Comparative Reconstruction Probabilistically: The Role of Inventory and Phonotactics

Andrei Munteanu
University of Toronto

1 Introduction

As the primary method for establishing language kinship, the comparative method has unveiled evidence of many languages that are not preserved in the physical record, such as proto-Indo-European and proto-Austronesian. However, it has not all been plain sailing in historical linguistics. Some proposals boasting a sizeable following in the past are now met with hesitation if not derision. Perhaps the most famous such example is that of proto-Altaic (Ramstedt, 1957; Poppe, 1960), which subsumes proto-Uralic, proto-Turkic, proto-Mongolic, proto-Tungusic, as well as Korean, and later Japonic (Starostin *et al.*, 2003; Robeets, 2005), with some even observing Mayan 'cognates' (Wikander, 1967). The proposal was met with scathing criticism (Doerfer, 1963; Vovin, 2005; Vovin, 2009), and scholarly opinion appeared to be split during the second half of the 20th century (Georg *et al.*, 1999). In the 21st century, it appears to be that while "negative criticism has been very influential, leading almost to a consensus that no Altaic language family exists, supporters of the Ramstedt-Poppe theory have by no means disappeared" (Norman, 2009).

Yet the Altaic hypothesis is not the most ambitious project in historical reconstruction. The Nostratic family (Illic-Svityc, 1963), originally proposed by Pedersen (1903), contains Altaic as a subbranch, along with Indo-European, Dravidian, Kartvelian, and Afro-Asiatic. Although most linguists tend to dismiss the idea altogether (Campbell, 1998; Ringe, 1998), publications arguing or assuming its validity have not subsided (Bomhard, 2008; Shields, 2011; Dolgopolsky, 2012). Hypotheses garnering even less support can also be found, from the most grandiose, such as proto-World or proto-Human (Ruhlen, 1994), to the relatively peripheral, such as Alarodian (Diakonoff & Starostin, 1986).

On the other end of the spectrum, not all 'mainstream' reconstructions are uncontroversial. For example, the membership of the Omotic in Afroasiatic has been called into question (Theil, 2012), while a comparison of two different Afroasiatic reconstructions revealed only 6% overlap (Ratcliffe, 2003). Likewise, the features shared between members of the Pama-Nyungan language family in Australia have been attributed to diffusion rather than genetic relatedness by some (Dixon & Aikhenvald, 2006).

Finally, for some language families, opinions are split more or less evenly. The Dene-Yeniseian hypothesis (Vajda, 2011; Vajda, 2016), linking the Yeniseian languages of the Old World with the Dene languages of the New World, appears to be one of these. The hypothesis is endorsed by some linguists (Kiparsky, 2014) while contested by others (Campbell, 2011; Starostin, 2012), while others still (Dunn, 2012), including the project's author (Vajda, 2011), recognize that more work is to be done before a conclusion can be reached. Other recent proposals, such as Totozoquean (Brown *et al.*, 2011), which links the Totonocan and Mixe-Zoquean language families of Mexico and later Chitimacha (Brown *et al.*, 2014), a language isolate in the US, have also not been evaluated to conclusion by the community.

The uncertainty regarding the validity of even the most rudimentary linguistic reconstructions stems from the fact that historical linguistics, while in possession of a universally accepted and rigorous methodology in the comparative method, lacks a universally accepted metric for evaluating its applications.

* I would like to thank my supervisor Peter Jurgec, as well as my thesis committee members Nathan Sanders and Barend Beekhuizen for their helpful feedback throughout this project. I would also like to thank members of the University of Toronto Phon Group and the audience at the 28th Manchester Phonology Meeting, who heard presentations overlapping in content.

Proceedings of *AMP 2021*

Although some probability-based arguments in historical linguistics have been made in the past (see Ringe, 1999 for a critique of mass comparison; Nichols & Peterson, 1996 for a critical analysis of the Amerind hypothesis), they do not engage with comparative reconstruction directly and vary widely in methods, scope, focus and generalizability. Current quantitative methods do not make provisions for sound change. Instead, these are usually heuristics for computing language similarity, and not evaluations of language kinship. There are two common quantitative alternatives to comparative reconstruction, lexicostatistics and multilateral comparison, which were both developed in the mid 20th century and inspired most subsequent approaches.

Lexicostatistics (Swadesh, 1955) is a method of comparative analysis in historical linguistics where the proportion of shared cognates is used as a stand-in for genetic proximity. The additional assumption that genetic proximity, as calculated by the lexicostatistic method, is correlated with the time-depth of the proto-language split is known as glottochronology. In practice, the two notions are closely intertwined, and early works on the topic (Swadesh, 1952; Swadesh 1955) suggest that the primary purpose of lexicostatistics was not to replace comparative reconstruction but instead to equip it with a tool for estimating time depth. In fact, in its original instantiation lexicostatistics presupposes the existence of a comparative reconstruction, as this is the implied tool for detecting cognacy (Swadesh, 1955:122).

Since its inception, lexicostatistic-like methods have been used not only to date language divergence (Rexova *et al.*, 2002; Gray & Atkinson, 2003; Chang *et al.*, 2015), but also to argue for the legitimacy of language families and for internal subgroupings (Miller, 1984; Sicoli & Holton, 2014). Cognacy is often taken directly from existing comparative reconstructions (Bowern & Atkinson, 2012; Chang *et al.*, 2015; Greenhill *et al.*, 2017) or determined heuristically through string similarity (Zhang & Gong, 2016). Most commonly, wordlists of 100-200 core vocabulary words are used, a sample which has been shown to sufficiently reflect a language's phonological patterns (Zhang & Gong, 2016). Technically, lexicostatistics is a misnomer, as the method does not require lexical comparison. To perform lexicostatistics, one simply needs to tally the presence or absence of elements across two or more languages. Core vocabulary does easily lend itself to this methodology, but so do typological features (Sicoli & Holton, 2014), or even recurrent sound correspondences (Hruschka *et al.*, 2015). Nevertheless, core vocabulary is particularly useful for this task as it has been shown to exhibit a slower rate of replacement than abstract phonological or syntactic features (Greenhill *et al.*, 2017).

Multilateral comparison, also known as mass comparison (Greenberg, 1987), is the other main method of comparative analysis in historical linguistics. In this method, the similarity of each putative cognate pair, or cognate set for comparisons of more than two languages, is calculated based on a phonological distance metric. The overall similarity score between word pairs in two wordlists is then used as a stand-in for genetic proximity. Greenberg himself stressed that the purpose of the methodology was not to replace the comparative method but to determine which languages to apply it to.

Strictly speaking, multilateral comparison need not be quantitative, and it is possible to simply assess cross language phonological similarity 'intuitively' (Greenberg, 1987; Ruhlen, 1994). In fact, many of the quantitative metrics in multilateral comparison have been first proposed by its critics (Nichols & Peterson, 1992; Baxter, 1995; Ringe, 1999: Kessler & Lehtonen, 2006) in an effort to impart onto it some mathematical rigour. At the heart of multilateral comparison lies the particular phonological similarity metric employed. There is no consensus on what constitutes the best method for phonological string comparison, with some suggesting multivalued articulatory phonetic features (Kondrak, 2003) and others employing basic edit distance (Holman *et al.*, 2011).

Despite the existence of alternatives, most researchers agree that the comparative method constitutes the gold standard when it comes to historical linguistic methodologies (McMahon & McMahon, 2003; Bostoen, 2007; Downey *et al.*, 2008; Kiparsky, 2015). Yet a manually assembled comparative reconstruction is not always tractable, and computational substitutes must occasionally be employed instead, as in the case of the Niger-Congo languages, for example (Rexova *et al.*, 2006). However, in cases when traditional and computational methods are applied in parallel, it is not clear how to interpret the results. For example, a Bayesian tree analysis of typological features in Dene-Yeniseian languages shows that typological diversity is greater within the Dene languages than between the Yeniseian languages and certain branches of the Dene languages, which the authors interpret as evidence for migration out of Beringia into Asia and not from Asia into North America (Sicoli & Holton, 2014). The original Dene-Yeniseian proposal (Vajda, 2011), supported

by the comparative method and genetic studies (Rubicz *et al.*, 2002), came to the opposite conclusion, and the idea of a migration into Asia has since been rejected by its author (Vajda, 2016). There currently exists no way to reconcile the two methodologies and, therefore, their respective conclusions.

This paper will attempt to fill this gap in the literature by presenting a method for evaluating the comparative reconstruction, rather than estimating between-language similarity. The comparative reconstruction itself, as an exhaustive series of changes, acts as the input to the model; the output is the probability that a randomly chosen wordlist would merit the same number and type of changes. In addition to bridging the gap between traditional and quantitative methodologies, the probabilistic model presented here yields several findings about comparative reconstruction more generally. One of these corollary findings, namely the effect of segment inventory size and phonotactics on the evaluation of a reconstruction, is presented here.

2 Methodology

To evaluate a comparative reconstruction from a mother language (proto-language) to some daughter language (attested language), we will estimate the probability that a random daughter language merits a reconstruction of the same size as the daughter or smaller. For our purposes, a language is simply an ordered list of phonological forms as in (1), where order corresponds to the semantic exponent. The particular order of words does not matter, so long as it is kept consistent between languages that share an analysis.

(1)

(Gloss)	Latin	Romanian
goat	kapra	kaprə
will of the gods	numen	--
spring	ve:ra	primavarə
summer	ae̯stas	varə

We define a reconstruction as an exhaustive series of transformations from the mother wordlist to some daughter wordlist. The reconstruction is assumed to be exhaustive, in that, with the mother wordlist as input to the transformations, the daughter wordlist must be the output. The comparative reconstruction literature is mostly concerned with regular sound change, and regular sound changes are certainly a part of the reconstruction as defined here. However, any well-defined transformation can also be part of the reconstruction. For example, in addition to sound changes, a reconstruction from Latin to Romanian in (1) would usually require semantic change (SPRING → SUMMER) and lexeme loss (WILL OF THE GODS). Other possible transformations include morphological change, analogical change, and calques.

Contrary to most other quantitative approaches to historical linguistics, this framework compares mother and daughter, not sister, wordlists. The analysis of sister wordlists derived from the same mother, i.e. a language family, requires independent applications of the framework, one for each mother-daughter pair. For our purposes, the existence of the proto-language, in the exact shape necessitated by the reconstruction, is assumed, and it is the link to each daughter language that is quantified. This is, in fact, the tacit standard in historical linguistics. A comparative reconstruction does not demonstrate that a proto-language existed; it shows how, if it had existed, one would derive living languages from it (Nichols, 1995).

The null hypothesis of a reconstruction is that the similarities between the mother and daughter are spurious and attributable to chance. This hypothesis assumes an understanding of what could have been produced by chance. It is somewhat common practice in computational historical linguistics to estimate the extent of random possibilities in a wordlist by simply permuting the words in that list in what is known as a Monte-Carlo simulation (Baxter, 1995; Kessler & Lehtonen, 2006; Croft, 2008; Kessler, 2008; Kessler, 2015; Hruschka *et al.*, 2015; Zhang & Gong, 2016). Such an approach is reasonable because, by definition, each phonological word in a wordlist is permissible in that language, and, through Saussarian arbitrariness (1916), phonology and semantics cooccur freely. Therefore, shuffling a wordlist is a reasonable approach to simulate phonologically constrained randomness. However, such an approach severely underestimates the degree of variation possible in language as it takes every accidental gap to be a systematic gap.

Thus, a different approach to wordlist randomness will be pursued here. We define a set of *synchronically equivalent* wordlists to the daughter. A synchronically equivalent wordlist is one that shares the same phonological and phonotactic properties, or, put differently, one that comprises a subset of the possible words in that language. Naturally, all permutations of a wordlist are synchronically equivalent to each other, as are many other wordlists. In general, a synchronically equivalent wordlist shares the segmental inventory of the daughter, as well as any phonotactic and phonological restrictions; these include restrictions on segment cooccurrence, harmony, neutralization, prosodic phenomena, restrictions on word length, etc. It is the null hypothesis that the daughter wordlist was drawn at random from this set; in other words, it is the null hypothesis that the phonological attributes of the daughter are predetermined, but the shape and order of lexical items are random.

Calculating the size of the synchronically equivalent wordlist set requires a detailed understanding of the phonology of the daughter language, as well as some mathematical subtlety. In almost all cases, the final number is extremely large, often on the order of 10^{500}. Such calculations will not be performed in this paper; instead, more general arguments about the relationship between certain phonological properties and their effect on reconstruction likelihood will be presented.

The alternative hypothesis of a reconstruction is that the similarities between the mother and daughter are unlikely to have developed through chance alone. In reference to a comparative reconstruction, this alternative hypothesis implies the existence of a distance metric between the mother and daughter. The distance metric employed should reflect the number and type of changes that must have occurred to yield the daughter wordlist from the mother wordlist, rather than a typological or phonological similarity measure. Thus, the alternative hypothesis states that the distance between the mother and daughter, i.e. the number of changes required to transform the former into the latter, is lower than what would be expected through chance alone.

We define a set of *diachronically equivalent* wordlists to the daughter in reference to a given reconstruction. A diachronically equivalent wordlist is one that could have been derived from the mother wordlist by the same number and type of transformations as the daughter wordlist. If the reconstruction in question is efficient, i.e. devoid of superfluous transformations, the set of diachronically equivalent wordlists is the smallest possible such set which also includes the daughter wordlist. Unlike synchronically equivalent wordlists, a wordlist can only be diachronically equivalent in reference to a reconstruction, since it is the distance from the mother to the daughter that defines membership to the set.

The greater the number of transformations required by a reconstruction, the larger the set of diachronically equivalent wordlists. As the set of diachronically equivalent wordlists grows, the reconstruction proposed becomes less and less convincing, because a reconstruction of the same magnitude is compatible with more and more wordlists. At a certain point, the proposed reconstruction does not bode much better than chance, as the set of diachronically equivalent wordlists contains, along with the daughter, almost any other relevant wordlists.

To evaluate the null hypothesis that any similarities between the mother and daughter wordlists are attributable to chance, one needs to evaluate the likelihood that a randomly generated wordlist is as or more diachronically proximate to the mother wordlist than the daughter. This is the same as the probability that a wordlist synchronically equivalent to the daughter is diachronically equivalent to the daughter given the reconstruction, or, assuming a uniform probability distribution, the proportion of diachronically equivalent wordlists in the set of synchronically equivalent wordlists. For example, if that proportion were .5, i.e. if half of synchronically equivalent wordlists are also diachronically equivalent, a wordlist that shares the phonological properties of the daughter merits a reconstruction of equal or lesser magnitude 50% of the time.

This intuition is captured more rigorously in (2). Let *S* be the set of wordlists synchronically equivalent to the daughter, and *D* the set of wordlists diachronically equivalent to the daughter given the proposed reconstruction. Through the formula for conditional probability, the likelihood that a member of *S* is also a member of *D* is equal to the cardinality of the intersection of the two sets divided by the cardinality of *S*.

(2)

$$P(D|S) = \frac{|D \cap S|}{|S|}$$

One may also conceptualize the formula in (2) more abstractly. Imagine a high-dimensional space of wordlists, where the position of each wordlist is determined by the values (segments) in each word. Distance between similar wordlists is shorter, and distance between dissimilar wordlists longer. A reconstruction starts with the point corresponding to the mother wordlist. Each transformation in the reconstruction alters the mother lexicon in some way, translating the starting point to some new point in the space. Alternative transformations of the same type may alter the wordlist to the same extent but in a different way, which can be thought of as a translation of the same distance but in a different direction. In this space, a reconstruction is a (not necessarily straight) path from the mother to the daughter. The set of diachronically equivalent wordlists is a high-dimensional sphere with the mother wordlist at its centre. With each new transformation the sphere grows to include the newest intermediate wordlist. The reconstruction stops when the latest layer of the sphere contains the daughter wordlist. At that point, the sphere corresponds to the entire set *D*, and the portion of the sphere containing synchronically equivalent wordlists corresponds to $|D \cap S|$. This abstraction is loosely approximated in Figure 1.

Notice that the formula in (2) contains no information that is exclusive to linguistics. Put simply, the formula estimates the likelihood that a state generated randomly in accordance with some restrictions can be derived through an equal or lesser number of transformations from some state as some other state. The same approach can be used to estimate how likely it is that two chess positions occurred in the same game, or that one image was derived from another through photo-editing (Munteanu, *in prep*.). However, for most other fields, detecting and listing the number of transformations between prior states and later states is not a common endeavor. Contrariwise, one of the chief objectives of historical linguistics is to identify and document all sound changes that have occurred in the past, making comparative reconstruction ideal for the application of this methodology.

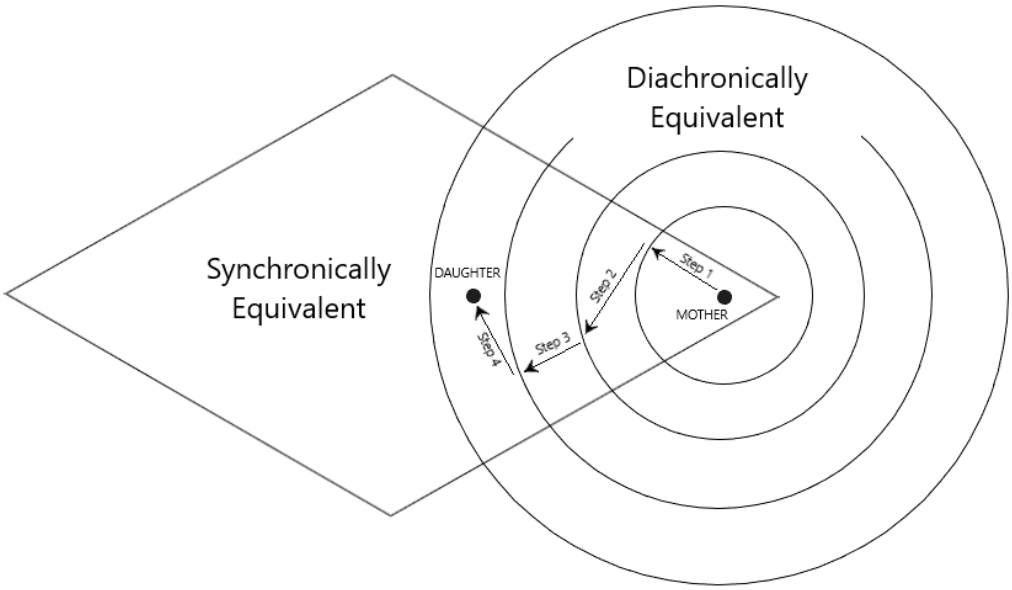

FIGURE 1: A two-dimensional abstract representation of (2). The reconstruction is marked as a path of transformations from the mother wordlist to the daughter wordlist. Each transformation introduces an additional layer of diachronically equivalent wordlists, until the newest layer envelops the daughter wordlist. |S| corresponds to the rhombus on the left. |D| corresponds to the circle on the right. The intersection of the two sets is the 'slice' of the rhombus that is contained within the circle.

Although $P(D|S)$ in (2) is not that same thing as a *p* value in statistics, the two are developed with the same purpose in mind. Both are the sum of the likelihood of the observed result (the daughter wordlist) and any result that is equally surprising or less surprising. Thus, just like a traditional *p* value, $P(D|S)$ should be thought of as an estimate as to how surprising an observation is rather than the probability of finding a particular wordlist or type of wordlist. This property of the $P(D|S)$ measure makes it suitable for hypothesis testing.

3 Results

3.1 *Estimating* $|S|$ Assume that the set of possible words in a language is finite though extremely large. This assumption follows from the fact that both the number of possible segments and their arrangements in a language are finite.

Although all words in a wordlist are selected from the same set of possible words, I will assume that they are otherwise phonologically independent. In other words, two words in a wordlist share the same phonological inventory and the same phonological restrictions, but the phonological shape of one word cannot be otherwise predicted from the phonological shape of another. Between-word phonological independence is conducive to a simple mathematical model, and this is the main reason for the assumption. Linguistically, at least two arguments about between-word dependence can be made.

Firstly, there is some evidence of between-word homophony avoidance in diachronic linguistics (Ogura & Wang, 2018; Silverman, 2009). Both computational modelling (Blevins & Wedel, 2009; Winter & Wedel, 2016) and language learning experiments (Yin & White, 2018) confirm that errors in language transmission tend to result in greater contrast between different word pairs. Homophony with a taboo word is particularly illicit, as has been recorded for many languages (Burridge & Benczes, 2018). This means that the exact same phonological string may be prohibited from appearing in multiple positions in the wordlist, and that this effect is stronger in certain position in the wordlist. Therefore, word shape within the wordlist may not be strictly independent.

Secondly, consider the Japanese words 電気 [denki] ‘electricity’, 病気 [bjouki] ‘sick’, and 天気 [tenki] ‘weather’. All of these words, and many more in the language, contain the phonological string [ki]. This incidence is not a result of phonological restrictions on word shape but is instead caused by morphology. The root 気 [ki] ‘energy’ appears in many Sino-Japanese compounds and can also stand on its own. Because the morpheme is so frequent, one may find that the phonological diversity of Japanese is lower than expected. In most wordlists of Japanese, the string [ki] is overrepresented, and, by comparison, the strings [ko] or [ke], underrepresented. In a particularly limited sample, the dearth of data may mimic a phonological restriction against [ko] or [ke] sequences. This sort of pattern is not exclusive to Japanese, as all languages exhibit morphology, and each language is expected to have one or more overrepresented morphemes. Thus, a sample wordlist from a natural language will exhibit similar strings more often than a wordlist generated by the simple concatenation of phones. Once again, this means that words in a wordlist are technically not independent.

It is likely the case that both homophony avoidance and morphology limit the phonological diversity of a language and, therefore, decrease the size of *S*. However, these arguments are rather subtle and difficult to implement mathematically. It is also up to discussion whether these are emergent trends in the lexicon or active psychological restrictions and whether their effect needs to be manifested in $|S|$. I make the choice of not including homophony avoidance or morphological frequency in the calculation and upholding the assumption that the shape of every word is independent. This is not a requirement of the framework and future implementations may seek a way to restrict *S* beyond what is given here.

Thus, under the assumptions that word shape is finite and independent, the way to calculate $|S|$ is given in (2), where *c* is *word complexity*, i.e. the number of possible words in the language, and *t* is *total wordlist length*. To form a member of *S*, for each position in the wordlist, select a word from the set of possible words in the daughter language.

(2)
$$|S| = c^t$$

3.2 *Phonological Factors* Combining the formulae in (2) and (3) reveals that the likelihood that a reconstruction from a mother wordlist to a daughter wordlist can be substantiated by chance is correlated with the number of possible words in the daughter language. Languages with fewer possible words are more likely to be substantiated by chance than languages with more possible words. The rate of growth of $P(D|S)$ is polynomial given a linear increase in word complexity c.

From the point of view of phonology, there are several factors which affect the number of possible words in a language: segmental inventory size, word length, and phonotactic restrictions. All three of these play a role in determining how likely it is that a reconstruction to a given daughter wordlist is spurious.

3.2.1 *Word Length* Most languages do not appear to have a set word length; however, word length is constrained in practice, as no language has words of infinite length, and word frequency is inversely correlated with length (Zipf, 1935) and decaying exponentially as a function of it (Sigurd *et al.*, 2004).

There is some suggestion that mean word-length is negatively correlated with segmental inventory size and phonotactic complexity, rendering word complexity relatively stable cross-linguistically (Nettle, 1995; Moran & Blasi, 2014; Pimentel *et al.*, 2020), as well as a similar negative correlation between complexity and speech rate (Pellegrino *et al.*, 2011). More broadly, the idea that a decrease in complexity in one linguistic domain must be coupled with an increase in complexity in another is known as the *compensation hypothesis* (Martinet, 1955). Even though there exists evidence in favor of compensation, word length still varies between languages with similar phonotactic complexity. Furthermore, there is no reason to believe that wordlists, such as those employed in comparative reconstruction, adequately capture all sources of linguistic complexity, such as syntax, pragmatics, etc. As such, it makes sense to ask how variation in word complexity in general and mean word-length in particular affect the $P(D|S)$ measure.

Assume two wordlists of the same length, segmental inventory, and phonotactic restrictions, from two languages, L_1 and L_2, both derived from the same mother wordlist through the same number and type of transformations (sound changes, semantic changes, borrowings, etc.). Note that the exact transformations must be different to yield two different results, but the differences may be *ad hoc* and not indicative of a greater diachronic distance for either of the two daughter wordlists. The only difference between the two languages is word-length.

We can calculate the ratio of $P(D|S)$ between L_1 and L_2 as in (4). Let us call this ratio R_w, for word-length. The symbol x shall refer to the number of possible segments in a given position and w_1 for word-length in L_1 and w_2 in L_2. Word complexity c is equal to x^w. This, along with (2), implies the statement in (4).

(4)
$$R_w = \frac{x^{w_1 t}}{x^{w_2 t}} = x^{w_1 - w_2}$$

Thus, $P(D|S)$ increases exponentially with a linear increase to word-length. For example, if words in L_2 are on average twice as long as word L_1 (say, 10 phonemes vs 5 phonemes or 4 syllables vs 2 syllables) then the $P(D|S)$ of L_2 can be calculated by dividing the $P(D|S)$ of L_1 by x^{w_1}. Incidentally, the effect on $P(D|S)$ from increases to word-length is identical as the one from to equivalent increases to wordlist length. Stated more intuitively, this means that doubling each word is the same as doubling the number of words, or that 100 words four syllables in length contain the same amount of complexity as 200 words two syllables in length.

3.2.2 *Segmental Inventory Size* The number of segments in a language varies greatly, from 11 in Pirahã and Rotokas to 141 in !Xu (Maddieson, 1984). As with word-length, there is some evidence that languages make up for a smaller inventory size with complexity in other domains (Nettle, 1995; Moran & Blasi, 2014). However, the correlation is rather weak (Pimentel *et al.*, 2020). As such, it makes sense to ask how variation in segmental inventory size affects the $P(D|S)$ measure.

Assume two wordlists of the same length, phonotactic restrictions, and word-length from two languages, L_1 and L_2, both derived from the same mother wordlist through the same number and type of transformations (sound changes, semantic changes, borrowings, etc.). The only difference between the two

languages is segmental inventory size. Note that it is not the case that every segment can occur in every position within a word, due to phonotactic restrictions. However, we will assume that an increase in segmental inventory size corresponds to an increase in the number of possible segments in a given phonological position, even if the correlation between the two is not one-to-one.

We can calculate the ratio of $P(D|S)$ between L_1 and L_2 as in (5). Let us call this ratio R_x. The symbol *x* shall refer to the number of possible segments in a given position, x_1 for L_1 and x_2 for L_2. Word-length is represented by *w*. Word complexity *c* is equal to x^w. This, along with (2), implies the statement in (5).

(5)
$$R_x = \frac{{x_1}^{wt}}{{x_2}^{wt}}$$

Thus, $P(D|S)$ increases polynomially with a linear increase to the average number of possible segments in a phonological position, which is strongly correlated with segmental inventory size. The effect of segmental inventory size on $P(D|S)$ increases as the mean word-length in the language increases and as total word list length increases.

3.2.1 *Phonotactics* In this context, phonotactics refers to any cooccurrence restriction between segments, as all such restrictions serve to reduce the number of possible words *c*, and, as per (2), increase $P(D|S)$. Restrictions on syllable size, onsets and codas, vowel-consonant interactions, neutralization, harmony, prosody, all restrict word shape in a language beyond what is implied by word-length and segmental inventory alone. The exact effect of phonotactics on the likelihood that a reconstruction is spurious is difficult to estimate, as it depends on the nature of the restriction, but also on word-length and segmental inventory size.

For example, imagine a vowel harmony system where a value for a feature ([±back], [±round], etc.) in the first vowel determines the value for the feature in all subsequent vowels, such as the case of Turkish (Kabak, 2011). Such a restriction effectively reduces the vowel inventory of the language by a factor of 2, but only for the second vowel onwards; the first vowel remains unrestricted. Thus, the number of possible words depends on the number of vowels but also on word-length.

The only conclusion that can be drawn with certainty is that additional restrictions can only decrease the set of synchronically equivalent wordlists and, therefore, increase $P(D|S)$. Increasing the number of synchronically equivalent wordlists by introducing phonotactic restrictions is impossible. Nevertheless, it is very likely that the effect of phonotactics on $P(D|S)$ is substantially smaller than that of word-length or segmental inventory size. This is because the exponential nature of (4) and polynomial nature of (5) often mean that even a tiny decrease to word-length or inventory size can result in a decrease in the number of possible words by a factor of 2 or greater. Meanwhile, it is difficult to imagine a phonotactic restriction that decreases the number of possible words by the same amount.

4 Conclusion

This paper introduced a novel quantitative methodology for evaluating manual comparative reconstructions. This method is incumbent on the existence of a manual comparative reconstruction and, unlike previous quantitative methods, cannot give a result contradictory to the reconstruction. The primary goal for this framework is to reconcile traditional and quantitative methodologies and act as an objective and accessible platform for comparative reconstruction, thereby extending the scope of historical linguistics further into the past.

This methodology can also be used to reason about comparative reconstruction more generally. A few theoretical corollaries of the framework have been presented in this paper. Namely, it has been shown that the likelihood that a reconstruction of equal size to the one proposed can be generated randomly, i.e. the likelihood the reconstruction is spurious, is related to some of the phonological properties of the daughter (descendent) language. This likelihood decreases exponentially as word-length increases and decreases

polynomially as segmental inventory increases. Additionally, active phonological processes and cooccurrence restrictions in the language that decrease the number of possible word-shapes – such as phonotactics, prosody, harmony, and neutralization – all serve to increase the likelihood that a reconstruction to that language is spurious.

References

Atkinson, Q., & Gray, R. (2005). Curious Parallels and Curious Connections—Phylogenetic Thinking in Biology and Historical Linguistics. *Systematic Biology, 54*(4), 513-526. doi:10.1080/10635150590950317

Baxter, W. H. (1995). A Stronger Affinity than could have been Produced by Accident: A Probabilistic Comparison of Old Chinese and Tibeto-Burman. *Journal of Chinese Linguistics Monograph Series, 8*, 1–39.

Blevins, Juliette, & Wedel, Andrew (2009). Inhibited sound change. *Diachronica, 26*(2), 143-183. doi:10.1075/dia.26.2.01ble

Bostoen, K. (2007). Pots, Words And The Bantu Problem: On Lexical Reconstruction And Early African History. *The Journal of African History, 48*(2), 173-199. doi:10.1017/s002185370700254x

Bomhard, A. (2008). Reconstructing Proto-Nostratic: Comparative Phonology, Morphology, and Vocabulary. *Leiden Indo-European etymological dictionary series.*

Bowern, C., & Atkinson, Q. (2012). Computational phylogenetics and the internal structure of Pama-Nyungan. *Language, 88*(4), 817-845. doi:10.1353/lan.2012.0081

Brown, C., Beck, D., Kondrak, G., Watters, J., & Wichmann, S. (2011). Totozoquean. *International Journal of American Linguistics, 77*(3), 323-372. doi:10.1086/660972

Brown, C. H., Wichmann, S., & Beck, D. (2014). Chitimacha: A Mesoamerican language in the lower Mississippi valley. *International Journal of American Linguistics, 80*(4), 425–474. https://doi.org/10.1086/677911

Burridge, K., & Benczes, R. A. (2019). Taboo as a driver of language change. In K. Allan (ed.), *The Oxford Handbook of Taboo Words and Language* (1st ed., pp. 180-199). Oxford University Press.

Campbell, L. (1998). Nostratic: A Personal Assessment. *Nostratic: Sifting the Evidence*.

Campbell, L. (2011). The Dene–Yeniseian Connection. *International Journal of American Linguistics, 77*(3), 445-451. doi:10.1086/660977

Chang, W., Cathcart, C., Hall, D., & Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language, 91*(1), 194-244. doi:10.1353/lan.2015.0005

Croft, W. (2008). Evolutionary Linguistics. *Annual Review of Anthropology, 37*(1), 219-234. doi:10.1146/annurev.anthro.37.081407.085156

Diakonoff, I., Starostin, S. (1986). Hurro-Urartian as an Eastern Caucasian Language. *Münchener Studien zur Sprachwissenschaft: Beiheft, 12*.

Dixon, R., & Aikhenvald, A. (2006). *Complementation a cross-linguistic typology*. Oxford: Oxford University Press.

Doerfman, G. (1963). Bemerkungen zur Verwandtschaft der sog. altaische Sprachen [Remarks on the relationship of the so-called Altaic languages]. *Türkische Und Mongolische Elemente Im Neupersische*, 51-105.

Dolgopolsky, A. (2012). *Nostratic dictionary*. Cambridge: McDonald Institute for Archaeological Research.

Downey, S., Hallmark, B., Cox, M., Norquest, P., & Lansing, J. (2008). Computational Feature-Sensitive Reconstruction of Language Relationships: Developing the ALINE Distance for Comparative Historical Linguistic Reconstruction. *Journal of Quantitative Linguistics, 15*(4), 340-369. doi:10.1080/09296170802326681

Dunn, M. (2012). Anthropological papers of the University of Alaska: The Dene-Yeniseian connection (review). *Language, 88*(2), 429-432.

Georg, S., Michalove, P., Ramer, A., & Sidwell, P. (1999). Telling general linguists about Altaic. *Journal of Linguistics, 35*(1), 65-98. doi:10.1017/s0022226798007312

Gray, R., & Atkinson, Q. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature, 426*(6965), 435-439. doi:10.1038/nature02029

Greenberg, J. H. (1987). *Language in the Americas*. Stanford Univ. Press.

Greenhill, S., Wu, C., Hua, X., Dunn, M., Levinson, S., & Gray, R. (2017). Evolutionary Dynamics of Language Systems. *Proceedings of the National Academy of Sciences*, *114*(42). https://doi.org/10.1073/pnas.1700388114

Holman, E., Brown, C., Wichmann, S., Müller, A., Velupillai, V., Hammarström, H., & Egorov, D. (2011). Automated Dating of the World's Language Families Based on Lexical Similarity. *Current Anthropology, 52*(6), 841-875. doi:10.1086/662127

Hruschka, D., Branford, S., Smith, E., Wilkins, J., Meade, A., Pagel, M., & Bhattacharya, T. (2015). Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution. *Current Biology, 25*(1), 1-9. doi:10.1016/j.cub.2014.10.064

Illic-Svityc, V. (1963). *Алтайские Дентальные* [The Altaic Dentals]. Jazykoznanie 6: 37–56

Kessler, B. (2008). The Mathematical Assessment of Long-Range Linguistic Relationships. *Language and Linguistics Compass, 2*(5), 821-839. doi:10.1111/j.1749-818x.2008.00083.x

Kessler, B. (2015). Computational and Quantitative Approaches to Historical Phonology. *Oxford Handbooks Online*. doi:10.1093/oxfordhb/9780199232819.013.030

Kessler, Brett & Annukka Lehtonen. 2006. Multilateral comparison and significance testing of the Indo-Uralic question. In Peter Forster & Colin Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*, 33–42. Cambridge, England: McDonald Institute for Archaeological Research. ISBN 978-1-902937-33-5

Kiparsky, P. (2015). New perspectives in historical linguistics. *The Routledge Handbook of Historical Linguistics*. doi:10.4324/9781315794013.ch2

Kondrak, G. (2003). Phonetic Alignment and Similarity. *Computers and the Humanities, 37*(3), 273-291. doi:10.1023/a:1025071200644

Maddieson, I. (1984). Patterns of sounds. Cambridge studies in speech science and communication. Cambridge: Cambridge University Press.

Martinet, André. (1955). *Économie des changements phonétiques: Traité de phonologie diachronique*. Francke.

Mcmahon, A., & Mcmahon, R. (2003). Finding Families: Quantitative Methods in Language Classification. *Transactions of the Philological Society, 101*(1), 7-55. doi:10.1111/1467-968x.00108

Miller, R. (1984). The Classification of the Uto-Aztecan Languages Based on Lexical Evidence. *International Journal of American Linguistics, 50*(1), 1-24. doi:10.1086/465813

Moran, S. & Blasi, D. (2014). Crosslinguistic comparison of complexity measures in phonological systems. In Frederick J. Newmeyer and Laurel B. Preston, editors, Measuring grammatical complexity, pages 217–240. Oxford University Press Oxford, UK.

Nettle, D. (1995). Segmental inventory size, word length, and communicative efficiency. Linguistics, 33:359–367.

Nichols, J., & Peterson, A. (1996). The Amerind Personal Pronouns. *Language, 72*(2), 336. doi:10.2307/416653

Norman, J. (2009). A New Look at Altaic. *Journal of the American Oriental Society, 129*(1), 83-89. doi:10.2307/40593870

Ogura, M. & Wang W. (2018). Evolution of Homophones and Syntactic Categories Noun and Verb. *Proceedings of the 12th International Conference on the Evolution of Language (Evolang12)*. doi:10.12775/3991-1.086

Pedersen, H. (1903). Türkische Lautgesetze [Turkish Phonetics]. *Zeitschrift Der Deutschen Morgenländischen Gesellschaft, 57*(3), 535-561.

Pellegrino, F., Chitoran, I., Marsico, E., & Coupé, C. (2011). A crosslanguage perspective on speech information rate. *Language*, 87(3):539–558.

Pimentel, T., Roark, B., & Cotterell, R. (2020). Phonotactic Complexity and Its Trade-offs. *Transactions of the Association for Computational Linguistics, 8*, 1-18. doi:10.1162/tacl_a_00296

Poppe, N. (1960). *Vergleichende Grammatik der altaischen Sprachen, Teil 1: Vergleichende Lautlehre* [Comparative grammar of the Altaic languages, Part 1: Comparative phonetics]. Wiesbaden: Harrassowitz.

Ramstedt, G. (1957). *Einführung in die altaische Sprachwissenschaft, I: Lautlehre* [Introduction to Altaic Linguistics, I: Phonetics]. Helsinki: Suomailais-Ugrilaisen Seura.

Ratcliffe, R. (2003). Afroasiatic Comparative Lexica: Implications for Long (and Medium) Range Language Comparison. *Proceeding of the Seventeenth International Congress of Linguists*.

Rexova, K., Frynta, D., & Zrzavy, J. (2003). Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics, 19*(2), 120-127. doi:10.1111/j.1096-0031.2003.tb00299.x

Ringe, D. (1999). How hard is it to match CVC-roots? *Transactions of the Philological Society, 97*(2), 213-244. doi:10.1111/1467-968x.00049

Robbeets, M. (2005). Is Japanese Related Turkic? *Wiesbaden: Harrasso witz*.

Rubicz, R., Melvin, K., & Crawford, M. (2002). Genetic Evidence for the Phylogenetic Relationship between Na-Dene and Yeniseian Speakers. *Human Biology, 74*(6), 743-760. doi:10.1353/hub.2003.0011

Ruhlen, M. (1994). The Origin of Language: Tracing the Evolution of the Mother Tongue. *Wiley*.

Shields, K. (2011). The "New Image" of Indo-European and the Nostratic Hypothesis: A Possible Reconciliation of Reconstructions. *Studia Etymologica Cracoviensia, 16*.

Sicoli, M., & Holton, G. (2014). Linguistic Phylogenies Support Back-Migration from Beringia to Asia. *PLoS ONE, 9*(3). doi:10.1371/journal.pone.0091722

Sigurd, B., Eeg-Olofsson, M., & Weijer, J. (2004). Word length, sentence length and frequency - Zipf revisited. *Studia Linguistica, 58*(1), 37-52. doi:10.1111/j.0039-3193.2004.00109.x

Silverman, D. (2009). Neutralization and Anti-Homophony in Korean. *Journal of Linguistics*, 46(2), 453–482, doi:10.1017/s0022226709990247.

Starostin, G. (2012). Dene-Yeniseian: A critical assessment. *Journal of Language Relationship, 8*(1), 117-138. doi:10.31826/jlr-2012-080109

Starostin, S., Dybo, A., & Mudrak, O. (2005). Etymological Dictionary of the Altaic Languages. *Handbook of Oriental Studies*, 8.

Swadesh, M. (1955). Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics, 21*(2), 121-137. doi:10.1086/464321

Theil, R. (2006). Is Omotic Afroasiatic? *Proceedings from the David Dwyer retirement symposium*.

Vajda, E. (2011). A Response to Campbell. *International Journal of American Linguistics*, *77*(3), 451-452. doi:10.
Vajda, E. (2016). Dene-Yeniseian. *Oxford Research Encyclopedia of Linguistics*. doi:10.1093/acrefore/9780199384655.013.31
Vovin, A. (2005). The End of the Altaic Controversy In Memory of Gerhard Doerfer. *Central Asiatic Journal*, *49*(1), 71-132.
Vovin, A. (2009). Japanese, Korean, and Other 'Non-Altaic' Languages. *Central Asiatic Journal*, *53*(1), 106-147.
Wikander, S. (1967). Maya and Altaic. *Ethnos*, *32*(1-4), 141-148. doi:10.1080/00141844.1967.9980993
Winter, B. & Wedel, A. (2016). The Co-evolution of Speech and the Lexicon: The Interaction of Functional Pressures, Redundancy, and Category Variation. *Topics in Cognitive Science*, *8*(2), 503-513. doi:10.1111/tops.12202
Yin, S. & White, J. (2018). Neutralization and homophony avoidance in phonological learning. *Cognition*, *179*, 89-101. doi:10.1016/j.cognition.2018.05.023
Zhang, M., & Gong, T. (2016). How Many Is Enough?—Statistical Principles for Lexicostatistics. *Frontiers in Psychology*, *7*. doi:10.3389/fpsyg.2016.019161086/660978
Zipf, G. (1935). The psycho-biology of language: An introduction to dynamic philology. MIT Press, Cambridge.