Journal of Quantitative Linguistics
2008, Volume 15, Number 4, pp. 340–369
DOI: 10.1080/09296170802326681

Routledge
Taylor & Francis Group

Computational Feature-Sensitive Reconstruction of Language Relationships: Developing the ALINE Distance for Comparative Historical Linguistic Reconstruction*

Sean S. Downey[1], Brian Hallmark[1], Murray P. Cox[2,3], Peter Norquest[4] and J. Stephen Lansing[1,3]
[1]Department of Anthropology, University of Arizona, USA; [2]Arizona Research Laboratories, University of Arizona, USA; [3]Santa Fe Institute, USA; [4]Department of Linguistics, University of Arizona, USA

ABSTRACT

Historical relationships among languages are used as a proxy for social history in many non-linguistic settings, including the fields of cultural and molecular anthropology. Linguists have traditionally assembled this information using the standard comparative method. While providing extremely nuanced linguistic information, this approach is time-consuming and labor-intensive. Conversely, computational approaches are appreciably quicker, but can potentially introduce significant error. Furthermore, current methods often use cognate sets that were themselves coded by historical linguists, thus reducing the benefit of computational approaches. Here we develop a method, based on the ALINE distance, to extract feature-sensitive relationships from paired glosses, datasets that require minimal contribution from trained linguists beyond transcription from primary sources. We validate our results by comparison with data generated independently via the comparative method, and quantify error rates using consistency indices. To showcase our method's utility and to demonstrate its robustness at local and regional scales, we apply it to two language datasets from eastern Indonesia. As linguistic datasets proliferate, scalable computational methods that mimic historical linguistic reconstruction will become increasingly necessary. Although at present we cannot disentangle all the processes driving linguistic change (e.g. lexical borrowing), our method provides a robust and accurate alternative to manual linguistic analysis. The feature-sensitive method adopted here accurately and automatically identifies emergent patterns hidden in

*Address correspondence to: J. Stephen Lansing, Department of Anthropology, University of Arizona, Tucson, AZ 85721, United States of America.
E-mail: lansing@santafe.edu

traditional word-lists by analysing critical phonetic information that is discarded (or required as prerequisite) by many current cognate-based computational methods. This approach is not intended to supplant manual linguistic analysis, but has an important role in quickly generating robust data for non-linguistic fields or interdisciplinary projects that require formal quantitative analysis of historical linguistic relationships. Our approach provides a workable approximate phylogeny in cases where a trained linguist is unavailable, or otherwise significantly reduces the time and effort required for manual classification.

INTRODUCTION

Determining how and why languages change is a central goal of modern linguistics. To understand these processes completely, it is important to determine how languages evolve relative to the historical processes acting in communities where they are spoken. Growing bodies of joint spatial, linguistic and genetic data now make such complex comparisons possible. However, shared quantitative frameworks are essential if datasets from multiple disciplines are to be compared and integrated. The magnitude of these datasets also requires the application of robust computational approaches.

The concept of distance is one of the most basic quantities used to compare data from different sources. While natural definitions exist for geography and several distance metrics have been defined for genetic data, there is no consistent use, or clear definition, of a corresponding linguistic distance. In addition to facilitating comparisons with other data types, such a statistic would allow the rich assemblage of distance-based statistical techniques developed by other fields to be transferred directly to linguistics. For many applications, a meaningful linguistic distance would allow automation of analysis, and therefore facilitate the study of large language datasets, which are time-prohibitive to analysis via traditional linguistic approaches.

Here, we focus specifically on the construction of a computational linguistic statistic, the ALINE distance, to compare linguistic information with other data sources quantitatively. This distance measure will allow automation of some historical linguistic tasks, such as tree-building and map-making. Results integrating our linguistic and genetic data are too extensive to be reported fully here, but are published elsewhere (Lansing et al., 2007). We show that the ALINE distance has several properties that are superior to previous distance statistics, and we suggest

new computational directions to allow the exploration of large linguistic datasets and link them with non-linguistic data.

Motivation

Our desire to develop the ALINE distance derives from research within the Austronesian Societies project, which seeks to integrate geographic, genetic, linguistic and cultural data from central and eastern Indonesia, and thereby understand the historical processes driving both genetic and linguistic change in this region. We currently possess paired geographic, genetic and linguistic data from 50 Indonesian communities, and further data collection is planned. Our linguistic data comprises a database of 200-word Swadesh lists from more than 400 Indonesian languages. While considerable historical research is already available for these languages, the standard comparative method does not naturally produce quantitative values that can be compared statistically with genetic and geographic data. The sheer number of languages in our project, and a desire to understand the joint history of gene and language co-evolution, drove us to develop the computationally-based distance statistic described here.

Linguistic Distances

Beginning with L. Luca Cavalli-Sforza (Cavalli-Sforza et al., 1988; Cavalli-Sforza, 1997), many studies have examined gene/language relationships on continental and regional scales (Sokal, 1988; Barbujani & Sokal, 1990; Cavalli-Sforza et al., 1992; Smouse & Long, 1992; Chen, Sokal & Ruhlen, 1995; Cavalli-Sforza, 1997; Cox, 2003; Nettle & Harriss, 2003; Hunley & Long, 2005), and more recently, at smaller geographical levels (e.g. Cox & Lahr, 2006). Some attempts (Cavalli-Sforza et al., 1988, 1992; Cavalli-Sforza, 1997) simply compared genetic and linguistic phylogenies, and identified general associations between subgroup clades. However, the difficulty of assigning statistical significance to phylogenetic comparisons at these large scales limited the power of this approach. Others have examined allele frequency data and looked to see if abrupt changes in gene frequencies correspond with linguistic barriers (Barbujani & Sokal, 1990; Hunley & Long, 2005). Later studies applied distance-based, Mantel correlation methods to examine relationships between geographic, genetic and linguistic distances (Sokal, 1988; Smouse & Long, 1992; Chen et al., 1995; Nettle & Harriss, 2003). However, these distance measures suffer from several deficiencies. Chen et al., (1995) used subjective distance estimates on a scale of 1–60, as assessed by a trained

historical linguist. Sokal (1988), Nettle and Harris (2003), and others used the number of branch nodes separating languages within an existing phylogeny to construct tree-based integer stepwise distances. At a slightly finer scale, Smouse and Long (1992) constructed a distance measure from the percentage of shared cognates between languages.

One clear deficiency of these distance measures is their lack of resolution. For example, many languages are only one branch apart in a tree-based distance, but their degree of similarity may be highly variable. Likewise, cognate-based distances suffer from their inability to distinguish similarity within cognate and non-cognate classes. While these distances may be fine enough to detect broad correlations between language families and genetic lineages at large scales, their inability to distinguish more subtle differences at local or regional scales is severely restrictive. As more data becomes available at the scale of communities, it will become increasingly important to have distance measures that distinguish closely related languages, and hence, closely related words. All these approaches also suffer from their requirement for pre-existing historical linguistics analysis – either to define a tree of language relationships, or to determine shared cognates. While the comparative method is unparalleled in the detail and nuance it can extract from linguistic datasets, it does not naturally produce quantitative measures. Furthermore, distances may also be difficult or impossible to compute for large datasets due to obligate manual implementation of the comparative approach. (Interestingly, phylogenetic trees are constructed in reverse fashion in genetics; trees are computed directly from distance matrices).

In our view, a comparative linguistic distance measure should satisfy the following requirements:

1. Be based on the finest phonetic detail feasible.
2. Quantify linguistic differences in a linguistically meaningful way.
3. Be easily computable for large datasets.
4. Be free from subjective decisions.
5. Require little or no preliminary analysis, such as word classification into cognate sets or construction of a language tree.
6. Reproduce results from standard historical methods as closely as possible.

The linguistic literature discusses several distance measures that meet some of these requirements (Kruskal, 1983; Covington, 1996, 1998;

Somers, 1998; Kondrak, 1999; Nerbonne, Heeringa & Kleiweg, 1999; Lebart & Rajman, 2000; Heeringa et al., 2006), and most are based on the alignment of character strings with subsequent scoring of aligned features. The simplest version of this type of measure, the Levenstein or edit distance, counts the number of changes required for one string to mutate into another (Kruskal, 1983). More sophisticated variations incorporate complex alignment and scoring algorithms. Some of these algorithms have been used for linguistic tasks like cognate identification and dialect research (Heeringa et al., 2006), and the Levenstein distance has previously been used to compare linguistic and geographic data (Nerbonne et al., 1999). However, none have been applied to linguistic comparisons with genetic data. The ability of these metrics to reproduce results from traditional linguistic methods also remains unexplored.

Kondrak (1999) reviewed the strengths and weaknesses of the main algorithms for aligning and scoring phonetic sequences, and demonstrated the performance benefits of the ALINE algorithm for this task. Taking this research as our foundation, we propose a new distance measure, the ALINE distance. This is based on feature-level phonetic similarity scores for word pairs, is computed with Kondrak's ALINE algorithm, meets the requirements above stated for a distance measure, and proves a meaningful measurement of linguistic distance. This linguistic distance is well suited to comparisons with genetic and geographic data. We assume no prior linguistic information, apart from lists of shared-meaning word pairs. Language data need not be categorized in any other way (e.g. into cognate sets or language phylogenies). The approach can be readily scaled to thousands of words and languages, and we show that the method is able to reconstruct traditional comparative historical linguistic classifications with some accuracy.

Overall, our approach promises to be useful in several contexts. We stress that we do not envisage ALINE replacing the traditional methods of trained linguists, but rather see the benefit of both approaches being used in conjunction. Our method can act as a first-pass operation for large datasets, therefore saving linguists many hours of initial classification. It also provides a way to quantify language relationships that have previously been determined only qualitatively. We imagine a number of data-mining and automated classification/mapping tasks that could initially be performed computationally, and later checked by a linguist who is familiar with the languages under study. Lastly, although our

research focuses on relationships between geography, genetics and language, we note no reason why this linguistic distance could not be compared with other data sources, such as economic, political or cultural information.

ALINE ALGORITHM

The ALINE algorithm (Kondrak, 1999, 2000) examines shared-meaning word pairs, and generates a similarity score for the optimal feature-based phonetic alignment of each word pair. Although other algorithms exist for this purpose, Kondrak (1999, 2000) demonstrated many important benefits of the ALINE approach. This algorithm has been implemented as software coded in C++, which is used here as the basis for the ALINE distance. Phonetic segments are represented as vectors of feature values in the following categories: prosodic (syllabic), place, manner, phonation, vowel length and colour. Features are locally aligned to determine the optimal phonetic similarity between word pairs, and an optimal similarity score is generated (for detailed description, see Kondrak, 1999, 2000).

ALINE uses ASCII characters to represent its IPA counterparts. For example, ⟨p⟩ represents [p], ⟨i⟩ represents [i], and so forth. When this notation fails, a set of modifiers (represented by capital letters) is available for combination with standard ASCII characters. These can represent features such as place and manner. The set of modifiers presently utilized by ALINE is listed in Table 1.

The place characters ⟨D, V, X, P⟩ represent the dental, palato-alveolar, retroflex and palatal places of articulation, respectively. Using plain characters in combination with place of articulation modifiers, ALINE can represent the following series of consonants based on the unmodified alveolar series (see Table 2).

Table 1. ALINE IPA to ASCII coding schema.

D	dental	N	nasal
V	palato-alveolar	A	aspirated
X	retroflex	H	long
P	palatal	F	front
S	fricative	C	central

Table 2. Examples of IPA character encodings for the unmodified alveolar series.

IPA	ALINE	IPA	ALINE	IPA	ALINE	IPA	ALINE	IPA	ALINE
t	⟨t⟩	t̪	⟨tD⟩	ʧ	⟨tsV⟩	ʈ	⟨tX⟩	c	⟨tP⟩
d	⟨d⟩	d̪	⟨dD⟩	ʤ	⟨dzV⟩	ɖ	⟨dX⟩	ɟ	⟨dP⟩
s	⟨s⟩	θ	⟨sD⟩	ʃ	⟨sV⟩	ʂ	⟨sX⟩	ɕ	⟨sP⟩
z	⟨z⟩	ð	⟨zD⟩	ʒ	⟨zV⟩	ʐ	⟨zX⟩	ʑ	⟨zP⟩
n	⟨n⟩	n̪	⟨nD⟩			ɳ	⟨nX⟩	ɲ	⟨nP⟩

Therefore, this system is representationally flexible. Table 3 shows the relationship between IPA characters and their corresponding ALINE encoding. This can capture differences within the traditional framework of distinctive features, as every potential feature may be assigned its own modifier. Moreover, modifiers may themselves be used in combinatorial fashion, as is the case with the interdental fricatives illustrated above.

Notably, the current implementation of features in ALINE is not exhaustive. For instance, the uvular place of articulation cannot currently be represented, and we are forced to discard this information by merging uvulars with plain velars in our study (Table 4).

However, this problem has a relatively straightforward solution: choose a new modifier (say, ⟨U⟩) for [uvular], and assign an appropriate weight in the ALINE algorithm (Table 5). This approach can be extended to any phonemes not currently included in the ALINE character set, and such modifications are natural extensions to future versions of the ALINE software.

LINGUISTIC METRICS FROM WORD-PAIR SIMILARITY

The ALINE algorithm returns a phonetic similarity score for a given word pair (Kondrak, 1999, 2000). However, this raw score is not suitable as a distance statistic, because scores from different word pairs are not directly comparable. Furthermore, the raw ALINE score measures only the similarity of optimally aligned phonetic segments of any two given words, not the words overall. Consequently, the raw ALINE score must be transformed mathematically into a ratio scale penalized for unaligned characters, if it is to be used as a metric of phonetic distances between words, or by extrapolation, representative distances between languages.

Table 3. IPA characters and corresponding ALINE encoding (minus ⟨ ⟩').

Obstruents		Sonorants		Vowels	
IPA	ALINE	IPA	ALINE	IPA	ALINE
p	p	m	m	i	i
t	t	ɳ	nD	e	e
ʈ	tD	n	n	æ	aF
ʧ	tsV	ɳ	nX	y	uF
ʈ	tX	ɲ	nP	œ	oF
c	tP	ŋ	gN	ɨ	iC
k	k	w	w	ə	eC
ʔ	q	l	l	ʌ	ac
p^h	pA	r	r	a	a
t^h	tA	j	y	u	u
$ʈ^h$	tDA			o	o
$ʧ^h$	tsVA				
$ʈ^h$	tXA				
c^h	tPA				
k^h	kA				
b	b				
d	d				
ɖ	dD				
ʤ	dzV				
ɖ	dX				
ɟ	dP				
g	g				
f	f				
s	s				
θ	sD				
ʃ	sV				
ʂ	sX				
ɕ	sP				
x	x				
v	v				
z	z				
ð	zD				
ʒ	zV				
ʐ	zX				
ʑ	zP				
ɣ	gS				

Consider the following examples, which demonstrate these problems and their resolution. When given the pairs of words, *api::api*, ALINE returns the score 65. However, *api::apila* also scores 65, because the last syllable is unpaired. Furthermore, the ALINE score has no maximum

Table 4. Exemplar data loss due to lack of ALINE encodings for uvular place of articulation.

IPA	ALINE	IPA	ALINE
k	⟨k⟩	q	⟨k⟩
g	⟨g⟩	G	⟨g⟩
x	⟨x⟩	χ	⟨x⟩
ɣ	⟨gS⟩	ʁ	⟨gS⟩
ŋ	⟨gN⟩	N	⟨gN⟩

Table 5. Possible solution to data loss problem.

IPA	ALINE	IPA	ALINE
k	⟨k⟩	q	⟨kU⟩
g	⟨g⟩	G	⟨gU⟩
x	⟨x⟩	χ	⟨xU⟩
ɣ	⟨gS⟩	ʁ	⟨gUS⟩
ŋ	⟨gN⟩	N	⟨gUN⟩

value, because the number of aligned sequences between any two words has no theoretical maximum. A very short pair of highly aligned words, such as *api::apik* (score = 65), may receive a lower score than a pair of long, but dissimilar words, such as *kalarita::makebela* (score = 75). Put another way, the ALINE algorithm cannot distinguish two identical words: *pu::pu* (score = 50) and *tausebasai::tausebasai* (score = 230).

Kondrak (1999) suggested normalizing the ALINE score by division with the length of the longest word and multiplication with the maximum possible similarity score between word segments. Here, we use a slightly different approach; we normalize the word-pair score by the average score of word self-comparisons:

$$s_{\text{norm}} = \frac{2s}{s_1 + s_2} \tag{1}$$

where s is the ALINE word-pair score, and s_1 and s_2 are the ALINE scores for each word compared to itself. This similarity score directly represents how similar two words are to each other. It ranges from 0–1, and converges to 1 as two words become more similar. Functionally, this normalization also accounts for length differences and penalizes

misaligned word portions. To convert this similarity score to a true distance (i.e. dissimilarity) statistic, we define the ALINE distance:

$$d_{\text{ALINE}} = 1 - \frac{2s}{s_1 + s_2} \tag{2}$$

The effect of these manipulations is summarized in Table 6.

The ALINE distance reaches its maximum at 1, and in some cases (such as comparison with geographic distance) it may be preferable to use:

$$d = -\ln\left(\frac{2s}{s_1 + s_2}\right) \tag{3}$$

Table 6. Examples of ALINE scores before and after normalization.

Word comparison	Raw ALINE score	ALINE distance	ALINE output
api::api	65	0	| a p i | | a p i |
apik::apik	100	0	| a p i k | | a p i k |
apila::apila	115	0	| a p i l a | | a p i l a |
api::apik	65	0.212	| a p i | - | a p i | k
api::apila	65	0.278	| a p i | - - | a p i | l a
apik::apila	65	0.395	| a p i | l a - | a p i | - - k
kalarita::kalarita	200	0	| k a l a r i t a | | k a l a r i t a |
kalara::kalara	150	0	| k a l a r a | | k a l a r a |
makebela::makebela	200	0	| m a k e b e l a | | m a k e b e l a |
kalarita::kalara	142.5	0.186	| k a l a r a | - - | k a l a r i | t a
kalarita::makebela	75	0.625	- - | k a - - l a | r i t a m a | k e b e l a | - - - -
kalara::makebela	82	0.531	- - | k a l a r a | m a | k e b e l a |

which ranges from 0 to ∞. We note that the Taylor expansion shows these distances measures are nearly equivalent for small values (approximately 0–0.6). However, in cases where two languages are very similar, this transformation will separate similar values and may yield better discrimination.

To scale multiple word-pair scores to inter-language divergence, we define the ALINE distance between two languages as the average of ALINE distances between corresponding word pairs. Intuitively, this distance can be represented as how similar, on average, word pairs are between the two languages. Although a weighted average may prove better for certain applications, we have no *a priori* reason to weight some word pairs more than others. Although we tried several normalization schemes in the construction of this distance statistic, we found this version performed best. Nevertheless, other normalizations and weighting schemes may favour certain applications. It also may prove useful to penalize dissimilarity in a more sophisticated way, such as weighting the location of dissimilar word portions or variably penalizing different types of word misalignment.

We realize that in transforming the raw ALINE similarity score into a distance, we are departing from Kondrak's original work. However, by using the ALINE algorithm unaltered and normalizing it with information from self-comparisons, we believe our approach preserves the advantages of a similarity score, as outlined by Kondrak (1999), while resulting in a more useful measurement of phonetic distance.

AUTOMATION

The ALINE algorithm is fast, and thus suitable for application to large datasets. However, the ALINE input file specification is not designed to handle more than two languages simultaneously. For our purposes, we used a database written in FileMaker Pro 8.0 to store word lists and language information, and to provide an interface for manual editing of word lists using a Unicode-compliant IPA character set. The database includes a lookup table and automatically generates the ASCII-compliant feature encoding scheme of ALINE. Figure 1 shows the steps required to transform paperbound Swadesh lists into ALINE distance matrices. Perl v5.8.6 scripts were used to prepare ALINE input files from a database export file, to batch input files for the ALINE executable, to

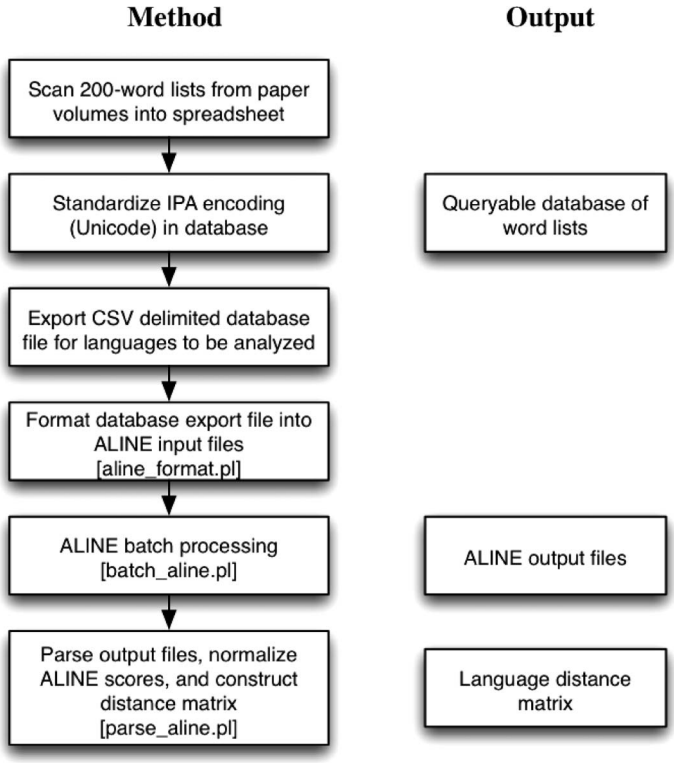


Fig. 1. Flow chart illustrating the methodology used to create a phonetic distance matrix. Products of the analysis are listed under ''Output''. Perl scripts used to automate steps are indicated in brackets.

parse ALINE output files into distance matrices, and to automate other downstream analyses. All of the word-pair score normalization and matrix-building options described herein are available through command-line interfaces to these Perl scripts.

USING THE ALINE DISTANCE: VALIDATION AND APPLICATION

Study Area and Data Sources

Our methodology was developed initially as part of an interdisciplinary project to investigate the joint history of languages and genes in

Austronesian-speaking communities across the Indonesian archipelago. To look for interesting patterns and correlations, our team collected hundreds of genetic samples and built a database of 200-word Swadesh lists for over 400 languages collected across the country by trained linguists. Many of the lists come from a series of sixteen volumes assembled by *Pusat Bahasa Departemen Pendidikan Nasional*, the Indonesian National Linguistic Centre in Jakarta. The sheer volume of these data is such that a through linguistic analysis is time prohibitive without computational assistance.

To examine the performance of the ALINE distance, we applied it to three goals, each focusing on a subset of our Indonesian language dataset:

1. We examined a reconstructed chronosequence of words from the village of Rindi, on the eastern Indonesian island of Sumba.
2. We constructed a language phylogeny for eighteen wordlists from Sumba.
3. We created a language classification map for Austronesian and non-Austronesian languages for the islands of Flores and Alor, also in eastern Indonesia.

These applications serve as a validation of the relationships reconstructed by the ALINE distance.

Chronosequence

Any reasonable linguistic distance should reflect and quantify historical linguistic change. Unlike cognate-based approaches, the ALINE distance can distinguish differences within cognate groups. In our case, the languages of Sumba all belong to the geographically extensive Austronesian family, which includes over a thousand languages distributed from Madagascar to Polynesia (Gordon, 2005). These languages trace their origin to proto-Austronesian (PAn) through a long chain of proto-languages, including new reconstructions of proto-Sumba (PS) and proto-East-Sumba (PES) derived from our own data. Many words within the Sumbanese languages remain cognates with their PAn counterparts even today. Because the evolution of these words is known with some certainty, we can apply the ALINE algorithm to these chronosequences to examine its performance.

The following simplified schema represents the historical evolution of Austronesian languages, beginning with proto-Austronesian and ending with the extant east-Sumbanese language, Rindi:

Proto Austronesian (PAn)
↓
Proto Malayo-Polynesian (PMP)
↓
Proto Central-Eastern Malayo-Polynesian (PCEMP)
↓
Proto Central Malayo-Polynesian (PCMP)
↓
Proto Sumba (PS)
↓
Proto Eastern Sumba (PES)
↓
Rindi

This reconstruction is based on the work of Robert Blust and our own data, and other branches of the accepted Austronesian family tree that are irrelevant to the following discussion are excluded. Previous work has tentatively associated PAn to archaeological evidence dating around 5000 BP, PMP to ~4500 BP, and CEMP to ~4000 BP (Bellwood, 1997, pp. 119–120), with the division into CMP occurring not long thereafter. Our new reconstructions for PS and PES presumably date after this time, but their chronology is not firmly established.

To better understand how the ALINE distance functions as a tool for understanding linguistic change, we applied it to reconstructed sequences of words starting with PAn and ending with modern words from Rindi. Table 7 shows five such chronosequences, and Figure 2 shows each word's trajectory from reconstructed PAn form to modern version. The table and chart show how words have evolved over time, and how ALINE distance renders those changes quantitatively. We emphasize that the x-axis is purely categorical, and does not reflect time other than in terms of general order.

Several trends are immediately apparent in this analysis. Vowel changes seem to have relatively little effect on ALINE distance scores (as expected from Table 1). Consonants have higher weightings than vowels, and consonant change has relatively more effect on overall scores.

Table 7. Exemplar evolution of five words from proto-Austronesian (PAn) to contemporary Rindi.

Node	Date	three	two	one	swim	louse
PAn[1]	5000 BP	təru	duʂa	isa	laŋuj	kuʈu
PMP	4500 BP	təlu	duha	isa	naŋuj	kutu
CEMP	4000 BP	təlu	dua	isa	naŋuj	kutu
CMP	<4000 BP	təlu	dua	isa	naŋuj	kutu
PS	<4000 BP	təlu	duada	isa	naŋi	utu
PES	<4000 BP	tailu	dambu	hau	ŋani	wutu
Rindi	Present	tailu	dambu	hau	ŋieni	wutu

[1]The following substitutions have been made in the traditional PAn inventory: ʈ for C, ʂ for S, l for N, and r for l. In addition, j is substituted for y in conformance with normal IPA usage.

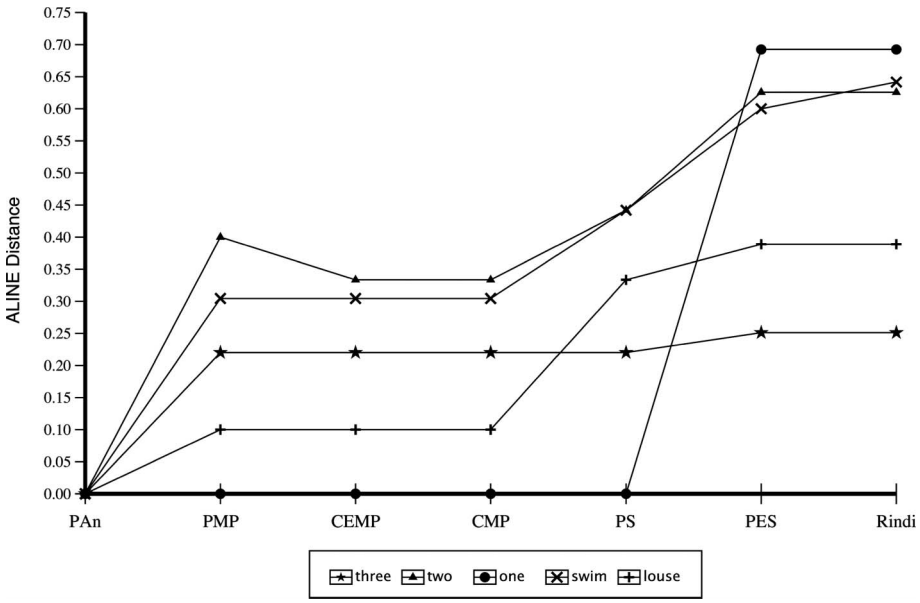


Fig. 2. Word evolution for five glosses based on ALINE distances from proto-Austronesian (PAn) to the contemporary language spoken at Rindi, Sumba, eastern Indonesia. The *x*-axis reflects evolutionary order, but is not directly scaled to time.

Section deletions have the most impact. The relative weighting of these changes is open to interpretation, and it may be preferable to change certain weighting values in future versions of the ALINE software.

However, these results suggest that the ALINE distance is capable of extracting high-level detail from the phonetic information contained in word lists. This is a significant improvement over cognate-based distance measurements, because it allows us to analyze change within cognate classes. Other linguistic distance methods discard this important phonetic information.

Figure 3 shows individual and average historical distances for 40 words from contemporary Rindi. Despite significant variation between individual trajectories, the average indicates an increasing distance measure from PAn to the present. Significant differences may underlie the phonetic evolution of individual words, but the average appears to be an acceptable representation of the overall processes, and matches previous

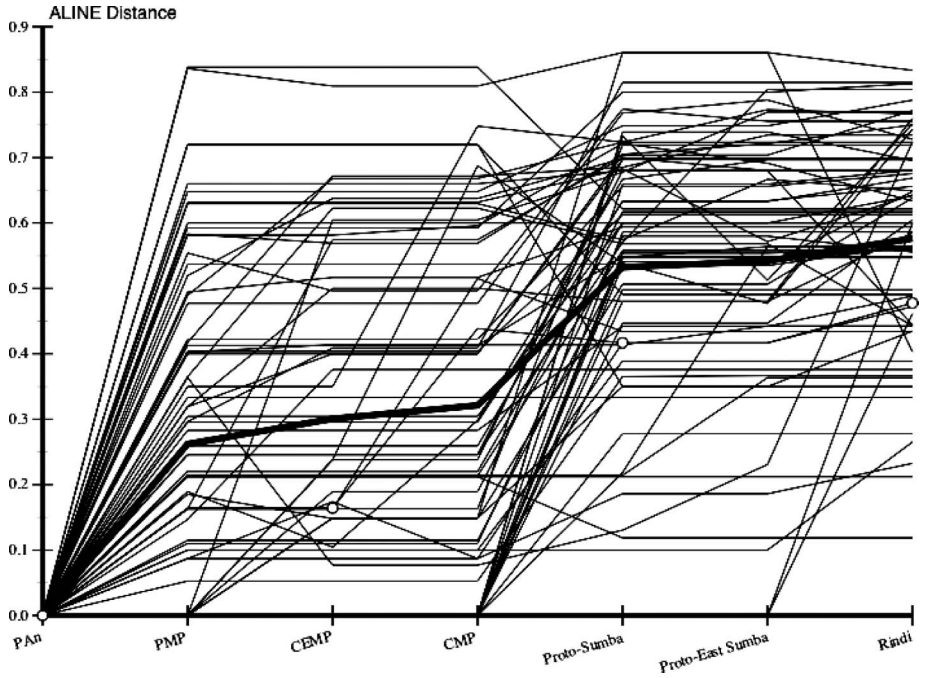


Fig. 3. Plot of the ALINE distance in the historical development of forty words leading to the contemporary language spoken at Rindi. Each thin black line represents the evolution of one word. The thick black line shows the average ALINE score, which shows increasing phonetic distance from PAn despite significantly different evolutionary trajectories for individual words.

work with these (proto-)languages using traditional comparative linguistic approaches.

As an aside, we note that line slopes cannot be compared as proxies for evolutionary rates, because the *x*-axis does not accurately reflect the relative chronology of these proto-languages. However, these results do highlight that phonetic evolution has no constant or average rate. This is clearly indicated in Figure 3, where significant differences appear in the evolutionary trajectories of forty Sumbanese words over the same time period. Nonetheless, the average of these distances suggests that the ALINE distance represents an increasing phonetic distance from proto-Austronesian to the present, as is consistent with the outcome of traditional historical linguistic methods. While some of these differences only reflect the internal weighting scheme implemented by ALINE, other results suggest interesting directions for further work using the ALINE distance to examine how and when words or languages have undergone phonetic change.

Notably, this graph also indicates that the ALINE distance can decrease from PAn to the present, with a word later in the sequence appearing closer to the original form than intermediaries. Four examples are shown in Table 8, where the node where distance measures change is indicated in bold. The trajectories of these words are charted in Figure 4.

These words illustrate the difficulty of reconstructing language evolution, but importantly, are minor contributors to the overall language scores. Although there is clearly room for algorithmic refinement (especially feature-weighting), ALINE satisfactorily evaluates phonetic distances between lexical items. Idiosyncrasies such as these are effectively integrated out at the language level, which is dominated by extensive information content (Figure 3). Combined, these results

Table 8. History from PAn to Rindi of four words that show decrease in the ALINE distance.

Node	leg	kill	right	blow
PAn	qaqaj	paʈaj	kawanal	ʂiup
PMP	qaqaj	bunuq	kawanan	hiup
CEMP	waqaj	bunuq	kawanan	upi
CMP	wai	bunuq	kawanan	upi
PS	wai	**pamate**	**kawana**	pui
PES	**witi**	pamati	kawana	**hapui**
Rindi	wihi	pameti	kawana	hapuji

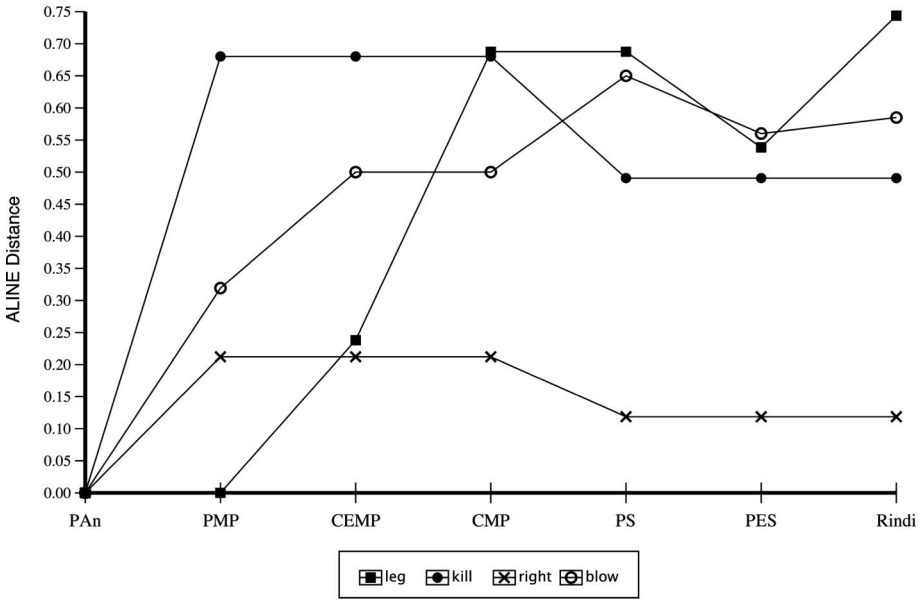


Fig. 4. Incidents of decrease in phonetic distance from PAn to contemporary Rindi.

indicate that the ALINE distance can quantify historical linguistic change in a meaningful way, and in conjunction with standard historical analysis, could highlight interesting phonological cases, or even identify possible coding errors.

Language Phylogeny

One important task of historical linguistics is the construction of language phylogenies showing historical relationships between languages. We constructed a language tree for eighteen languages on Sumba, east Indonesia using the traditional comparative method (Figure 5A). A number of methods have been developed in evolutionary biology to produce trees from distance matrices (Felsenstein, 2004). We examined the performance of the ALINE distance in tree construction by building both UPGMA (unweighted pair group method with arithmetic mean; Figure 5B) and neighbor-joining trees (Figure 5C) from our distance matrices using the Phylip 3.65 program NEIGHBOR (Felsenstein, 2005). We constructed trees using the complete dataset (Figure 5B–C) as well as trees using glosses identified as PAn cognates (not shown). Since the languages still trace a significant portion of their lexicon to

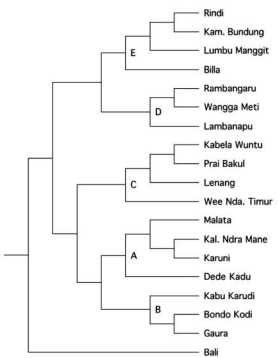
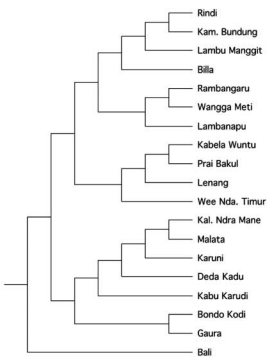
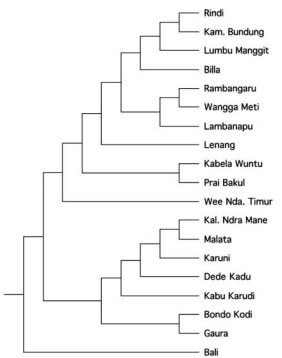
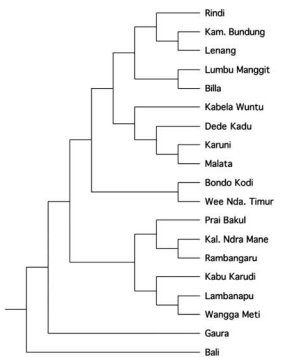
A. Historical Tree	B. ALINE Tree (UPGMA)	C. ALINE Tree (Neighbor-Joining)	D. ALINE Tree (Random)
			
R-F Distance	6	10	32
MAST Distance	6	6	11
P	<.0001	<.0001	.559
% matching nodes/edges	87.9	80.7	46.5

Fig. 5. Side-by-side comparison of four Sumbanese language trees. A. Traditional tree manually constructed by a linguist using the comparative method. B. ALINE tree built using the UPGMA algorithm. C. ALINE tree built using the neighbor-joining algorithm. D. ALINE tree built after words were shuffled within each language. Distances are between the ALINE tree and the traditional tree (A). R-F distance is the Robinson-Foulds symmetric distance. The MAST distance is the number of taxa that need to be removed to get the maximal agreement subtree. The % similarity score calculates the number of matching nodes/edges. Trees were rooted with Bali as an outgroup language, and are shown without branch lengths for easier topological comparison. Branch lengths were not used to calculate distances because they are not available for the historical tree (a limitation of the comparative approach).

PAn, this procedure allowed us to examine the role of a cognate-only subset in determining tree topology.

To assess the robustness of the ALINE distance at the level of language families, we examined trees produced by three Monte Carlo simulation methods. We considered the degree to which ALINE identifies gross phonological similarity between languages by shuffling words and meanings within each language and comparing shuffled word lists. We shuffled words between languages within a given gloss category to produce random 'languages' from our dataset. Finally, to examine the spread of tree topologies present in our distances, we constructed pairwise language distances by sampling with replacement from the 200 individual gloss comparisons and averaging that sample. This process is equivalent to looking at different random weighted means instead of taking the unweighted mean of the 200 word comparisons.

We compared our trees using three methods. The Robinson-Foulds symmetric distance (Robinson-Foulds, 1981) was computed with Phylip's TREEDIST. The MAST distance is the number of branches that must be removed to obtain the maximal agreement sub-tree, and was computed with Phylonet 1.2 (Than et al., 2008). The percentage similarity score is computed with Nye's online phylogeny comparison applet (Nye et al., 2006). We note that the symmetric distance repeats in two-step multiples, so a single branch difference between trees yields a distance of 2. Since all these algorithms produce unrooted trees, we determined the root by using the Balinese language as an outgroup language. Balinese is an appropriate choice because it is also an Austronesian language, but belongs to Western Malayo-Polynesian, a different branch of PMP than that which leads to PS via CEMP. We omitted branch lengths since the historical method does not produce them, a limitation of the standard comparative approach. Figure 5 shows the results of the comparisons below each computed tree and an exemplar random tree. The clade designations A-E (Figure 5A) are based on the original tree constructed through traditional means, and the order of clade lettering follows the order in which divisions are understood to have taken place in actual historical sequence (i.e. group A was the first to diverge from the proto-Sumba unity, group B was the second, and so on).

UPGMA proves to be a more accurate method than neighbour-joining for phylogeny reconstruction, and produces a tree surprisingly close to the historical reconstruction. There are only three differences: it gets the five major subgroups nearly correct, placing Kabu Karudi with Group A

(rather than Group B); it places group C closer to groups D and E (rather than A and B); and it also attaches Kalembu Ndra Mane closer to Malata (than to Karuni within Group A). These differences are minor, and can be quantified by the close distance of this ALINE tree to the historical tree. Although outperformed by UPGMA, the neighbour-joining tree is also close to the historical phylogeny. Given the observed dissimilarities, it is important to note that tree construction is more difficult than it may at first appear, primarily due to the large number of possible tree topologies even for small numbers of terminal taxa. In our case, there are an immense 3.77×10^{21} rooted or 1.01×10^{20} unrooted trees possible for our 19 languages. In this context, the ALINE trees are extremely close to the tree inferred via the standard comparative approach. The MAST distance reflects this with a low probability value; i.e. the likelihood that one would get a tree equally close to the true tree by selecting any tree at random.

The three discrepancies between the traditionally-constructed tree (A) and UPGMA ALINE tree (B) illustrate the difficulties of tree reconstruction – either by computer or by a trained historical linguist. The greatest error in the tree involves the attachment of group C. In the traditional tree, it is attached to the subgroup comprising groups A + B, whereas in the ALINE tree it is attached to the subgroup comprising groups D + E. There is more lexical replacement in the A + B subgroup than in C, therefore resulting in the superficial appearance that C is closer to the D + E subgroup. This similarity is an artefact of the degree of lexical retention in groups C, D and E, which ALINE cannot differentiate from innovations in subgroup A + B. Other criteria, such as shared phonological innovations and an older layer of innovated vocabulary, indicate that C groups closer to A + B than D + E, but that A + B has undergone marked repopulation of its lexicon under the influence of a non-Austronesian stratum in western Sumba.

The second discrepancy is the attachment of Kabu Karudi. Tree A assigns Kabu Karudi to group B, whereas tree B assigns Kabu Karudi to group A. In similar fashion to the first discrepancy, the other group B languages, Gaura and Bondo Kodi, have undergone significant lexical innovation while their parent was still a single language. Kabu Karudi is therefore comparatively more conservative, which gives the impression that it is more similar to the languages of group A (which also often share this conservative vocabulary). Compounding this dilemma is the fact that Kabu Karudi is in close geographical proximity with, and has borrowed

heavily from, Dede Kadu, a group A language. Since ALINE (along with other computational approaches) cannot identify loans, borrower languages are artificially skewed towards donor languages. (We show below that ALINE still handles this situation with relative efficiency).

The third discrepancy is the attachment of Kalembu Ndara Mane to Karuni (tree A) or to Malata (tree B). This is the most difficult discrepancy to assess, as there seems to be evidence supporting both perspectives. Arguments for attaching Kalembu Ndara Mane to Karuni include shared, arbitrary phonological innovations: (Table 9).

However, at least one phonological innovation and several lexical similarities also suggest closeness to Malata (Table 10).

On balance, tree A probably gives the correct grouping on the grounds that idiosyncrasies shared by Kalembu Ndara Mane and Karuni are more obviously derived from a common ancestor. However, it should be noted that this was one of the most difficult sub-grouping assessments to make under the traditional comparative method, and ALINE cannot be faulted for providing an alternative analysis which has so much supporting evidence.

Table 9. Shared characteristics linking Kalembu Ndara Mane and Karuni.

Gloss	English	KNM	Karuni	Malata	Dede Kadu
beri	"give"	jai	jai	jani	jani
tulang	"bone"	ruwiwi	ruwi	ri	ri
hantam	"crush"	tauwa	tauwa	tausa	taʔa
debu	"dust"	pautana	paʔutana	kombura	kobura
lebar	"wide"	gellara	gallara	belleka	belleka
ini	"this"	nawwa	nagha	ne	naj

Table 10. Shared characteristics linking Kalembu Ndara Mane and Malata.

Gloss	English	KNM	Malata	Karuni	Dede Kadu
kanan	"right"	kawano	kawano	kawana	kawana
pohon	"tree"	wasu	ghasu	pu	puʔu
potong	"divide"	loppu	loppu	ropo	katupu
benar	"correct"	hinna takka	hinatakka	ghinnad^{t}o	patub^{t}ana

In summary, two phenomena obscure relationships between languages and lead ALINE to less than optimal reconstructions. Firstly, substantial lexical replacement in some languages or subgroups, sometimes through language contact. This process creates an extra degree of distance that is phylogenetically artificial. Secondly, diffusion of sound changes across the boundaries of genetic language groups. Since sound changes act upon phonemes (the direct units of analysis in ALINE), any changes, which occur across a group of languages will serve to separate affected from unaffected languages. In situations of strictly phylogenetic inheritance, this is unproblematic. However, in language contact situations (as on Sumba), sound changes can diffuse across phylogenetic boundaries. Both situations can confuse the ALINE algorithm.

We also constructed trees using all words and just PAn cognates; the UPGMA tree was unchanged and the neighbour-joining tree was slightly improved. This result was surprising because it is commonly thought that the presence of loans can significantly affect the generation of trees using automated methods. Our results suggest that this tree-building method is robust to the presence of loans to a significant degree, and that under many circumstances, it may be unnecessary to remove loans prior to tree construction. To understand the affect of loans on the accuracy of trees, it would be necessary to perform additional analyses using datasets with differing amount of external influence, and this is planned for future research.

Finally, Figure 6 shows the distribution of symmetric distances from the historical tree to the resampled UPGMA and neighbour-joining trees, as well as random trees. Although the distributions do not include the historical tree (distance = 0), they are remarkably close.

In summary, the ALINE distance performs well when used with distance-based tree-building algorithms, especially UPGMA. Many previous linguistic distances are based on an existing historical tree; here, we can computationally reconstruct the historical tree with reasonable accuracy and no prior knowledge of language relationships. We emphasize, however, that we do not envision the ALINE distance replacing standard comparative methods. Furthermore, refinements of the distance statistic would likely yield better performance. However, in cases where the historical tree is uncertain, the ALINE tree could give a reasonable first guess for a historical linguist attempting a manual reconstruction (thereby providing significant time savings in the case of large datasets), or the ALINE tree could be used as a prior for other computational methods such as that of Atkinson and Gray (2006).

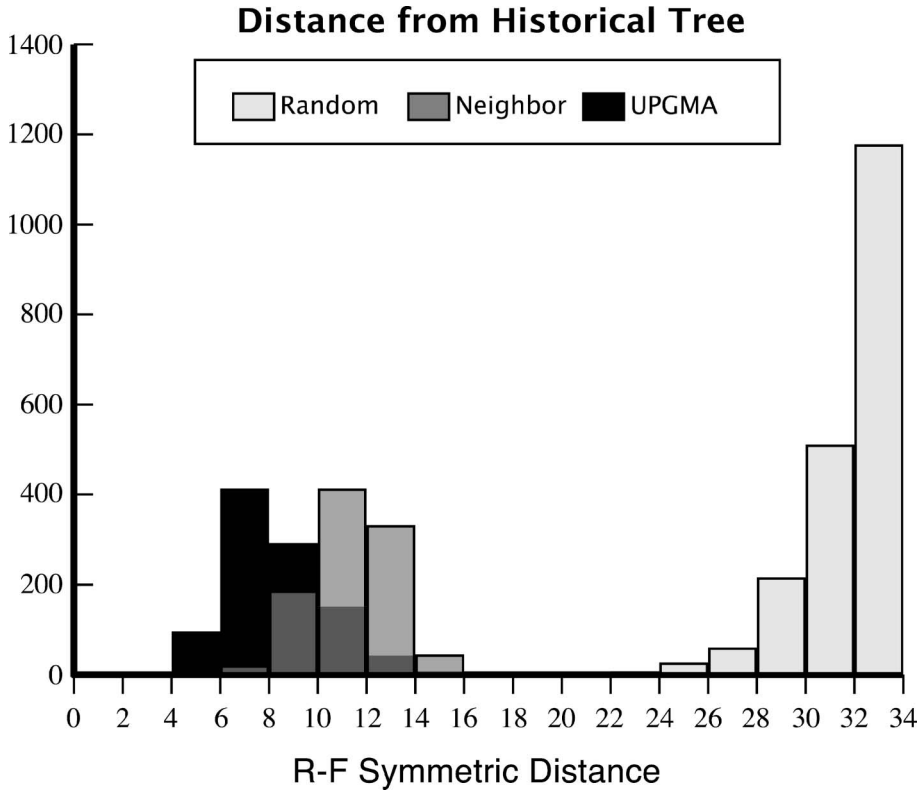


Fig. 6. Robinson-Foulds symmetric distance from historical tree for 4000 randomized and re-sampled trees (1000 in each category). Random (far right) includes trees produced by randomly shuffling words within languages and words with glosses as described in the text. The UPGMA (far left) and neighbour-joining (middle) trees were produced by sampling with replacement from 200 pairwise word comparisons for each language comparison, functionally resulting in using random weighted averages as the language distance. The unweighted averages produced distances of 6 and 10 respectively. The distribution gives an idea of the amount of variance in the data and the overall performance of the method. The additional dark area shows where the two leftmost distributions overlap.

Linguistic Similarity Map

Another useful task in historical linguistics is the creation of maps showing the spatial extent of languages or language families. Comparing maps integrating linguistic, geographic and genetic data is important for understanding how these datasets relate to each another. Of great interest to our project are the islands of eastern Indonesia, where there is a mosaic

of Austronesian and Papuan languages. To examine the spatial distribution of language groups in eastern Flores and Alor using ALINE, we calculated the distance of each contemporary language from PAn and mapped the result (Figure 7). The ALINE distance was log transformed to increase the distance between mean clusters and improve the performance of the interpolation algorithm. Transformed scores were imported into a geographic information system application (ArcMap 9.0) and interpolated by kriging (default options). The interpolation was clipped using an outline of the Indonesian coastline, and the predicted ALINE distances stretched along a grayscale. The mean distances for PAn ($d = 0.715$) and Papuan ($d = 0.935$) clusters were significantly different ($p < 0.001$, $t = 17.02$, two-independent sample t-test, SPSS 12.0.1).

To determine how well the ALINE distance reconstructs boundaries between major language families, we compared the ALINE map to a well-known historical map of the area, the *Language Atlas of the Pacific* (Wurm & Hattori, 1981). The ALINE map captures the main distinction between the Papuan languages to the east, and the Austronesian languages to the west. Wurm and Hattori's map shows an Austronesian-speaking coastal region along the western islands of Alor that is not captured by the ALINE map. However, plotting the location of each language list clearly illustrates that missing data cause this lack of detail.

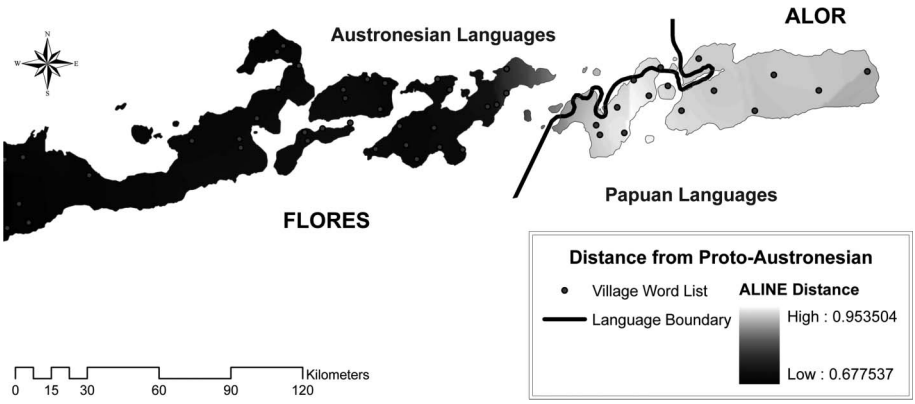


Fig. 7. Interpolated map based on the ALINE distance between proto-Austronesian and 51 200-word lists from Flores and Alor, eastern Indonesia. The ALINE distance was transformed using a log-linear function as described in the text. The boundary between Austronesian and Papuan languages was drawn by Wurm and Hattori (1981) using traditional methods.

Overall, the ALINE distance performs remarkably well in a mapping context. In future, similar maps integrating genetic data will allow us to better understand how genetic and linguistic datasets are interrelated.

DISCUSSION

While computational approaches hold promise for automated processing of hundreds or even thousands of word lists, they are not currently used routinely for this purpose. Several important limitations have restricted the use of computational approaches in historical linguistic analysis. Most importantly, there is no widely accepted model framework describing language change. While some models of sound change do exist, they have not been used to build predictive models of phonological evolution because the historical processes of interaction between languages are often more significant factors shaping the structure of languages. Computational models are by nature simplifications of the real-world processes they are designed to emulate, and as such, will always exclude information. In our analysis, for example, ALINE uses only a subset of all the possible phonological features observed in our word lists, and this may result in missing key evidence for linkages between languages that a traditional historical analysis might discover. We also did not have access to comparative syntactical information for any of the languages under consideration, and therefore syntax could not be incorporated into either the computational or traditional analyses. Other limitations of the computational approach include the presence of loans, short words and internal cognacy. Manual analysis by a linguist using traditional methods can avoid many of the problems associated with automated methods, but even the best historical linguist is limited by the amount of information that they can process within a given timeframe.

Despite limitations, computational methods permit large-scale comparisons of languages that must remain beyond the reach of traditional techniques. This is particularly true when the intention is to combine historical linguistic analysis with other data, such as from genetics, geography or archaeology. For this kind of combined analysis of extensive datasets, the problem of scale looms large. For example, an interdisciplinary project may begin with a partial Mantel test that correlates matrices of geography, genetic and linguistic distances. The number of cells in each distance matrix increases quadratically as sample

locations are added. A distance matrix of 18 Sumbanese languages, based on a 200-word Swadesh list, requires 30,600 word comparisons (153 pairwise language comparisons, 200 words per comparison), and is amenable to traditional analysis. Conversely, our complete dataset includes word lists for 405 Indonesian languages, and the resulting distance matrix requires 16.3 million word comparisons. Such comparisons are beyond standard linguistic approaches, whereas the method discussed here can easily handle such sizable datasets.

Our hope in adapting ALINE as a measurement of phonetic distance is to develop a computational method that can move beyond cognate-only approaches by preserving as much of the rich store of phonetic information as possible. Thus, we emphasize that the results discussed here do not hinge on identification of cognates or loans. Our maps and trees are based purely on phonological similarity, and the ALINE algorithm has no prior information about historical relationships of words or languages. This is both a strength and weakness of our methodology. The weakness is that our analysis can potentially be biased by the presence of loans and borrowing (although, at least on Sumba, we have demonstrated that the bias introduced by loans is minimal). The strength is this method's computational basis, its scalability, and its close correspondence to results from more traditional linguistic approaches.

CONCLUSIONS

The method presented here suggests that the ALINE distance can be used as a ratio scale measurement of the dissimilarity between shared-meaning word pairs. Our method is based on a normalized version of the ALINE algorithm, and represents an absolute phonetic distance between 0 and 1. The method is suitable for analysing large datasets because it can be fully automated, runs quickly, and can easily be scaled as more data becomes available. The principle output of our method is a matrix of distances between language pairs. These linguistic distances are superior to previous quantifications, and are suitable for direct comparison with genetic and geographic distances. We have shown that this method can be used to generate language trees and classification maps, and we have validated these results using traditional, manual analysis via the comparative method. Monte Carlo simulation and bootstrapping results add statistical support to our claim that the language phylogeny based on

the ALINE distance is robust. In addition, results to be published elsewhere use the ALINE distance in relation to genetic and geographic pairwise distances (Lansing et al., 2007).

The ALINE method is proven to be useful for detecting patterns of phonetic similarity at multiple scales. At the level of 18 languages and a single island, we were able to faithfully assemble a language tree with the same number of major sub-groupings as was found independently using the comparative method. At a regional scale, we have presented a language classification map for a portion of the Indonesian Archipelago spanning Komodo, Flores, and Alor, and were able to detect the major language groups suggested by previous work using the comparative method. The results discussed in this paper suggest that the ALINE distance will be a useful, and indeed necessary, method for analysing large language datasets as these increasingly become available.

ACKNOWLEDGEMENTS

The authors would like to thank Grzegorz Kondrak for his technical assistance and for the use of the ALINE source code. John Schoenfelder provided the spatial information necessary to create the maps, and Gary Christopherson reviewed the interpolation procedure. Joseph Watkins provided important feedback on drafts of this paper. The sometimes tedious task of scanning, entering, and formatting the Indonesian words into IPA was conducted by Eleanor McCallum and Abby Dowling. This research was supported by the National Science Foundation, the James McDonnell Foundation Robustness program at the Santa Fe Institute, and the Eijkman Institute for Molecular Biology, Jakarta Indonesia. Swadesh word lists for Sumbanese languages were provided by the National Language Center of the (Indonesian) National Department of Education.

REFERENCES

Atkinson, Q. D., & Gray, R. D. (2006). Are accurate dates an intractable problem for historical linguistics? In C. P. Lipo, M. J. O'Brien, M. Collard & S. J. Shennan (Eds), *Mapping Our Ancestors* (pp. 269–296). New Brunswick: Aldine Transaction.

Barbujani, G., & Sokal, R. R. (1990). Zones of sharp genetic change in Europe are also linguistic boundaries. *Proceedings of the National Academy of Sciences USA*, *87*, 1816–1819.

Bellwood, P. (1997). *Prehistory of the Indo-Malaysian Archipelago*. Honolulu: University of Hawai'i Press.

Cavalli-Sforza, L. L., Piazza, A., Menozzi, P., & Mountain, J. (1988). Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences USA*, *85*, 6002–6006.

Cavalli-Sforza, L. L., Minch, E., & Mountain, J. L. (1992). Coevolution of genes and languages revisited. *Proceedings of the National Academy of Sciences USA*, *89*, 5620–5624.

Cavalli-Sforza, L. L. (1997). Genes, peoples, and languages. *Proceedings of the National Academy of Sciences USA*, *94*, 7719–7724.

Chen, J., Sokal, R. R., & Ruhlen, M. (1995). Worldwide analysis of genetic and linguistic relationships of human populations. *Human Biology*, *67*, 595–612.

Covington, M. A. (1996). An algorithm to align words for historical comparison. *Computational Linguistics*, *22*, 481–496.

Covington, M. A. (1998). Alignment of multiple languages for historical comparison. *Proceedings of COLING–ACL'98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (pp. 275–280). San Francisco: Morgan Kaufmann Publishers.

Cox, M. P. (2003). *Genetic patterning at Austronesian contact zones*. Unpublished PhD thesis, University of Otago, New Zealand.

Cox, M. P., & Lahr, M. M. (2006). Y-Chromosome diversity is inversely associated with language affiliation in paired Austronesian- and Papuan-speaking communities from Solomon Islands. *American Journal of Human Biology*, *18*, 35–50.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland: Sinauer Associates Inc.

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.65. Retrieved August 18, 2008, from http://evolution.genetics.washington.edu/phylip.html

Gordon, R. G., Jr. (2005). *Ethnologue: Languages of the World*. Dallas: SIL International.

Heeringa, W., Kleiweg, P., Gooskens, C., & Nerbonne, J. (2006). Evaluation of string distance algorithms for dialectology. In J. Nerbonne & E. Hinrichs (Eds), *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics* (pp. 51–62). Sydney: Association for Computational Linguistics.

Hunley, K., & Long, J. C. (2005). Gene flow across linguistic boundaries in Native North American populations. *Proceedings of the National Academy of Sciences USA*, *102*, 1312–1317.

Kondrak, G. (1999). Alignment of phonetic sequences. Technical Report CSRG–402. University of Toronto. Retrieved August 18, 2008 from ftp://ftp.cs.toronto.edu/pub/reports/csri/402/

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics*. San Francisco: Morgan Kaufmann Publishers.

Kruskal, J. (1983 [1999]). An overview of sequence comparison. In D. Sankoff & J. Kruskal (Eds), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* (pp. 1–44). Stanford: CSLI Publications.

Lansing, J. S., Cox, M. P., Downey, S. S., Gabler, B. M., Hallmark, B., Karafet, T. M., et al. (2007). Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proceedings of the National Academy of Sciences*, *104*(41), 16022–16026.

Lebart, L., & Rajman, M. (2000). Computing similarity. In R. Dale, H. Moisl & H. Somers (Eds), *Handbook of Natural Language Processing* (pp. 477–505). Basel: Dekker.

Nerbonne, J., Heeringa, W., & Kleiweg, P. (1999). Comparison and classification of dialects. *Proceedings of the 9th Meeting of the European Chapter of the Association for Computational Linguistics* (pp. 281–282). Bergen: EACL. Retrieved August 18, 2008 from http://www.let.rug.nl/∼heeringa/dialectology/papers/eacl99.pdf

Nettle, D., & Harriss, L. (2003). Genetic and linguistic affinities between human populations in Eurasia and West Africa. *Human Biology*, *75*, 331–344.

Nye, T., Liò, P., & Gilks, W. (2006). A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics*, *22*, 117–119.

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*, 131–147.

Smouse, P. E., & Long, J. C. (1992). Matrix correlation analysis in anthropology and genetics. *American Journal of Physical Anthropology*, *35*, 187–213.

Sokal, R. R. (1988). Genetic, geographic, and linguistic distances in Europe. *Proceedings of the National Academy of Sciences USA*, *85*, 1722–1726.

Somers, H. L. (1998). Similarity metrics for aligning children's articulation data. *Proceedings of COLING–ACL'98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (pp. 1227–1231). San Francisco: Morgan Kaufmann Publishers.

Than, C., Ruths, D., & Nakleh, L. (2008). Phylonet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, *9*(322).

Wurm, S. A., & Hattori, S. (1981). *Language Atlas of the Pacific Area*. Canberra: Australian Academy of the Humanities (in collaboration with the Japan Academy). Part I 1981 and Part II 1983.