Levenshtein Distances Fail to Identify Language Relationships Accurately

Simon J. Greenhill*
The University of Auckland

The Levenshtein distance is a simple distance metric derived from the number of edit operations needed to transform one string into another. This metric has received recent attention as a means of automatically classifying languages into genealogical subgroups. In this article I test the performance of the Levenshtein distance for classifying languages by subsampling three language subsets from a large database of Austronesian languages. Comparing the classification proposed by the Levenshtein distance to that of the comparative method shows that the Levenshtein classification is correct only 40% of the time. Standardizing the orthography increases the performance, but only to a maximum of 65% accuracy within language subgroups. The accuracy of the Levenshtein classification decreases rapidly with phylogenetic distance, failing to discriminate homology and chance similarity across distantly related languages. This poor performance suggests the need for more linguistically nuanced methods for automated language classification tasks.

1. Introduction

There are around 7,000 languages spoken in the world today (Lewis 2009) that trace thousands of years of human cultural and linguistic evolution. Historical linguistics discovers the relationships between these languages using the comparative method to identify systematic sound correspondences between words in different languages. These correspondences distinguish the words that have descended from a common ancestor of the languages (i.e., are "cognate"), and are representative of the phonological or morphological innovations in the shared history of those languages (Durie and Ross 1996). Take, for example, the following words meaning "three": Javanese *telu*, Mussau *tolu*, and Tongan *tolu*. One could therefore postulate that the two Oceanic languages, Mussau and Tongan, are more closely related to each other than to Javanese (in schematic form: Mussau,Tongan|Javanese) following a merger of Proto-Malayo-Polynesian **e/-aw* into Proto-Oceanic **o* (Blust 2009).

Correctly identifying the relationships between these languages is not an exercise in mere phylogenetic classification but a tool for investigating human prehistory. As an example, linguists have used the comparative method to trace the spread of the 1,200 or so Austronesian languages across the Pacific back to Taiwan (Blust 1999; Pawley 2002). Recently, computational phylogenetic methods from evolutionary biology have been

* Computational Evolution Group, The University of Auckland, Auckland 1142, New Zealand. E-mail: s.greenhill@auckland.ac.nz.

Submission received: 27 October 2010; revised submission received: 18 February 2011; accepted for publication: 14 March 2011.

applied to this lexical cognate data. For example, we tested different scenarios of Pacific settlement and found compelling support for an origin of the Austronesian language family in Taiwan around 5,200 years before present (Gray, Drummond, and Greenhill 2009). This combination of computational phylogenetic methods and linguistic data promises to be a powerful way of exploring human prehistory and linguistic and cultural evolution (Greenhill, Blust, and Gray 2008; Currie et al. 2010; Gray, Bryant, and Greenhill 2010).

Using the comparative method to identify phylogeny requires a lot of lexical data, detailed knowledge about phonology (in general, and in the languages in question), and a lot of time (Durie and Ross 1996), however. It has recently been suggested that the Levenshtein distance can be used to subgroup languages without the intensive time requirement of the comparative method (Brown et al. 2007) or the potential subjectivity involved in identifying sound correspondences (Serva and Petroni 2008). The Levenshtein distance (Levenshtein 1966) is a string comparison metric that counts the number of edit operations (replacements, insertions, and deletions) required to transform one string into another (Kruskal 1983). For example, the Levenshtein distance between the Mussau and Tongan cognate words *tolu* is 0, and the difference between *tolu* and Javanese *telu* is 1 (a replacement of the /e/ with /o/). This is usually normalized by dividing by the length of the longest word (Brown et al. 2007; Serva and Petroni 2008), so the distance between *tolu/telu* is 0.25.

This Levenshtein classification has recently been applied to the Indo-European (Serva and Petroni 2008; Tria et al. 2010), Austronesian (Petroni and Serva 2008), Turkic (van der Ark et al. 2007), Indo-Iranian (van der Ark et al. 2007), Mayan, Mixe-Zoque, Otomanguean, Huitotoan-Ocaina, Tacanan, Chocoan, Muskogean, and Austro-Asiatic language families (Brown et al. 2007; Holman et al. 2008; Bakker et al. 2009). The results of Levenshtein classification have even been used to explore broader questions such as the relationship between population size and the rates of language evolution (Wichmann and Holman 2009), the dates of human population expansions (Serva and Petroni 2008; Wichmann and Holman 2009; Wichmann et al. 2010), whether languages arise and go extinct at a constant rate (Holman 2010), and to triangulate the homelands of language families (Wichmann, Müller, and Velupillai 2010).

Proponents of Levenshtein classification have claimed that the results are very similar to that of the comparative method. However, to date there has been no rigorous attempt to quantify the performance of the Levenshtein distance at classifying languages. For example, Petroni and Serva (2008) use the Levenshtein distance to classify 50 Austronesian languages, and claim that their obtained language phylogeny is "similar" to the results of the comparative method. However, close inspection of their classification reveals some puzzling incongruities. Their tree correctly places the Atayalic subgroup of the Formosan languages at the base of the tree (Blust 1999), but the next subgrouping on the tree is the large Oceanic subgroup before the rest of the Formosan languages are encountered. According to the comparative method, the Austronesian language family has a highly rake-like structure, and the Oceanic subgroup should be nested within Central-Eastern Malayo-Polynesian and Eastern Malayo-Polynesian (Blust 1993, 2009; Gray, Drummond, and Greenhill 2009). Instead, the Levenshtein classification looks more like the results from a lexicostatistical analysis (Swadesh 1952; Embleton 1985), which incorrectly inferred the base of the Austronesian phylogeny to be in Near Oceania (Dyen 1965). This alternative tree topology has been largely discounted as a methodological error due to an inability of the lexicostatistical methodology to handle differences in the rates of lexical change (Bergsland and Vogt 1962; Blust 2000; Greenhill and Gray 2009).

The incongruities between the Levenshtein classification and the comparative method demonstrates the need for a quantitative evaluation of the accuracy of the Levenshtein distance for genealogically subgrouping languages. The large Austronesian language family is a good test case for evaluating the accuracy of comparative methods (Greenhill and Gray 2009; Greenhill, Drummond, and Gray 2010). First, the major subgroups of the Austronesian language family are well established (Dempwolff 1934, 1937, 1938; Grace 1959; Pawley 1972; Dahl 1973; Blust 1978, 1991, 1993, 1999, 2009; Ross 1988; Ross and Næss 2007). Second, most subgroups in the Austronesian family can be identified from the basic vocabulary commonly used for Levenshtein classification (Greenhill, Drummond, and Gray 2010). Third, there is a large-scale database of Austronesian vocabulary available for such a test: the Austronesian Basic Vocabulary Database (Greenhill, Blust, and Gray 2008). In this article, I attempt to evaluate the accuracy of the Levenshtein classification method for identifying genealogical subgroups.

2. Method

The Austronesian Basic Vocabulary Database (ABVD) (Greenhill, Blust, and Gray 2008) is a large collection of basic vocabulary word lists from over 650 Austronesian languages. Each word list comprises 210 items such as words for body parts, kinship terms, colors, numbers, verbs, nouns, and so forth. These items are thought to be highly stable over time and resistant to being borrowed between languages (Swadesh 1952). From this database I extracted the word lists for 473 languages that belonged to the Austronesian language family and were not reconstructed proto-languages.

To compare the performance of the Levenshtein classification to the traditional subgroupings I subsampled triplets of languages from these data 10,000 times. The correct classification of triplets is the simplest possible subgrouping task, with only four possible subgroupings: language A is more similar to language B than C (A,B|C), A is closer to C (A,C|B), B is closer to C (B,C|A), and no language is closer to the other (A,B,C). Restricting the comparison to triplets has the advantage of not requiring a tree construction algorithm to infer the topology, and hence avoids adding uncertainty caused by the phylogenetic reconstruction process (Susko, Inagaki, and Roger 2004).

For each language triplet I obtained the expected linguistic subgrouping from the Ethnologue database (Lewis 2009). The Ethnologue is the primary catalogue of language information about the world's languages. Each language has a classification string associated with it that categorizes the language into a nested set of subgroups that are derived from primary historical linguistic research. It could be argued that the Ethnologue classification lags behind linguistic research (Campbell and Grondona 2008). However, the deeper structure of the Austronesian language family has been well established for a long time (Dempwolff 1934, 1937, 1938; Grace 1959; Pawley 1972; Dahl 1973; Blust 1978, 1991, 1993, 1999, 2009; Ross 1988), and the recent release of the Ethnologue has updated the classification to match even quite newly identified language subgroups like Temotu (Ross and Næss 2007).

The normalized Levenshtein distance between each language pair in the three-language sample was calculated as follows. In each of the 210 words in the ABVD word lists, one entry was selected for each language. Where the ABVD had multiple entries for a word in a language one of the entries was sampled at random. The Levenshtein distance was then calculated for each pair of words (word 1 in language A vs. word 1 in language B; word 2 in language A vs. word 2 in language B; etc.), and normalized by dividing by the length of the longer word (Brown et al. 2007; Holman et al. 2008; Serva and Petroni 2008). When one of the language pairs had no entries for a given word the

word pair was ignored. The average normalized Levenshtein distance for all 210 words between each pair of languages in the three-language sample was taken as a measure of relatedness between the language pairs. These averaged distances were then used to select the most appropriate of the four possible subgroupings (e.g., A,B|C, or A,C|B or B,C|A or A,B,C). The subgrouping estimated from the Levenshtein distances was then evaluated as correct when it matched the subgrouping obtained from the Ethnologue.

A potential problem with this large-scale comparison is that the language data in the ABVD have varying orthographies. These differences in orthographies might play a crucial role in the accuracy of the Levenshtein distance to correctly infer language relationships. The ABVD contains word lists from many different sources of data, but there are some subsets that contain consistent orthography. To assess the effect of orthography on the Levenshtein distance comparison, I repeated the given procedure while sampling from three different subsets of languages. The first of these subsets is composed of word lists for 121 languages across the Austronesian region obtained from the linguist Robert Blust (Blust 2000, 2009), and these are transcribed in Blust's standard orthography. In contrast, the second two subsets are composed of detailed language surveys of 91 languages from the Solomon Islands described in Tryon and Hackman (1983), and 45 Philippines languages described in Reid (1971).

3. Results

The average Levenshtein distance between each sampled pair of languages in the full 473-language sample was 0.847 (s.d., 0.063). In the Blust language subsample the average Levenshtein distance was 0.846 (s.d., 0.047). The average distance for the Tryon and Hackman (1983) and Reid (1971) languages was 0.772 (s.d., 0.120) and 0.771 (s.d., 0.121), respectively. Cronbach's alpha (Nerbonne and Heeringa 2009) calculated on all inter-item distances between 1,000 random language pairs showed an alpha of 0.99, suggesting that the signal measured by the Levenshtein distance was consistent.

The ability of the Levenshtein distance to identify the correct subgrouping from the three possible combinations was low. In the full 473-language subsample the Levenshtein distance correctly identified 41.3% of the subgroupings. In the Blust language subsample the accuracy was 36.9%. In the Tryon and Hackman (1983) and Reid (1971) languages, the accuracy was higher at 65.8% and 64.3%, respectively.

It may be unfair to penalize the Levenshtein classification for resolving a subgroup that Ethnologue did not. Discarding the comparisons when the expected classification was unresolved (A,B,C) increased the accuracy marginally: 47.8% Full data set, 44.7% Blust, 77.0% Tryon and Hackman (1983), and 76.8% Reid (1971).

4. Discussion

The performance of the Levenshtein distance at classifying the Austronesian languages using basic vocabulary data was poor, with the correct subgroup chosen only 41.3% of the time. The more standardized orthography in the Solomon Islands and Philippines language subsets increased the accuracy of the classification to around 65%. This was not the case for the Blust subsample, however, which scored the lowest accuracy of 36.90%. The relatively poor performance on the Blust subsample when compared to the Solomons and Philippines samples suggests that further standardization might not increase accuracy. The Blust languages are sampled right across the range of Austronesian languages. In contrast, the other subsets are sampled from distinct regions and contain languages within a restricted set of subfamilies (e.g., Temotu, Meso-Melanesian, and

South-East Solomonic for the Solomon Islands subset; and Greater-Central Philippines and North Luzon for the Philippines subset).

The major limiting factor for classification accuracy using the Levenshtein distance appears to be phylogenetic distance: how closely the languages are related. Figure 1 plots the average number of Ethnologue classification nodes in each three-language subset (Treelength) against the proportion of correct Levenshtein classifications for the two language subsets spanning the largest range of subgroups (the Full data set, and the Blust subsample). This figure shows that as the phylogenetic distance between languages increases, the Levenshtein distance is less accurate at choosing the correct classification. This finding is concordant with studies that have applied the Levenshtein distance to dialect-level data and have generally shown strong congruence between the subgroupings estimated by traditional dialectology methods and the Levenshtein distance (Kessler 1995; Gooskens and Heeringa 2004; Heeringa et al. 2006; Nerbonne and Heeringa 2009). Interestingly, Figure 1 hints at an increase in accuracy with very large phylogenetic distances, perhaps suggesting a potential role for the Levenshtein distance for discriminating between different language families (van der Ark et al. 2007).

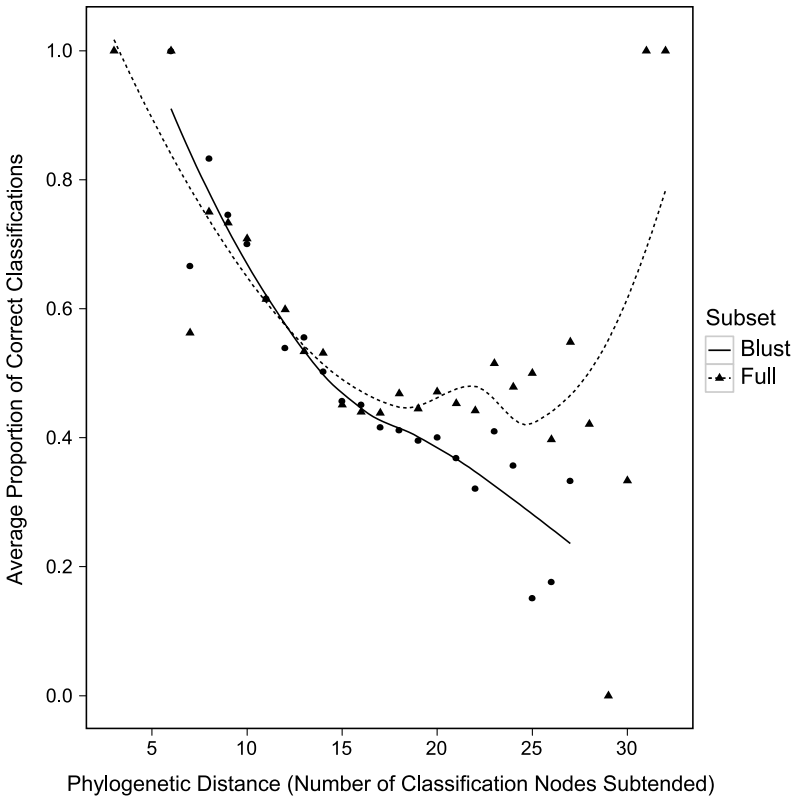

Figure 1
Scatter plot showing the accuracy of the Levenshtein classification approach as a function of phylogenetic distance. Phylogenetic distance is measured by the average number of Ethnologue classification nodes subtended by each language triplet. The points are drawn from the two language subsets spanning the largest range of subgroups (the full data set and the Blust subsample) with LOESS curves of best fit (Full data set: triangles, dotted line; Blust data set: circles, line).

Why is the accuracy of the Levenshtein method so low? The major cause of this poor performance is that the Levenshtein distance is linguistically naive in at least four ways. First, the Levenshtein distance blurs the distinction between cognate and non-cognate words. Dialect studies generally explore the change within a cognate set where the Levenshtein distance between entries is generally one or two character changes. In contrast, when classifying languages, the metric conflates two different processes: change within a cognate set and change between cognate sets. The distance between two cognate words (e.g., *tolu* and *telu*) is small (0.25), however, calculating the distance between two different cognate sets *tolu* and (for example) Marovo *hike* gives a maximally different string comparison. When words are very different, then the Levenshtein distance is more likely to reflect chance similarity.

Second, the Levenshtein distance identifies the surface similarity between words. Historical linguistics is skeptical about surface similarity for geneaological classification (Ringe 1992; Durie and Ross 1996), however, as surface similarity could reflect borrowing, sound symbolism, onomatopoeia, nursery forms, or chance instead of phylogenetic relationships (Campbell and Poser 2008).

Third, processes like metathesis, reduplication, and fossilization of affixes can involve multiple character differences but only occur as one change. For example, Malay *takut* 'to fear' is cognate with Proto-Malayo-Polynesian **ma-takut* 'fearful, afraid' (Blust 2009). The *ma- is a stative prefix that has a tendency to become fossilized (Blust 2009). Rather than a single change, the Levenshtein distance would represent this as two or three insertions/deletions. The third issue is that, under the Levenshtein, all phonological changes are equally likely and occur at the same rates. In reality, some changes occur very rarely whereas others occur frequently and repeatedly (e.g. /t/ to /k/ is thought to have occurred independently at least 20 times within Austronesian [Blust 2004]). In a good model, these frequent changes should have lower penalties than other changes that are more unlikely.

A final cause of this poor classification performance is a result of subgrouping languages according to an overall distance metric. Intuitively, clustering according to minimal distance makes sense, although this has two consequences. First, distance metrics ignore the distinction between forms that are common retentions from an ancestral language, and the forms that are shared innovations in a set of languages. This distinction—a "fundamental tenant of the comparative method" (Blust 2000, page 314)—is critical for correct subgrouping (Brugmann 1884; Hennig 1966; Blust 2000; Greenhill and Gray 2009). Second, distance metrics remove a large proportion of the signal in the data and better classification performance is achieved when raw data are used (Steel, Hendy, and Penny 1988). By using a distance-based subgrouping method the Levenshtein is susceptible to exactly the same problems that classical lexicostatistics faced (Swadesh 1952). In particular, there is a very high vulnerability to variation in retention rates (Bergsland and Vogt 1962; Blust 2000; Greenhill and Gray 2009). Dyen (1965) applied lexicostatistics to the Austronesian language family and interpreted the results as indicating an origin in Melanesia, possibly in the Bismarck Archipelago north of New Guinea. This contradicts sharply with the view from historical linguistics which suggests Taiwan as the homeland (Blust 1999; Gray, Drummond, and Greenhill 2009). Blust (2000) has demonstrated that retention rates in Malayo-Polynesian languages range from 5% to 60% over the last 4,000 years or so. The languages which have had more replacement in basic vocabulary are predominantly those of Melanesia, where there has been much contact-induced change as incoming Austronesian languages encountered non-Austronesian languages (Ross 1996). Subgrouping the languages with lower similarities placed these Melanesian languages at the base of the tree, and led to

the incorrect inference of a Melanesian homeland (Blust 2000; Greenhill and Gray 2009). A Levenshtein-derived classification of the Austronesian languages (Petroni and Serva 2008) makes an identical mistake, suggesting that this approach is just as adversely affected by variations in retention rate. It is therefore crucial to use methods that can use all the cognate data, and account for variation in rates of lexical change over time—like Bayesian phylogenetic methods (Greenhill and Gray 2009).

Using the Levenshtein distance to automatically classify languages has many attractions. These results indicate that better methods and approaches need to be explored, however. One approach would be to make a less linguistically naive distance. For example, the algorithm could be altered to allow larger changes such as metathesis and reduplication. Or the transposition weights could be modified to make transitions between common sound changes less penalized. These weights could be adjusted using evidence from historical linguistics (Blevins 2004), or using a more general acoustic distance measure (Mielke 2005). One step towards a more nuanced distance has been taken by Wieling, Prokić, and Nerbonne (2009), who modified the Levenshtein distance to handle increased rates of metathesis in Bulgarian dialects and found a small increase in performance on a string alignment task. A further modification using a Pointwise Mutual Information method to learn the weights of phonetic transitions performed better and had fewer errors than the modified Levenshtein.

Regardless of potential improvements, using a static metric like the Levenshtein is still problematic. Processes like metathesis and reduplication are not necessarily common in other languages or language families, and common sound changes are not only phonetically constrained but are also constrained by the sounds already occurring in a language. A superior approach would be to use an adaptive algorithm to learn the transition weights from the data before classifying the languages using methods like naive Bayesian classifiers (Dunning 1994; Ellison 2007), dynamic Bayesian networks (Kondrak and Sherif 2006), or stochastic transducers (Ristad and Yianilos 1998). A recent technique uses a probabilistic model of phonological change trained on a corpus of data with an expectation-maximization algorithm to infer protoforms directly based on a phylogeny (Bouchard-Côté et al. 2008; Bouchard-Côté, Griffiths, and Klein 2009). Given these new methods which promise substantial increases in performance, and the large collection of established cognate data contained in databases like the ABVD that can be used to train these algorithms, there is no reason to continue using the Levenshtein distance to classify languages.

Acknowledgments

I thank Quentin Atkinson, Russell Gray, Robert Ross, Annik van Toledo, and four anonymous reviewers for discussion and suggestions. Funding was provided by a Royal Society of New Zealand Marsden grant (no. UOA0709).

References

Bakker, D., A. Müller, V. Velupillai, S. Wichmann, C. H. Brown, P. Brown, D. Egorov, R. Mailhammer, A. Grant, and E. W. Holman. 2009. Adding typology to lexicostatistics: a combined approach to language classification. *Linguistic Typology*, 13:167–179.

Bergsland, K. and H. Vogt. 1962. On the validity of glottochronology. *Current Anthropology*, 3(2):115–153.

Blevins, J. 2004. *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge University Press, Cambridge.

Blust, R. 1991. The greater central Philippines hypothesis. *Oceanic Linguistics*, 30:73–129.

Blust, R. 1993. Central and Central-Eastern Malayo-Polynesian. *Oceanic Linguistics*, 32:241–293.

Blust, R. 1999. Subgrouping, circularity and extinction: Some issues in Austronesian comparative linguistics. In E. Zeitoun and P. J-K. Li, editors, *Selected Papers from the Eighth International Conference on Austronesian Linguistics*. Symposium

Series of the Institute of Linguistics, Academia Sinica, Taipei, Taiwan, pages 31–94.
Blust, R. 2000. Why lexicostatistics doesn't work: The 'universal constant' hypothesis and the Austronesian languages. In C. Renfrew, A. McMahon, and L. Trask, editors, *Time Depth in Historical Linguistics*. McDonald Institute for Archaeological Research, Cambridge, chapter 13, pages 311–331.
Blust, R. 2004. *t to k: An Austronesian sound change revisited. *Oceanic Linguistics*, 43:365–410.
Blust, R. 2009. *The Austronesian Languages*. Pacific Linguistics, Canberra.
Blust, R. A. 1978. Eastern Malayo-Polynesian: A subgrouping argument. In *Second International Conference on Austronesian Linguistics: Proceedings*, pages 181–234, Canberra.
Bouchard-Côté, A., T. L. Griffiths, and D. Klein. 2009. Improved reconstruction of protolanguage word forms. In *Proceedings of Human Language Technologies*, pages 65–73, Boulder, CO.
Bouchard-Côté, A., P. Liang, T. L. Griffiths, and D. Klein. 2008. A probabilistic approach to language change. *Advances in Neural Information Processing Systems*, 20:1–8.
Brown, C. H., E. W. Holman, S. Wichmann, and V. Velupillai. 2007. Automated classification of the World's languages: A description of the method and preliminary results. *STUF-Language Typology and Universals*, 61:285–308.
Brugmann, K. 1884. Zur Frage nach den Verwandtschaftsverhältnissen der Indogermanischen Sprachen. *Internationale Zeitschrift für allgemeine Sprachewissenschaft*, 1:226–256.
Campbell, L. and V. Grondona. 2008. Ethnologue: Languages of the world. review. *Language*, 84:636–641.
Campbell, L. and B. Poser. 2008. *Historical Linguistics: History and Method*. Cambridge University Press, Cambridge.
Currie, T. E., S. J. Greenhill, R. D. Gray, T. Hasegawa, and R. Mace. 2010. Rise and fall of political complexity in island south-east Asia and the Pacific. *Nature*, 467(7317):801–804.
Dahl, O. C. 1973. *Proto-Austronesian*. Scandinavian Institute of Asian Studies, Monograph Series No. 1, Lund.
Dempwolff, O. 1934. Vergleichende lautlehre des austronesischen wortschatzes. 1. induktiver aufbau einer indonesischen ursprache. *Zeitschrift für Eingeborenen-Sprachen*, 15.
Dempwolff, O. 1937. Vergleichende lautlehre des austronesischen wortschatzes. 2. deductive anwendung des urindonesischen auf austronesische einzelsprachen. *Zeitschrift für Eingeborenen-Sprachen*, 17.
Dempwolff, O. 1938. Vergleichende lautlehre des austronesischen wortschatzes. 3. austronesisches wörterverzeichnis. *Zeitschrift für Eingeborenen-Sprachen*, 19.
Dunning, T. 1994. Statistical identification of language. Technical report MCCS 94-273, Computing Research Laboratory, Las Cruces, NM.
Durie, M. and M. D. Ross. 1996. *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*. Oxford University Press, Oxford.
Dyen, I. 1965. *A Lexicostatistical Classification of the Austronesian Languages*. Waverly Press, Inc, Baltimore, MD.
Ellison, T. M. 2007. Bayesian identification of cognates and correspondences. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, SigMorPhon '07, pages 15–22, Stroudsburg, PA.
Embleton, S. M. 1985. *Statistics in Historical Linguistics*. Studienverlag Brockmeyer, Bochum.
Gooskens, C. and W. Heeringa. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, 16:189–207.
Grace, G. W. 1959. *The Position of the Polynesian Languages within the Austronesian (Malayo-Polynesian) Language Family*. (Indiana University Publications in Anthropology and Linguistics) Waverly Press, Baltimore, MD.
Gray, R. D., D. Bryant, and S. J. Greenhill. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society, B*, 365:3923–39233.
Gray, R. D., A. J. Drummond, and S. J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, 323(5913):479–483.
Greenhill, S. J., R. Blust, and R. D. Gray. 2008. The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4:271–283.
Greenhill, S. J., A. J. Drummond, and R. D. Gray. 2010. How accurate and robust are the phylogenetic estimates of

Austronesian language relationships? *PLoS One*, 5(3):e9573.

Greenhill, S. J. and R. D. Gray. 2009. Austronesian language phylogenies: Myths and misconceptions about Bayesian computational methods. In K. A. Adelaar and A. Pawley, editors, *Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust*. Pacific Linguistics, Canberra, pages 375–397.

Heeringa, W., P. C. J. Kleiweg, C. Gooskens, and J. Nerbonne. 2006. Evaluation of string distance algorithms for dialectology. In *Proceedings of the Workshop on Linguistic Distances*, pages 51–62, Sydney.

Hennig, W. 1966. *Phylogenetic Systematics*. University of Illinois Press, Urbana. Translation by D. Davis and R. Zangerl.

Holman, E. W. 2010. Do languages originate and become extinct at constant rates? *Diachronica*, 27(2):214–225.

Holman, E. W., S. Wichmann, C. H. Brown, V. Velupillai, A. Müller, and D Bakker. 2008. Explorations in automated lexicostatistics. *Folia Linguistica*, 42:331–354.

Kessler, B. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the Seventh conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 60–67, Dublin.

Kondrak, G. and T. Sherif. 2006. Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification. In *Proceedings of the Workshop on Linguistic Distances*, pages 43–50, Sydney.

Kruskal, J. B. 1983. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM Review*, 25(2):201–237.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.

Lewis, P. M., editor. 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, 16th edition.

Mielke, J. 2005. Modeling distinctive feature emergence. In J. Alderete, C.-H. Han, and A. Kachetou, editors, *Proceedings of the West Coast Conference on Formal Linguistics*. Cascadilla Proceedings Project, Somerville, MA, pages 281–289.

Nerbonne, J. and W. Heeringa. 2009. Measuring dialect differences. In J. E. Schmidt and P. Auer, editors, *Language and Space: Theories and Methods*. Mouton De Gruyter, Berlin, pages 550–567.

Pawley, A. 1972. On the internal relationships of eastern Oceanic languages. In R. C. Green and M. Kelly, editors, *Studies in Oceanic Culture History*, volume 3. Bernice P. Bishop Museum, Honolulu, HI, pages 3–106.

Pawley, A. 2002. The Austronesian dispersal: Languages, technologies and people. In P. Bellwood and C. Renfrew, editors, *Examining the Farming/Language Dispersal Hypothesis*. McDonald Institute for Archaeological Research, Cambridge, pages 251–273.

Petroni, F. and M. Serva. 2008. Language distance and tree reconstruction. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(08):P08012.

Reid, L. A. 1971. *Philippine minor languages: Word lists and phonologies*. University of Hawai'i Press, Canberra.

Ringe, D. 1992. On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society*, 82:1–110.

Ristad, E. S. and P. N. Yianilos. 1998. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532.

Ross, M. and Å. Næss. 2007. An Oceanic origin for Äiwoo, a language of the Reef Islands. *Oceanic Linguistics*, 46:456–498.

Ross, M. D. 1988. *Proto-Oceanic and the Austronesian Languages of Western Melanesia*. Pacific Linguistics, Canberra.

Ross, M. D. 1996. Contact-induced change and the comparative method: Cases from Papua New Guinea. In M. Durie and M. D. Ross, editors, *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*. Oxford University Press, New York, pages 180–217.

Serva, M. and F. Petroni. 2008. Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005.

Steel, M., M. D. Hendy, and D. Penny. 1988. Loss of information in genetic distances. *Nature*, 336:118.

Susko, E., Y. Inagaki, and A. J. Roger. 2004. On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. *Molecular Biology and Evolution*, 21:1629–1642.

Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, 96(4):452–463.

Tria, F., E. Caglioti, V. Loreto, and A. Pagnani. 2010. A stochastic local search approach to language tree reconstruction. *Diachronica*, 27(2):341–358.

Tryon, D. T. and B. D. Hackman. 1983. *Solomon Islands Languages: An Internal Classification*. Pacific Linguistics, Canberra.

van der Ark, R., P. Mennecier, J. Nerbonne, and F. Manni. 2007. Preliminary Identification of Language Groups and Loan Words in Central Asia. In *Proceedings of the RANLP Workshop on Computational Phonology*, pages 12–20, Borovetz.

Wichmann, S. and E. W. Holman. 2009. Population size and rates of language change. *Human Biology*, 81:259–274.

Wichmann, S., E. Holman, A. Müller, V. Velupillai, J.-M. List, O. Belyaev, M. Urban, and D. Bakker. 2010. Glottochronology as a heuristic for genealogical language relationships. *Journal of Quantitative Linguistics*, 17(4):303–316.

Wichmann, S., A. Müller, and V. Velupillai. 2010. Homelands of the world's language families: A quantitative approach. *Diachronica*, 27(2):247–276.

Wieling, M., J. Prokić, and J. Nerbonne. 2009. Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)*, pages 26–34, Athens.