show a correspondence. The detecting process can be conceived as a Bernoulli process. The probabilities of showing correspondences in all exemplars follow a 0–1 distribution, and the probabilities for a particular correspondence to occur different times in all exemplars follow a binomial distribution.

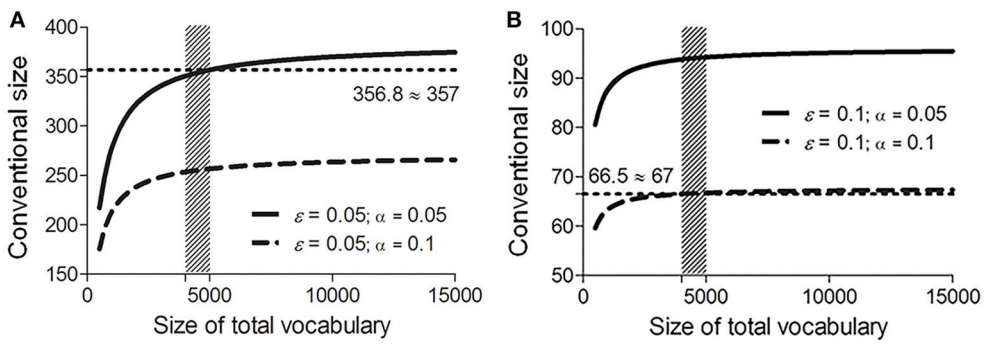

FIGURE 2 | Conventional sizes of the vocabulary list under fixed error rate ε [0.05 (A) and 0.1 **(B)]** and significance level α [0.05 (solid lines) and 0.1 (dash lines)], and various total vocabulary sizes N (500–15000). Shade areas mark the range where N is between 4000 and 5000. Dotted lines mark the range of the conventional size n (round-up to the closest integers) calculated using Equation (8).

list because comparison accuracy decreases considerably when using a 100-word list. Our results suggest that the Swadesh 100-word list and the Swadesh 200-word list are both reliable, since their sizes lie in the conventional size range. Note that this does not mean that the two lists are interchangeable. In fact, they reach different significance levels (α). As shown in **Figure 3**, the Swadesh 200-word list shows a better performance than the Swadesh 100-word list; under the same sampling requirements (predefined significance level and error rate), the significance levels of the Swadesh 200-word list are consistently lower than those of the Swadesh 100-word list. By contrast, the reliabilities of the smaller lists containing 15, 33, 35, or 40 concepts are much lower (see **Figure 3** and **Table 1**, under a total vocabulary of 4000–5000 words, the significance levels of those sizes are above 0.2, and their confidence levels are below 80%). In other words, these lists could not reliably reflect the distribution of sound correspondences in the languages being compared.

In addition, as shown in **Figure 2**, given fixed α and ε, the conventional size n increases steadily with the total vocabulary size N, yet such tendency becomes less explicit under a much bigger N. For example, in the most rigorous condition ($\alpha = 0.05$, $\varepsilon = 0.05$; the solid line in **Figure 2A**), under a total of 15,000 words, the conventional size remains below 400. This indicates that a much bigger vocabulary list containing over 400 concepts offers no additional benefit in reflecting the distributions of sound correspondences in languages being compared. This is consistent with linguistic discussions (e.g., Ringe, 1992; Kessler, 2001). For example, Embleton points out that a word list having more than 500 items does not bring additional advantage in language comparison (Embleton, 1986).

Furthermore, since the above statistical analysis does not consider semantics, in principle, this conventional size range is instructive to cognate collection based on other linguistically acceptable meaning lists. Some linguists advocate using 300 concepts to sample potential cognates in Austronesian or

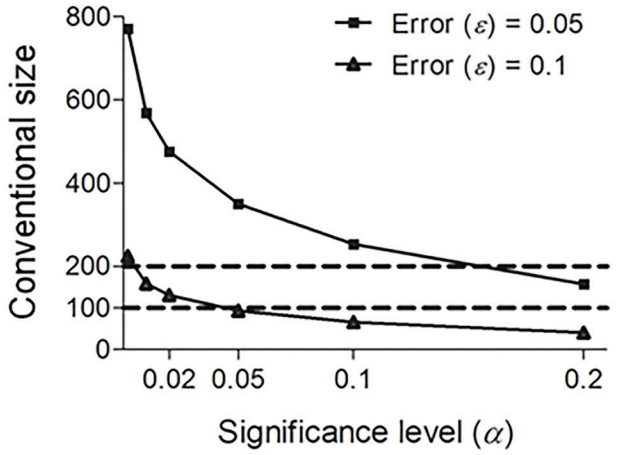

FIGURE 3 | Conventional size *n* under fixed error rates ε (0.05 and 0.1) and total vocabulary size *N* (4000), and various significance levels α.

Sino-Tibetan languages (e.g., Huang, 1997). This is because some concepts in the Swadesh lists (e.g., "bark," "to swim," "to lie," or "because," "in," "at," "with," "if") have no corresponding word forms in some Tibeto-Burman, Miao-Yao, or Zhuang-Dong languages (Jiang, 2007), whereas words encoding other concepts not in the Swadesh lists (e.g., "hemp," "bamboo") are arguably stable and resistant to borrowing (Li, 1995). Apart from these linguistic considerations, our quantitative analyses suggest that at least the size (300) of this vocabulary list is sufficient to collect potential cognates (see **Figure 3** and **Table 1**, the confidence levels of size 300 under a total vocabulary of 4000 or 5000 words are above 90%).

Generality of the Swadesh 100-Word List

After proposing the Swadesh 200-word list in 1952, Swadesh published the 100-word list in 1955. He stressed that the 100-word list contained more stable concepts and the corresponding word forms tended to be less prone to borrowing (though the inherent stabilities of these words may differ, see Tadmor et al., 2010). Leaving aside linguistic considerations, after showing that both Swadesh lists are acceptable for cognate assembly, an immediate next question naturally arises: whether the Swadesh 100-word list is a special sub-list of the Swadesh 200-word list, in terms of the distribution of detected sound correspondences. Answer to this question is informative to cases where certain word forms in the Swadesh 100-word list do not exist or are generally hard to obtain. Such cases are common in reality (cf. Jiang, 2007).

Hypothesis testing in statistical inference serves as a useful means to address this question. In particular, the Ansari-Bradley test (Lunneborg, 2005) gives a 0–1 decision to the null hypothesis that two samples having common medians come from the same distribution. Decision "0" means that the null hypothesis cannot be rejected, and "1" that the null hypothesis can be rejected (that is to say, the two samples do not come from the same distribution) at a predefined significance level (say, 0.05). The Ansari-Bradley test is non-parametric and distribution-free; the samples for comparison do not need to have finite variances, identical sizes, or show normal distributions. Hence, this test is suitable for comparing the distributions of sound correspondences in assembled words, which follow binomial distributions. In addition, the Spearman's rank correlation coefficient (a.k.a Spearman's rho; Kornbrot, 2014) is another non-parametric measure of statistical correlation between two samples. It returns a value within $[-1, 1]$, indicating the degree of negative or positive mutual dependence between any pairs of the data from the two samples. This measure is also suitable for revealing statistical dependence between sets of assembled words.

We discuss the generality of the Swadesh 100-word list based on the word forms assembled, respectively, in English and Latin according to the Swadesh 100- and 200-word lists (the data are extracted from Ringe, 1992; **Table 2**). For the sake of simplicity, we only consider sound correspondences appearing at fixed positions of assembled words. To be specific, we only consider the potential word-initial consonant correspondences shown in the assembled words.

Based on Spearman's rho, we compare the distribution of the sound correspondences in the assembled words following the Swadesh 100-word list with that in the assembled words following the Swadesh 200-word list. Spearman's rho in this test is 0.7384, with $p < 0.0001$. The p value indicates the chance for random sampling to reach the observed correlation if there is really no correlation between the two samples. The high Spearman's rho with an extremely low p-value reveals a high correlation between the distributions of the sound correspondences in the assembled words following the two Swadesh lists. This is also partially observed in **Table 3**. Based on the same setting and the statistical principle, identified recurrent correspondences based on the Swadesh 100-word list are largely consistent to those based on the Swadesh 200-word list.

Based on the Ansari-Bradley test, we compare the distribution of sound correspondences in the Swadesh 100-word list with those in 10,000 100-word lists generated by random sampling from the Swadesh 200-word list. In these 10,000 comparisons, we count the total number of "1" decisions returned by the test. Such number is 11, corresponding to a p-value of $11/10{,}000 \approx 0.0011$. This indicates that in most situations the randomly sampled 100-word lists exhibit a similar distribution to the Swadesh 100-word list. In other words, the Swadesh 100-word list is not statistically distinct from other sub-lists of the same size.

Finally, we extend the above test by considering randomly generated sub-lists ranging in size from 20 to 200, incremented at a step of five words. For each list size n, we first sample a list of size n from the Swadesh-200 list, and then create another 10,000 sub-lists of n words. After that, we compare the distributions of the sound correspondences in the assembled words between the first list and each of the 10,000 sub-lists, and count the total number of "1" decisions returned by the test.

Figure 4 shows that in most randomly-created sub-lists containing 60 or more members the distributions of the sound correspondences are similar (at the predefined significance level of 0.05). This conclusion also holds for the Swadesh 100-word list; the distribution of the sound correspondences in any of the 100-word sub-list is similar.

TABLE 2 | Word-initial consonant correspondences (CCs) between English (left) and Latin (right) (extracted from Ringe, 1992) following the Swadesh 100- and 200-word lists.

Index	Concept	CC	Index	Concept	CC	Index	Concept	CC
1	all(pl.)	Ø-Ø	49	leaf	l-f	99	you(sg.)	y-t
2	ashes	Ø-k	50	lie	l-y	100	yellow	y-f
3	bark[of tree]	b-k	51	liver	l-y	101	and	Ø-Ø
			52	long	l-l	102	animal	Ø-Ø
4	belly	b-w	53	louse	l-p	103	at	Ø-Ø
5	big	b-m	54	man	m-w	104	back[nn]	b-t
6	bird	b-Ø	55	many	m-m	105	bad	b-m
7	bite	b-m	56	moon	m-l	106	because	b-k
8	black	b-Ø	57	mountain	m-m	107	blow[vb, wind]	b-f
9	blood	b-s	58	mouth	m-Ø			
10	bone	b-Ø	59	name	n-n	108	breathe	b-s
11	breast(s)	b-m	60	neck	n-k	109	child	c-p
12	burn[intr]	b-Ø	61	new	n-n	110	count	k-n
13	claw	k-Ø	62	night	n-n	111	cut	k-s
14	cloud	k-n	63	nose	n-n	112	day	d-d
15	cold	k-f	64	not	n-n	113	dig	d-f
16	come	k-w	65	one	w-Ø	114	dirty	d-s
17	die	d-m	66	path	p-s	115	dull	d-h
18	dog	d-k	67	rain[nn]	r-p	116	dust	d-p
19	drink	d-b	68	red	r-r	117	fall	f-k
20	dry	d-s	69	root	r-r	118	far	f-p
21	ear	Ø-Ø	70	round	r-r	119	father	f-p
22	earth	Ø-t	71	sand	s-h	120	few	f-p
23	eat	Ø-Ø	72	say	s-d	121	fight	f-p
24	egg	Ø-Ø	73	see	s-w	122	five	f-k
25	eye	Ø-Ø	74	seed	s-s	123	flow	f-f
26	fat[nn]	f-Ø	75	sit	s-s	124	flower	f-f
27	feather	f-p	76	skin	s-k	125	fog	f-n
28	fire	f-Ø	77	sleep	s-d	126	four	f-k
29	fish	f-p	78	small	s-p	127	freeze	f-g
30	flesh	f-k	79	smoke	s-f	128	fruit	f-p
31	fly[vb]	f-w	80	stand	s-s	129	grass	g-g
32	foot	f-p	81	star	s-s	130	guts	g-Ø
33	full	f-p	82	stone	s-l	131	he	h-Ø
34	give	g-d	83	sun	s-s	132	heavy	h-g
35	good	g-b	84	swim	s-n	133	here	h-h
36	green	g-w	85	tail	s-k	134	hit	h-f
37	hair[of head]	h-k	86	that(nt.)	ð-Ø	135	hold	h-t
			87	this(nt.)	ð-h	136	hunt[vb]	h-w
38	hand	h-m	88	tongue	t-l	137	husband	h-m
39	head	h-k	89	tooth	t-d	138	ice	Ø-g
40	hear	h-Ø	90	tree	t-Ø	139	if	Ø-s
41	heart	h-k	91	two	t-d	140	in	Ø-Ø
42	horn	h-k	92	walk	w-Ø	141	knife	n-k
43	hot	h-k	93	water	w-Ø	142	lake	l-l
44	human[nn]	h-h	94	we	w-n	143	laugh	l-r
45	I	Ø-Ø	95	what	w-k	144	left[-hand]	l-s
46	kill	k-Ø	96	white	w-Ø	145	mother	m-m
47	knee	n-g	97	who	h-k	146	narrow	n-Ø
48	know	n-s	98	woman	w-m	147	near	n-p
148	now	n-n	165	sky	s-k	183	think	θ-k
149	old	Ø-w	166	smell[tr]	s-Ø	184	three	θ-t
150	other	Ø-Ø	167	smooth	s-l	185	throw	θ-y
151	play	p-l	168	snake	s-Ø	186	tie	t-l
152	pull	p-t	169	snow	s-n	187	true	t-w
153	push	p-t	170	some(pl.)	s-Ø	188	vomit	v-w
154	right[-hand]	r-d	171	spit	s-s	189	wash	w-l
			172	split	s-f	190	wet	w-Ø
155	river	r-f	173	squeeze	s-p	191	wide	w-l
156	rotten	r-p	174	stab	s-f	192	wife	w-Ø
157	rub	r-f	175	stick[nn]	s-b	193	wind[nn]	w-w
158	salt	s-s	176	straight	s-r	194	wing	w-Ø
159	scratch	s-s	177	suck	s-s	195	wipe	w-t
160	sea	s-m	178	swell	s-t	196	with	w-k
161	sew	s-s	179	there	ð-Ø	197	woods	w-s
162	sharp	s-Ø	180	they	ð-Ø	198	worm	w-w
163	short	s-b	181	thick	θ-k	199	you(pl.)	y-w
164	sing	s-k	182	thin	θ-t	200	year	y-Ø

(Continued)

TABLE 2 | Continued

The first 100 concepts are from the Swadesh 100-word list. "Ø" denotes zero consonant.

TABLE 3 | Potential and recurrent word-initial consonant correspondences (CC) in the assembled words from English (left) and Latin (right) following the Swadesh 100- and 200-word lists (extracted from Ringe, 1992).

α	Swadesh list	Potential CC	Recurrent CC	CC items
0.01	100	62	6	Ø-Ø, f-p, h-k, l-y, n-n, r-r
	200	108	7	Ø-Ø, f-p, m-m, l-y, n-n, r-r, s-s
0.05	100	62	10	Ø-Ø, b-m, f-p, h-k, l-y, n-n, r-r, s-s, t-d, y-t
	200	108	15	Ø-Ø, b-m, f-p, h-k, l-y, m-m, n-n, p-t, r-r, s-s, ˙s-b, ð-Ø, θ-t, t-l, t-d

"Ø" denotes zero consonant.

These results indicate that although the meanings in the Swadesh 100-word list are carefully chosen according to linguistic considerations, the distribution of correspondences in the assembled words following this list is not significantly distinct from that following the Swadesh 200-word list or any 100-word sub-list of the Swadesh 200-word list. In other words, the Swadesh 100-word list is not statistically special.

The above tests verify the practice of using the Swadesh 100-word list in language comparison. Although some correspondence(s) may not exist in the assembled words following this list (see **Table 3**, which shows the potential and recurrent sound correspondences identified based on our second principle discussed in the next section), the findings remain statistically reliable. In addition, these tests also support the assumption that the semantics-phonetics mappings are largely arbitrary, such that word sampling is less confined by meanings. Furthermore, these tests give fieldwork linguists freedom to replace concepts in the Swadesh 100-word list that have no

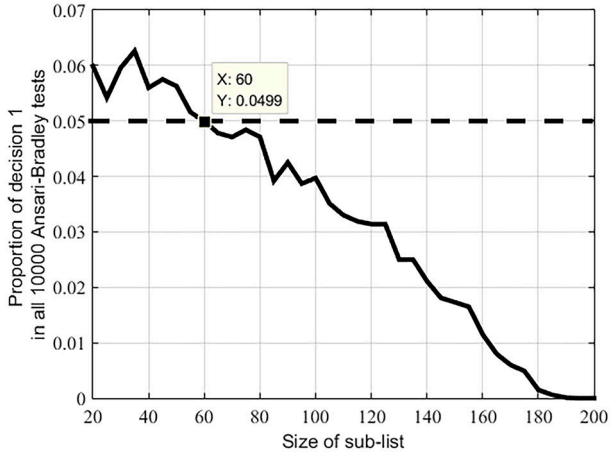

FIGURE 4 | Proportions of decision 1 in 10000 Ansari-Bradley tests under different sizes of sub-lists of the Swadesh 200-word list. Dotted line indicates the significance level 0.05.

word forms with those in the Swadesh 200-word list (or other linguistically-acceptable list) that have word forms, and still obtain statistically similar results.

Dynamic Threshold of Recurrent Sound Correspondence

As proved in Section Conventional Size of the Vocabulary List to Assemble Potential Cognates, if words for comparison are gathered following a vocabulary list having a conventional size, language affinity can be measured based on the number of detected recurrent sound correspondences in the collected words. Here, we revisit the statistical permutation principle (Ringe, 1992) to calculate the threshold for identifying recurrent sound correspondences among potential correspondences detected in assembled words. This principle is derived from the previous estimations of the degree of similarity that two (or more) languages are expected to show by chance (Bender, 1969; Oswalt, 1971; Ringe, 1993; Kessler, 2001). The logic behind the principle is as follows. If (1) a sound correspondence occurs in a few pairs of assembled words with semantic equivalence and (2) the accidental probability for such correspondence to randomly occur these many times is considerably low, the correspondence can be identified as a recurrent one. In addition, the minimum number of occurrences that keeps the accidental probability below a predefined significance level α can be assigned as the threshold for identification of recurrent correspondence.

Revisiting **Figure 1**, in the first step of lexicostatistics, the main concern is whether there are sound correspondences in the exemplars. Accordingly, we adopt the statistics of binomial distribution to calculate the conventional number of exemplars. In the second step, the focus shifts to counting the occurrences of a particular sound correspondence in the exemplars. In a Bernoulli process, the probability distribution of the number of occurrences follows a binomial distribution. Thus, based on the statistics of binomial distribution and the observed frequencies of relevant segments in the assembled words, we can calculate:

(1) Accidental probability of a sound correspondence randomly occurring n times in the assembled words;
(2) Minimum number of occurrence to keep the accidental probability of the sound correspondence below a considerably low level.

For two-language comparison, each occurrence of a sound correspondence (say, segment a in language A corresponds to segment b in language B) in the assembled words results from a combined Bernoulli process. The accidental probability P for such correspondence to occur randomly at least k times in the aligned words can be calculated as in Equation (11):

$$P\left(X \geq k\right) = 1 - \sum_{i=0}^{k-1} \binom{n}{i} p^i (1 - p)^{n-i} \tag{11}$$

Here, n is the number of assembled words, identical to the size of the vocabulary list used for collecting cognates, and p is the combined occurring frequency of the segments involved in this correspondence.

Without prior knowledge of the relations between languages being compared and considering the arbitrariness in mappings between semantics and phonetics, we can safely assume that the occurring frequencies of segments in these languages are independent, at least in the assembled words. Thus, the combined occurring frequency p of the segments in the assembled words can be calculated as the product of the occurring frequency of segment a in the assembled words from A (p_A) and that of segment b in the assembled words from B (p_B):

$$p = p_A p_B \tag{12}$$

Note, that if there is a large degree of borrowing between the two languages, a dependent correlation between the occurring frequencies of segments in these languages can be easily observed.

Figure 5 traces the step-wise thresholds calculated by Equation (11) for two-language comparison using a Swadesh 100-word list. As shown in **Figure 5**, the smallest acceptable threshold is two (based on two matching instances). Compared to the other threshold values, the smallest threshold remains valid for the widest range of combined occurring frequency [0.0002, 0.0015). This indicates that the threshold two can reliably identify recurrent sound correspondences, provided that the combined occurring frequencies of these correspondences fall in this wide range. In addition, our principle suggests that the approach of using a fixed threshold throughout is not recommended. The threshold value must increase dynamically along with increase in combined occurring frequency.

Our principle (Equations 11, 12) enables an automatic assignment of threshold of recurrent sound correspondence. For example, if the combined occurring frequency of two segments falls in the range [0.0015, 0.0043), the threshold is set to three (according to **Figure 5**). Accordingly, if there are three or more pairs of the assembled words that recurrently show correspondence between the two segments, we can classify this correspondence as a recurrent one and those pairs of assembled words as cognates.

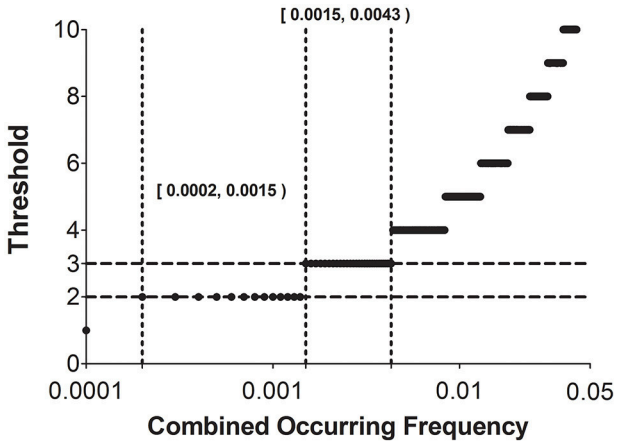

FIGURE 5 | Thresholds of recurrent sound correspondences under different combined occurring frequencies of involved sound segments (0.0001–0.05, with a step 0.0001). Two regions marked by dotted lines [0.0002, 0.0015) and [0.0015, 0.0043), respectively, denote the ranges of combined occurring frequencies where the threshold can be set to two and three.

We use two examples that are much simpler than those in real cases to illustrate how to apply this principle. Using such simple examples avoids unnecessary distractions from linguistic aspects.

In the first example, the empirical data are word forms assembled, respectively, from English and French following the Swadesh 100-word list (extracted from Ringe, 1993; see **Table 4**). We apply our principle to calculate the accidental probability of a word-initial consonant correspondence /f/-/p/ in the assembled words.

In this example, p_A is the occurring frequency of segment a (/f/) in the assembled words from English, and p_B is that of segment b (/p/) in the assembled words from French. In the assembled words from English, the word-initial consonant /f/ appears ten times; in those from French, the word-initial consonant /p/ appears six times. Following Equation (12), the combined occurring frequency p of the two segments is 0.006(=0.1×0.06). Following Equation (11), the accidental probability P for the correspondence /f/-/p/ to occur randomly at least four times is:

$$\begin{aligned} P(X \geq 4) &= 1 - \sum_{i=0}^{3} \binom{100}{i} 0.006^i (1-0.006)^{100-i} \\ &= 1 - \binom{100}{0} 0.006^0 (1-0.006)^{100} \\ &\quad - \binom{100}{1} 0.006^1 (1-0.006)^{99} \\ &\quad - \binom{100}{2} 0.006^2 (1-0.006)^{98} \\ &\quad - \binom{100}{3} 0.006^3 (1-0.006)^{97} \\ &= 0.0032 \end{aligned} \tag{13}$$

Statistically speaking, the accidental probability 0.0032 is much lower than the general significance level α (0.01 or 0.05), and among 100 pairs of the assembled words, there are four pairs that exhibit such correspondence (i.e., father vs. *père*, fish vs. *poisson*, foot vs. *pied*, and full vs. *plein*). Therefore, we can safely claim that the correspondence /f/-/p/ is recurrent and those four pairs of words are cognates. Similarly, as shown in **Figure 5**, since its combined occurring frequency is 0.006, the threshold for classifying it as a recurrent one should be four.

In the second example, the empirical data are word forms assembled, respectively, in Latin and English according to the Swadesh 100− and 200-word lists (extracted from Ringe, 1992; see **Table 2**). We use our principle to evaluate all detected word-initial consonant correspondences (see **Table 3**), under two significance levels, 0.01 and 0.05.

At the significance level 0.01, in the words assembled following the Swadesh 100-word list, we detect 62 correspondences. Based on the occurring frequencies of related consonants in the assembled words and **Figure 5**, our

TABLE 4 | Word-initial consonant correspondences (CCs) between English (left) and French (right) (extracted from Ringe, 1993).

Index	Concept	CC	Index	Concept	CC	Index	Concept	CC
1	all(pl.)	Ø-t	34	good	g-b	67	path	p-s
2	ashes	Ø-s	35	grease	g-g	68	rain[nn]	r-p
3	bark[of tree]	b-Ø	36	green	g-v	69	red	r-r
4	belly	b-v	37	hair[of head]	h-š	70	root	r-r
5	big	b-g	38	hand	h-m	71	round	r-r
6	bird	b-Ø	39	head	h-t	72	sand	s-s
7	bite	b-m	40	hear	h-Ø	73	say	s-d
8	black	b-n	41	heart	h-k	74	see	s-v
9	blood	b-s	42	horn	h-k	75	seed	s-g
10	bone	b-Ø	43	hot	h-š	76	sit	s-Ø
11	breast(s)	b-s	44	human[nn]	h-p	77	skin	s-p
12	burn[intr]	b-b	45	I	Ø-m	78	sleep	s-d
13	claw	k-g	46	kill	k-m	79	small	s-p
14	cloud	k-n	47	knee	n-ž	80	smoke	s-f
15	cold	k-f	48	know	n-s	81	stand	s-d
16	come	k-v	49	leaf	l-f	82	star	s-Ø
17	die	d-m	50	lie	l-Ø	83	stone	s-p
18	dog	d-š	51	liver	l-f	84	sun	s-s
19	drink	d-b	52	long	l-l	85	swim	s-n
20	dry	d-s	53	louse	l-p	86	tail	s-k
21	ear	Ø-Ø	54	man	m-Ø	87	that(nt.)	ð-s
22	earth	Ø-t	55	many	m-b	88	this(nt.)	ð-s
23	eat	Ø-m	56	meat	m-v	89	tongue	t-l
24	egg	Ø-Ø	57	moon	m-l	90	tooth	t-d
25	eye	Ø-Ø	58	mountain	m-m	91	tree	t-Ø
26	feather	f-p	59	mouth	m-b	92	two	t-d
27	fire	f-f	60	name	n-n	93	water	w-Ø
28	fish	f-p	61	neck	n-k	94	we	w-n
29	fly[vb]	f-v	62	new	n-n	95	what	w-k
30	foot	f-p	63	night	n-n	96	white	w-b
31	full	f-p	64	nose	n-n	97	who	h-k
32	give	g-d	65	not	n-n	98	woman	w-f
33	go	g-Ø	66	one	w-Ø	99	you(sg.)	y-t
						100	yellow	y-ž

"Ø" denotes zero consonant.

principle identifies six recurrent correspondences, matching exactly those identified by Ringe (1992). At the same significance level, in the words assembled following the Swadesh 200-word list, we detect 108 correspondences. Based on the occurring frequencies of related consonants and Equations (11) (set $n = 200$) and (12), our principle identifies seven recurrent correspondences. There are differences between the two sets of recurrent correspondences obtained following the two lists: /h/-/k/ in the Swadesh 100-word list, whereas /m/-/m/ and /s/-/s/ in the Swadesh 200-word list. At the significance level 0.05, our principle identifies more recurrent correspondences (10 in the Swadesh 100-word list and 15 in the Swadesh 200-word list).

The above examples were first used by Ringe (1992, 1993), who attempted to show that no matter what size the vocabulary list has the relative frequency between related and unrelated words remained the same and the numbers of matches expected by chance were proportional to the number of words (Kessler, 2001). Our results demonstrate two things: (1) under the same vocabulary list, different significance levels lead to identification of different sets of recurrent sound correspondences and (2) under the same significance level, size differences in vocabulary lists also lead to differential identification outcomes. These results suggest that our principle is dependent on a number of numerical parameters, including the size of the vocabulary list for word assembly, the occurring frequencies of related sound segments in the assembled words, and the predefined significance level.

In addition, if one sticks to the smallest threshold (two), some correspondences would be incorrectly classified as recurrent. For example, in the assembled words following either list, the correspondences /b/-Ø and /f/-Ø occur four and two times, respectively. Based on the threshold two, both correspondences would be deemed as recurrent. However, linguists have proved that neither of them is recurrent. By contrast, according to our principle, due to their high combined occurring frequencies, the threshold of these correspondences should be set as a value much bigger than their occurring numbers in the assembled words. Then, in line with linguistic considerations, both correspondences are not judged as recurrent.

DISCUSSION

In lexicostatistics, linguistic intuitions and subjective experiences have been the primary factors determining (1) which words should be collected for comparison, (2) how many words are needed for comparison, and (3) whether a specific number of matching instances is sufficient to confirm a "recurrent" correspondence for identifying cognates (Hock and Joseph, 1996; Baxter and Ramer, 2000; Brown et al., 2013). We use statistical principles to independently validate the reliability and generality of the results pulled from the commonly used Swadesh 100- and 200-word lists. We also use statistical theorems to provide objective answers to questions (2) and (3): we propose a method to quantify the conventional size of vocabulary lists for word assembly, as well as a method to assign reasonable thresholds for detecting recurrent sound correspondences. Our results show that (1) the widely-adopted practice of using 100 or 200 words following the Swadesh lists for cognate assembly can render reliable comparison; (2) the Swadesh 100-word list is statistically invariant to the Swadesh 200-word list and other sub-lists having comparable sizes; and (3) the threshold based on at least two matching instances remain valid for a wide range of cases, yet such threshold must increase dynamically with increase in combined occurring frequencies of relevant sound segments.

Our study reveals that in addition to linguistic considerations, statistical criteria (e.g., significance level and error rate) are also critical for comparison outcome. Based on different sampling requirements, the same datasets may render distinct results. Therefore, respecting and applying statistical criteria is necessary and beneficial to verify, replicate, and discuss language comparison studies based on the same datasets. As social scientists, linguists generally receive less mathematical training in developing or evaluating statistical approaches (Baxter and Ramer, 2000). Nonetheless, their linguistic intuitions and rich experiences turn out to be reliable to a certain extent to bring forth informative understanding about historical relations among languages. With more and more large-scale datasets being available, our study can guide future research based on such datasets.

Identifying recurrent sound correspondence is an important step not only in lexicostatistics to detect cognates, but also in comparative method to reconstruct linguistic features of protolanguage. Among the available probability approaches adopted to do this task, such as Chi-square calculations (Ross, 1950; Kessler, 2001), Binomial approach (Ringe, 1992) and Shift test (Oswalt, 1971), our statistical principle follows Ringe's Binomial approach. This approach explicitly assigns the meaning to the parameter p (probability for two segments to exhibit a sound correspondence). In addition, the Binomial approach can also be efficiently applied in the cases where there are very low expected numbers in the slots of the table for comparison. In such cases, Chi-Squared test cannot be used (McMahon and McMahon, 2005). Note that despite of its advantages, some linguists criticize that the Binomial approach could be too rigorous to confirm close relationship within Indo-European languages (Greenberg, 1993: p. 89). Furthermore, the rationale of our method is similar to that of Bayesian approach. In principle, Bayesian approach is to estimate a posterior probability with respect to a prior probability based on the given data. It depends excessively on the prior distribution, and determination of this distribution is subject to the intuition and experience of researchers. In our study, the frequencies of the segments occurring in a given word list can be treated as the prior distribution. Whether a correspondence is recurrent or not cannot be simply determined by the occurrence of such correspondence. Instead, it needs to take into consideration not only the number of occurrence of the correspondence but also the number of occurrence of the segments involved in the correspondence, the latter of which could be biased due to word collection.

Mathematicians and statistical physicists have developed powerful approaches, many of which have potential applications in linguistics. Our study attempts to bridge the gap between linguistics and other disciplines, by exemplifying how to

employ statistical knowledge and approaches to tackle linguistic issues. There have been more and more attempts like this (e.g., Bouchard-Côté et al., 2013; Hruschka et al., 2015). Integration of multidisciplinary approaches has become imperative to evaluate data collection and classification methods widely-adopted in linguistics and other social science disciplines.

AUTHOR CONTRIBUTIONS

MZ and TG designed the research, MZ carried out the study. MZ and TG analyzed the results. MZ and TG wrote the paper.

ACKNOWLEDGMENTS

MZ is supported by the projects from Postdoctoral Science Foundation of China (2015M570316) and National Natural Science Foundation of China (31501010). TG is supported by the US NIH Grant (HD-071988). The research has been supported in part by the MOE Project of the Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies. The authors thank Li Jin and Wuyun Pan from Fudan University, Ken Pugh and Dave Braze from Haskins Laboratories, and Yicheng Wu from Zhejiang University for support and comments, and Dave Kush from Haskins Laboratories for proofreading.

REFERENCES

Baxter, W. H., and Ramer, A. M. (2000). "Beyond lumping and splitting: probabilistic issues in historical linguistics," in *Time Depth in Historical Linguistics*, Vol. 1, eds C. Renfrew, A. McMahon, and L. Trask (Cambridge: McDonald Institute for Archaeological Research), 167–188.

Bender, M. L. (1969). Chance CVC correspondences in unrelated languages. *Language* 45, 519–536. doi: 10.2307/411436

Bergsland, K., and Vogt, H. (1962). On the validity of glottochronology. *Curr. Anthrop.* 3, 115–153. doi: 10.1086/200264

Bouchard-Côté, A., Hall, D., Griffiths, T. L., and Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proc. Natl. Acad. Sci. U.S.A.* 110, 4224–4229. doi: 10.1073/pnas.1204678110

Brown, C. H., Holman, E. W., and Wichmann, S. (2013). Sound correspondences in the World's languages. *Language* 89, 4–29. doi: 10.1353/lan.2013.0009

Campbell, L. (2013). *Historical Linguistics*. Edinburgh: Edinburgh University Press.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

De Saussure, F. (1983). *Course in General Linguistics*, Edited by C. Bally and A. Sechehaye. La Salle, IL: Open Court.

Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Lang. Ling. Compass* 6, 654–672. doi: 10.1002/lnc3.361

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., and Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends Cogn. Sci.* 19, 603–615. doi: 10.1016/j.tics.2015.07.013

Dolgopolsky, A. B. (1986). "A probabilistic hypothesis concerning the oldest relationships among the language families in northern Eurasia," in *Typology, Relationship and Time: A Collection of Papers on Language Change and Relationship*, eds V. V. Shevoroshkin and T. L. Markey (Ann Arbor, MI: K*aroma)*, 27–50.

Dyen, I., Kruskal, J., and Black, P. (1997). *FILE IE-DATA1*. Available online at: http://www.wordgumbo.com/ie/cmp/iedata.txt (Accessed N*ovember* 30, 2015).

Embleton, S. M. (1986). *Statistics in Historical Linguistics*. Bochum: Brockmeyer.

Eska, J. F., and Ringe, D. (2004). Recent work in computational linguistic phylogeny. *Language* 80, 569–582. doi: 10.1353/lan.2004.0123

Greenberg, J. H. (1990). The American Indian language controversy. *Rev. Archaeol.* 11, 5–14.

Greenberg, J. H. (1993). Observations concerning Ringe's calculating the factor of chance in language comparison. *Proc. Am. Philos. Soc.* 137, 79–90.

Greenhill, S. J., Blust, R., and Gray, R. D. (2008). The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evol. Bioinform.* 4, 271–283.

Heggarty, P. (2010). Beyond lexicostatistics: how to get more out of 'word list' comparisons. *Diachronica* 27, 301–324. doi: 10.1075/dia.27.2.07heg

Hock, H. H., and Joseph, B. D. (1996). *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics*. Berlin: Mouton de Gruyter.

Hockett, C. F. (1960). The origin of speech. *Sci. Am.* 203, 89–97. doi: 10.1038/scientificamerican0960-88

Hoijer, H. (1956). Lexicostatistics: a critique. *Language* 32, 49–60. doi: 10.2307/410652

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., and Bakker, D. (2008). Explorations in automated language classification. *Folia Linguist.* 42, 331–354. doi: 10.1515/FLIN.2008.331

Hruschka, D. J., Branford, S., Smith, E. D., Wilkins, J., Meade, A., and Pagel, M. (2015). Detecting regular sound changes in linguistics as events of concerted evolution. *Curr. Biol.* 25, R41–R43. doi: 10.1016/j.cub.2014.10.064

Huang, B. F. (1997). Range and standard for defining comparative cognate lists: an example for Sino-Tibetan languages. *Min. Lang. China* 4, 10–16.

Hurford, J. R. (2012). *The Origins of Grammar*. Oxford: Oxford University Press.

Hymes, D. H. (1960). Lexicostatistics so far. *Curr. Anthropol.* 1, 3–44. doi: 10.1086/200074

Jiang, D. (2007). *On Evolutionary Models of Sound Changes for Sino-Tibetan Languages: Theores and Methods of Historical Linguistics.* Beijing: Social Science Academy Press.

Kessler, B. (2001). *The Significance of Word Lists: Statistical Tests for Investigating Historical Connections between Languages*. Stanford, CA: CSLI Publications.

Kornbrot, D. (2014). "Spearman's Rho," in *Wiley StatsRef: Statistics Reference Online*, eds N. Balakrishnan, T. Colton, B. Everitt, G. Molenberghs, W. Piegorsch, and F. Ruggeri (New York, NY: John Wiley & Sons). Available online at: http://onlinelibrary.wiley.com/book/10.1002/9781118445112

Laufer, B., and Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Read. Foreign Lang.* 22, 15–30.

Li, P. J. K. (1995). "Is Chinese genetically related to Austronesian?" in *The Ancestry of the Chinese Language*, ed W. S-Y. Wang (Berkeley, CA: Project on Linguistics Analysis), 93–112.

Lohr, M. (2000). "New approaches to lexicostatistics and glottochronology," in *Time Depth in Historical Linguistics*, Vol. 1, eds C. Renfrew, A. McMahon A, and L. Trask (Cambridge: McDonald Institute for Archaeological Research), 209–223.

Lunneborg, C. E. (2005). "Ansari–Bradley Test," in *Encyclopedia of Statistics in Behavioral Science*, eds B. Everitt, and D. C. Howell (New York, NY: John Wiley & Sons), 93–94.

McMahon, A., and McMahon, R. (2005). *Language Classification by Numbers*. Oxford: Oxford University Press.

McMahon, A., and McMahon, R. (2006). "Why linguists don't do dates: evidence from Indo-European and Australian languages," in *Phylogenetic Methods and the Prehistory of Languages*, eds P. Forster and C. Renfrew (Cambridge: McDonald Institute for Archaeological Research), 153–160.

Nation, P., and Warning, R. (1997). "Vocabulary size, text coverage and word lists," in *Vocabulary: Description, Acquisition and Pedagogy*, eds N. Schmitt, and M. McCarthy (Cambridge: Cambridge University Press), 6–19.

Newman, P. (1995). *On Being Right: Greenberg's African Linguistic Classification and the Methodological Principles Which Underlie It*. Bloomington, IN: Institute for the Study of Nigerian Languages and Cultures, Indiana University.

Oswalt, R. L. (1971). Towards the construction of a standard lexicostatistic list. *Anthropol. Linguist.* 13, 421–434.

Ringe, D. R. (1992). On calculating the factor of chance in language comparison. *Trans. Am. Philos. Soc.* 82, 1–110. doi: 10.2307/1006563

Ringe, D. R. (1993). A reply to Professor Greenberg. *Proc. Am. Philos. Soc.* 137, 91–109.

Ross, A. S. (1950). Philological probability problems. *J. R. Stat. Soc. B*, 19–59.

Ruhlen, M. (1994). *The Origin of Language: Tracing the Evolution of the Mother Tongue*. New York, NY: John Wiley & Sons.

Starostin, S. A. (2000). "Comparative-historical linguistics and lexicostatistics," in *Time Depth in Historical Linguistics*, Vol. 1, eds C. Renfrew, A. McMahon, and L. Trask (Cambridge: McDonald Institute for Archaeological Research), 223–266.

Swadesh, M. (1952). Lexicostatistic dating of prehistoric ethnic contacts. *Proc. Am. Philos. Soc.* 96, 452–463.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Linguist.* 21, 121–137. doi: 10.1086/464321

Tadmor, U., Haspelmath, M., and Taylor, B. (2010). Borrowability and the notion of basic vocabulary. *Diachronica* 27, 226–246. doi: 10.1075/dia.27.2.04tad

Walpole, R. E., Myers, R. H., Myers, S. L., and Ye, K. (2011). *Probability and Statistics for Engineering and Scientists*. New York, NY: Prentice Hall.

Wichmann, S., Müller, A., Wett, A., Velupillai, V., Bischoffberger, J., Brown, C. H., et al. (2013). *The ASJP Database (ver. 16).* Available online at: http://asjp.clld.org/ (Accessed January 1, 2016).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.