

Assessing the lexical evidence for a Central Solomons Papuan family using the Oswalt Monte Carlo Test*

Michael Dunn^{a,c} and Angela Terrill^{a,b}

^aMax Planck Institute for Psycholinguistics /^bRadboud University Nijmegen / ^cDonders Institute for Brain, Cognition, and Behavior

In the absence of comparative method reconstruction, high rate of lexical cognate candidates is often used as evidence for relationships between languages. This paper uses the Oswalt Monte Carlo Shift test (a variant of Oswalt 1970) to explore the statistical basis of the claim that the four Papuan languages of the Solomon Islands have greater than chance levels of lexical similarity. The results of this test initially appear to show that the lexical similarities between the Central Solomons Papuan languages are statistically significant, but the effect disappears when known Oceanic loanwords are removed. The Oswalt Monte Carlo test is a useful technique to test a claim of greater than chance similarity between any two word lists — with the proviso that undetected loanwords strongly increase the chance of spurious identification.

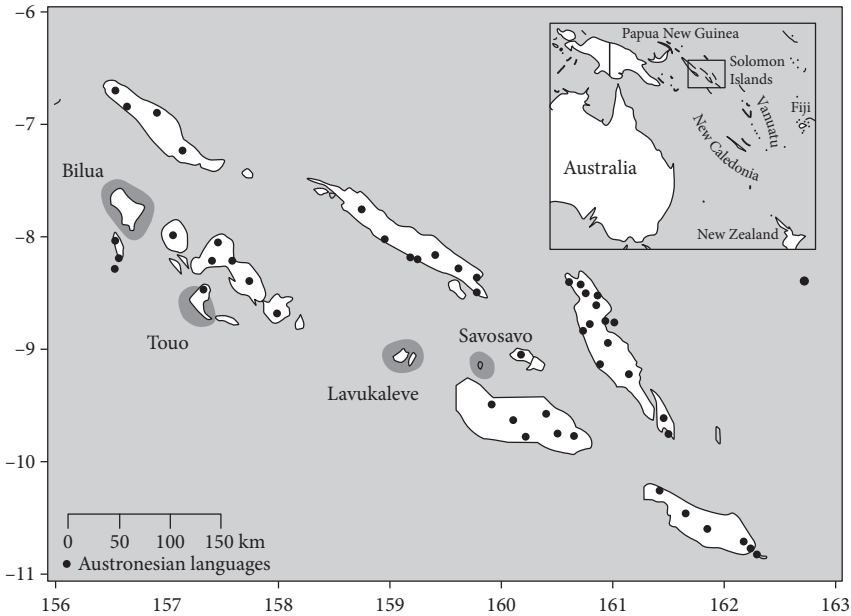
Keywords: Levenshtein distance, lexical comparison, Oswalt test, Papuan languages, randomization, Solomon Islands languages

1. Introduction

1.1 The Central Solomon Islands Papuan languages

Of the 90-plus languages spoken in the Central Solomon Islands, all but four are members of the Austronesian family. The remaining four languages, Bilua, Touo, Lavukaleve and Savosavo, are called Papuan, a non-genetic label used for the non-Austronesian indigenous languages of the New Guinea continent and surrounding areas. Evidence for genealogical grouping including the four Papuan

* This work was supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) and the Max Planck Society. We are grateful to Simon Kirby, Eva Lindström, Andrew Pawley, Edgar Suter and four anonymous reviewers for comments on early versions of this paper.



Map 1. The four Papuan languages of the Central Solomon Islands.

languages of the Central Solomon Islands group has been proposed on typological as well as on morphological/lexical grounds (Codrington 1885, Schmidt 1902, Ray 1926, 1928, Lanyon-Orgill 1953, Wurm 1972, Todd 1975, Greenberg 1971), but no comparative method reconstruction of a Proto-Central Solomons group has ever been published. The most recent inquiries into the relatedness of the Central Solomons Papuan languages have been carried out in the structural phylogenetics framework (Dunn et al. 2005, 2008, Dunn 2009), which uses structural features to investigate potential genealogical relationships between these languages.

Until recently, several other languages of the Solomon Islands were thought to be Papuan. Reanalysis of published materials on the extinct Kazukuru language of New Georgia has shown that it is Oceanic Austronesian, most closely related to the other Oceanic languages spoken around its historical range (Usher 2002: 66, Dunn & Ross 2007). New fieldwork on Äiwoo, spoken in the Reef Islands (Ross & Næss 2007, Næss & Boerger 2008) has resolved an old debate about the affiliation of the Santa Cruz languages (Lincoln 1978, Wurm 1978) by showing that they form a subgroup of Eastern Oceanic.

In this paper we use a formal measure of lexical similarity between word lists to compare the aggregate similarity of semantically aligned wordlists to the distribution of similarity scores between a large sample of semantically unaligned wordlists (Baxter 1995, Kessler 2007, Justeson & Stephens 1980, Oswalt 1970, 1998, also McMahon & McMahon 2005:66–68). This gives us a mechanism to

make a statistical evaluation of the claim that there are greater than chance levels of similarity between the Papuan languages of the Solomon Islands.

We demonstrate that the lexical similarities between the Central Solomons Papuan languages appear superficially to be significant, but that the statistical effect shrinks or disappears when probable Oceanic loanwords are removed. A particular point of reference here is Greenberg's Indo-Pacific Hypothesis (Greenberg 1971), which includes a Central Solomons group, embedded within an East Papuan group. We demonstrate that the positive identification produced by Greenberg's methodology is an artefact of Papuan — Austronesian language contact. Finally, despite failing to find greater than chance levels of lexical relatedness in this study, it is noteworthy that in previous research we have identified patterns of typological variation which are apparently robust to contact. This suggests that rather unusually, the East Papuan languages potentially show a greater degree of structural stability than lexical stability (Dunn et al. 2005, Dunn 2009).

1.2 Population history

At least two major population movements must have come through the Solomons. The first human settlement was at least 40,000 years ago (Spriggs 1997); more recent was the Lapita expansion in the region perhaps as far back as 3,500 years ago (Kirch 1997), bringing the speakers of the Oceanic languages, perhaps in more than one wave (Ross 1988).

The population genetics of the Solomons show a long history of biological continuity, with assimilation rather than replacement of populations being the norm over the entire period of human habitation. Friedlaender's work on a range of biological markers suggests that there might be more than one of these pre-Austronesian settlements in island Melanesia, followed by a lesser degree of differentiation in situ (Friedlaender 1987: 355). More recent work suggests there will be no easy correlation of population genetics and language groupings in the Solomon Islands. A small-scale genetic study by Cox & Lahr (2006: 35) — examining the association between language affiliation and genetic markers in two island groups of the Solomon Islands, Malaita (solely Oceanic-speaking) and southern Rendova (speakers of the Papuan language Touo) — found no relationship on this scale between Y-chromosome lineages and language affiliation. Thus it was not possible to predict from Y-chromosome DNA whether an individual belonged to an Austronesian or Papuan-speaking population. For the wider area of Island Melanesia, Hunley et al. (2008) conclude that linguistic and genetic exchange have, for the most part, erased evidence of early settlement history, with the exception of the rugged hinterland of New Britain.

1.3 Typological arguments

Most of the recent claims for relationship between the Papuan languages of the Solomon Islands rest on typological features. Capell (1969), who does not in fact make a claim for the relatedness of these languages, nonetheless sets the stage by discussing in detail the noun class systems. He notes that while the four languages all have noun classes (i.e. gender), these classes are expressed very differently in each language. Wurm (1971) also notes the similarities in noun classification systems, with Terrill (2002) reaching a similar conclusion.

Wurm (1972) argued strongly for the use of 'typological and structural' characteristics in Papuan language classification on the grounds that lexicons of the languages in the region have been too much influenced by borrowing to be usable in determining language relationship. He makes an explicit argument for grouping the Solomon Islands Papuan languages with the rest of the Papuan languages of Island Melanesia, including Reefs and Santa Cruz. The structural features amassed by Wurm include: moderately complex phonologies; overt noun classification with concord; gendered pronouns; inclusive–exclusive contrast; number marked on nouns or noun particles; dual number; subject marked by preposed particles or prefixes, and object by suffixes; elaborate TAM, and apparently no medial verbs (however cf. Terrill 2003a on Lavukaleve and Wegener 2008 on Savosavo).

Dunn et al. (2002) describe various typological features of the languages, many as given in Wurm (1975), noting similarities and dissimilarities, but not committing to whether or not the languages form a genetic unit. Two more recent investigations into these languages directly address the question of the group's genetic unity. It is notable that Ross (2001) using pronominal evidence and Dunn et al. (2005, 2008) using structural features both offer much more tentative conclusions than earlier scholars. The typological argument is most fully presented in Dunn (2009), where it is shown that typological similarity between homogeneous pairs of Papuan-Papuan and Oceanic-Oceanic languages can be predicted by geographic proximity, but that the typological similarity between heterogeneous (Papuan-Oceanic) cannot. This means that the patterns of typological similarity between the Papuan languages of island Melanesia have not been caused by shared patterns of Oceanic contact, otherwise Papuan languages would tend to be structurally more similar to their nearer Oceanic neighbours. A historical explanation for the patterns of similarity between Papuan languages must come down to (i) common ancestry or (ii) a period of contact prior to their mutual isolation by the spread of the Oceanic languages. While the similarities between these four languages, based on their typological features, are certainly suggestive, such features alone cannot make a watertight case for the genetic unity of these languages. Other types of evidence are needed: pronouns and lexicon are the two other areas in which a case for the unity of these languages has been made.

1.4 Evidence from pronoun reconstruction

There have been two published cases in support of the genetic unity of these four Solomons Papuan languages based on pronoun comparison. Greenberg states that “the most powerful indication of the unity of the Central Solomons group comes from the pronominal system” (1971: 816), and Ross (2001), identified formal resemblances among the pronouns and made reconstructions.

Table 1 presents the pronoun paradigms of the Central Solomon languages. Cognate sets identified by Greenberg are in bold. Reconstructions of particular forms where relevant are from Ross (2001).

Table 1. Central Solomons pronoun comparison.

	SG	DU	PL
1 INCL	[blb] <i>anja</i> [lvk] <i>ɲai</i> (Greenberg <i>ɲai</i> / <i>ɲa</i>) [tqu] <i>yei</i>	[blb] <i>aniqe</i> [lvk] <i>el</i> [tqu] <i>be</i> [svs] <i>mai</i>	[blb] <i>(ani)me</i> [lvk] <i>me</i> (Greenberg: also <i>mee</i>) [tqu] <i>memw</i> [svs] <i>mai</i>
1 EXCL	[svs] <i>anyi/ai</i> (Greenberg: <i>ani</i>) [svs] *ny <i>i</i> ; [lvk], [blb] *ɲa	[blb] <i>eqe</i> [lvk] <i>mel</i> [tqu] <i>yere</i> [svs] <i>aghe</i> (Greenberg: <i>ange</i>)	[svs] *mai, [lvk], [blb] *me [blb] <i>aniɲe</i> (Greenberg: <i>(ani)vo</i>) [lvk] <i>e</i> (Greenberg <i>ee</i>) [tqu] <i>yebw</i> [svs] <i>ave</i> [svs] *ve, [lvk] *e [svs], [blb] *ge
2	[blb] <i>ɲo</i> [lvk] <i>inu</i> (Greenberg <i>ɲo</i>) [tqu] <i>noe</i> (Greenberg: <i>no</i>) [svs] <i>no</i> [svs], [tqu] *no, [lvk], [blb] *ɲo	[blb] <i>qe</i> [lvk] <i>imil</i> [tqu] <i>bere</i> [svs] <i>pe</i>	[blb] <i>me</i> [lvk] <i>imi</i> (Greenberg: also bound form <i>me</i>) [tqu] <i>mebw</i> [svs] <i>me</i> [svs], [blb] *me, [lvk] *m(e,i)
3	[blb] <i>nei</i> (Greenberg: <i>o</i>) [lvk] <i>foina</i> (Greenberg <i>o</i> (possessive)) [tqu] <i>zo</i> [svs] <i>lo/la</i> [lvk] *o (from possessive <i>o</i>)	[blb] <i>nioqi</i> [lvk] <i>foinala</i> [tqu] <i>zere</i> [svs] <i>to/ta</i>	[blb] <i>ni</i> (Greenberg <i>se</i>) [lvk] <i>foiva</i> (Greenberg: <i>ma, mi</i> (bound forms)) [tqu] <i>mw</i> [svs] <i>ze/zepo/za</i> (Greenberg <i>mi</i> object) [svs] *ze, [tqu], [lvk] *ma

Notes to the table:

- 1SG: If Greenberg had had the correct form we do not know whether he would have still counted the Savosavo forms with the Lavukaleve and Bilua. Judging by the cognate set for the 2SG, which conflates *no* with *ŋo*, he might well have. For Ross, Greenberg’s 1SG cognate set is actually two cognate sets, reconstructable to **ni* and **ŋa*.
- 2SG: Again for Ross this is two cognate sets, not one as with Greenberg.
- 3SG: Ross makes no reconstruction for Bilua *o*, but reconstructs Lavukaleve’s 3SG *o* as **o*.
- 1DU excl: Both Ross and Greenberg have these two forms as a cognate set.
- 1PL incl: Greenberg’s single cognate set is two, **mai* and **me*, for Ross.
- 1PL excl: the three words in this cognate set are separate for Ross, who reconstructs **ve* for Savosavo *ave*; **e* for Lavukaleve *ee* (actually *e*); and nothing for Bilua (*ani*)*vo*. It is not clear what the *anivo* form is: Ross does not have it, he just has *anime*, and Obata (2003) seems not to have it either.
- 2PL: Greenberg and Ross agree that this is a single cognate set.
- 3PL: Ross gives **ze* for Savosavo, and **ma* for Lavukaleve, and Touo.
- Lavukaleve (lvk) 3rd person pronouns are masculine distance-neutral; Lavukaleve data from Terrill (2003).
- Touo (tqu) pronouns are all masculine. Touo paucal pronouns are not included. Initial *y* followed by a vowel indicates a breathy voice vowel. *W* indicates *ɔ*. Touo data comes from Dunn/Terrill field notes.
- Bilua (blb) 3rd person pronouns are masculine proximate; Bilua data from Obata (2003).
- Savosavo (svs) 3rd person pronouns are masculine; Savosavo data from Wegener (2008).

In conclusion, only two of the cognate sets proposed by Ross and Greenberg, over the same sets of pronoun paradigms, agree. Ross’s work is much easier to evaluate because he publishes the entire pronoun paradigms he is working with, whereas Greenberg only publishes the putative cognate sets.

1.5 Lexicostatistical comparison

Wurm (1975:788) claimed to have identified high levels of shared vocabulary between these languages: levels in the “low thirties to low forties”, which he considers evidence that these languages descended from a common ancestor. These levels were considerably higher than those found by Tryon & Hackman (1983), who produced counts based on c.300 word lists yielding figures in the teens at most. Surprisingly Wurm (1975:788) believes that the presence of Austronesian loans deflates the true estimate of lexical cognacy: “undoubtedly their original lexical interrelationship used to be much higher than the figures referred to above may suggest”. Wurm does not indicate the number of words compared in this study but Todd’s (1975) comparison of an 180-item word list of the four Papuan languages yields very few potential cognates, most of which are clearly recent Austronesian loans. Todd also notes difficulties inherent in such an exercise, establishing regular sound correspondences and the influence of Austronesian borrowings.

Tryon & Hackman (1983) published modified Swadesh word lists of 321 items for all the Solomon Islands languages and carried out lexicostatistical comparison on a subset of 200 words. The percentages of shared lexicon between the central Solomon Island Papuan languages are low, as shown in Table 2. These figures are

approximately one standard deviation below the mean (24%, SD=17) for all Solomon Islands languages.

Table 2. Lexical cognate percentages from Tryon & Hackman (1983).

Savosavo	13.7%		
Touo	8.2%	10.1%	
Bilua	7.6%	7%	9.1%
	Lavukaleve	Savosavo	Touo

Greenberg makes the most detailed case for the relatedness of the four central Solomon Islands languages, based almost entirely on lexical items, and by listing his putative cognates he gives us an opportunity to critically appraise his proposed cognate sets. The following section discusses in detail Greenberg's method of lexical comparison, and the data used.

1.6 Mass comparison

Greenberg classified the languages of the Central Solomon Islands, together with the rest of the Papuan languages of Island Melanesia, with the Papuan languages of New Guinea, and the languages of Tasmania and the Andaman Islands, into "a large group of genetically related languages" (Greenberg 1971: 807): the Indo-Pacific languages. The larger classification has never been widely accepted, rejected by Pawley (2007), Crowley & Dixon (1981) and other regional subgroup specialists. While we too reject the larger group, what concerns us here is the smaller group of four 'Papuan' languages of the Central Solomon Islands, posited as a subgroup of Indo-Pacific. We ignore the proposal that the Santa Cruz languages are another subgroup joined to the Central Solomons Papuan languages at the next highest taxonomic level, since the Santa Cruz languages have been convincingly shown to be Austronesian (Ross & Næss 2007, Næss & Boerger 2008).

Greenberg's method has been known by several names; first 'collateral comparison', later 'mass comparison', and from 1987 'multilateral comparison'. In the Indo-Pacific Hypothesis (1971) Greenberg was, in his own words, doing 'mass comparison', which involves "the rapid inspection of hundreds of words and grammatical forms from each of hundreds of languages, with languages showing sizeable numbers of resemblances being grouped together" (Trask 1996: 383). Note that Greenberg himself did not calculate lexical percentages between the central Solomons Papuan languages.

Multilateral versus pairwise comparison. Greenberg's proposed method of multilateral comparison cannot be reduced to a series of pairwise comparisons of languages. Greenberg himself clearly thought about using both types of method; the descriptions of the proposed quantitative techniques quoted above are all bilateral methods, considering all permutations of pairs of languages. Greenberg (2001) and others have argued that bilateral comparisons have lower statistical power than multilateral comparisons, and that there might be relationships apparent in a multilateral comparison that are not evident in any or all of the bilateral comparisons of the same set of languages. Kessler & Lehtonen (2006) and Kessler (2007) discuss a statistically rigorous way of carrying out a multilateral comparison using computational, quantitative techniques. In the present paper we do not attempt an implementation of a multilateral method, since we are only attempting to test the significance of the lexical similarities between the four Central Solomon Islands Papuan languages; we are not testing the Indo-Pacific Hypothesis as a whole.

2. Testing lexical similarity

2.1 The Oswald Monte Carlo test

Oswald shift test. The Oswald shift test is frequently referred to (cf. McMahon & McMahon 2005), but rarely carried out. It comprises the computational implementation of a test for greater than chance levels of similarity between wordlists. In this test, two semantically aligned wordlists — A and B — are scored for the number of apparent shared cognate terms. If item 1 on list A is called A1, item 2 is A2 and so on to A100 (for a 100 word list), then in the comparison of aligned lists, A1 is compared with B1, A2 with B2, etc. After the comparison of aligned lists, list B is rotated by one place, so that A1 is compared with B2, A2 with B3, through to A100 with B1. This rotation is done 99 times, so that each word in list A has been compared with each word in list B. The apparent cognate scores for the 99 semantically unaligned lists gives a sample distribution of the levels of apparent cognacy that would be expected by chance. If the cognate count for the semantically aligned list is higher than any of the semantically unaligned lists, then the similarity between these lists is greater than would be expected by chance, with a p-value less than 0.01.

The computationally difficult part of this technique is the comparison of forms. It is highly desirable that the comparison be done computationally, because there are a large number of pairwise comparisons to be done, and so that the comparisons are objective and unbiased by the expectations of the person doing the comparison. A computational implementation of explicit matching criteria avoids

these problems. The Oswalt shift test is an attempt to balance the statistical requirements of the test (a very large number of comparisons) against the limitations of the computational resources available in the 1970s. The current state of personal computing is vastly improved, and some of the compromises made by Oswalt in his experimental design are no longer necessary. The major theoretical weakness of the shift test is the limitation on the number of semantically unaligned comparisons which can be made. With a 100-word list only 99 unaligned comparisons are made. Further, it is unclear how the shift test could cope with, e.g., more than one word in a language for a particular meaning — limiting the lists to one form for each meaning means that the experimenter must throw away data, and it is unclear what grounds could be used for choosing one form over others as the exponent of a particular meaning. This is of course a problem with any lexical comparison.

Monte Carlo variant. The main criticisms of the Oswalt shift test are that it would be tedious to do by hand and it would be impossible to blind; that is, a linguist capable of identifying cognate candidates could almost always work out when the semantically aligned comparison was being made. The mechanics of the shift test may also lead to unexpected behaviour: the list order for each language is fixed, not random, and so the set of comparisons cannot be treated statistically as a random sample of a distribution. This latter point is easily remedied: we replace the ‘shift’ operation of the test with true randomization. Instead of carrying out 100 fixed order comparisons we can do, for example, one thousand comparisons of randomly shuffled lists. With the computer power now available we can carry out a fully randomized version of the shift test. Algorithms which use random sampling to solve problems which are insoluble using deterministic methods are called ‘Monte Carlo’ methods (Metropolis 1987, Metropolis & Ulam 1949), and so we call the revised test the Oswalt Monte Carlo test. Instead of limiting the semantically unaligned comparisons to $N-1$ comparisons for each word, we can produce a much larger sample of comparisons from randomized lists. It would be possible to do this test exhaustively, and compare every possible combination, but this would be overkill — if we wish to make a statement about the distribution of chance similarities, a large random sample gives us the same statistical validity. The randomization procedure so described resembles a standard Monte Carlo resampling procedure in statistics, to which apply standard statistical tests and standards of evaluation (Good 2006). Doing the Oswalt test with fully randomized comparisons it is furthermore possible to eliminate the constraint that there can only be ‘one word per cell’, i.e. under the Monte Carlo version of the test a language can have multiple words associated with each meaning on the list.

Following Oswalt’s original paper, we use a computational method of counting cognates. We test a number of different models for identifying cognate candidates,

all variants of the Levenshtein algorithm (below) widely used in computational linguistics (Kruskal 1983, Nerbonne et al. 1999, Sankoff & Kruskal 1983).

Greenberg's Indo-Pacific Hypothesis is useful here because it provides a set of matched cognates from the languages under investigation. Our analysis takes these cognacy judgements, and uses them to calibrate the parameters of a string comparison algorithm, which we then use in the Monte Carlo test. In the process it may be possible to learn something about how much evidence is needed to produce a claim of greater than chance levels of similarity between putatively related languages.

2.2 Recognising cognate candidates

The first requirement for the Oswalt Monte Carlo test is a computational implementation of the cognate matching criteria. In this section we describe a number of models for matching potential cognates and test the fidelity of these models with a range of parameters against Greenberg's data. An algorithm which adequately replicates Greenberg's matching criteria is identified from this set for use in the tests with fuller lexical data.

The Indo-Pacific Hypothesis paper includes tables of matches (Greenberg 2005a:205). In traditional comparative linguistics these would be termed 'cognate candidates'; in the Greenbergian paradigm they are part of a larger, semi-statistical comparison meant to exclude the possibility of accidental similarity without producing an actual reconstruction. Greenberg's method assumes that relationships within a subgroup are established from above and below simultaneously: cognates are not necessarily attested more than once within a subgroup — evidence can equally come from matches with languages from other groups of the Indo-Pacific macro-family.

Greenberg's paper does not include all his data. Critically, only positive matches are reported: we have no negative evidence, i.e. comparisons of words which are not considered to be matches. This considerably inhibits our ability to test Greenberg's claims. However, methodological questions aside, we are ultimately interested in the real-world status of the Central Solomons Papuan languages, rather than simply testing Greenberg's method. From this point of view, it would be inappropriate to limit our investigation only to materials which were available to Greenberg in the 1960s. Below we attempt to compensate for our imprecise re-creation of Greenberg's original data by using the much more complete set of lexical data available to us today. In this section we consider only the terms presented in Greenberg's paper since our concern here is to parameterize our model to replicate Greenberg's match judgements in a way that can be used computationally in a huge number of comparisons. In later sections we use the more complete Solomon Islands data in Tryon & Hackman (1983), which better represents the linguistic situation.

Matching algorithm. This paper uses an approach to identify cognate candidates based on measures of phonological distance. Various measures of phonological similarity will be tested against a list of Greenberg’s matched word pairs and against a list of Greenberg’s (presumed) non-matching pairs. We identify a method (i) maximizing the number of ‘true’ matches identified (according to Greenberg), while (ii) minimizing the number of false positives. We test several variations on the Levenshtein distance metric (Levenshtein 1966), which is a standard tool used in computational dialectometry and quantitative historical linguistics amongst other disciplines (see for example Ellison & Kirby 2006, Heeringa, Kleiweg, Goo-skens & Nerbonne 2006, Kondrak & Sherif 2006, Heeringa 2004, Kessler 1995).

In its classic form, the Levenshtein distance is simply the minimum number of changes needed to convert one string into another string. This is also known as the ‘edit distance’. For example, to move from Touo *zavaŋa* to Lavukaleve *savun* ‘beach’, four steps are required:

Table 3. Calculating Levenshtein distances.

Touo	<i>z</i>	<i>a</i>	<i>v</i>	<i>a</i>	<i>ŋ</i>	<i>a</i>	
	↓			↓	↓	↓	number of changes = 4
Lavukaleve	<i>s</i>	<i>a</i>	<i>v</i>	<i>u</i>	<i>n</i>	<i>ø</i>	

One weakness of the calculation is that if the forms being compared are very different in length then the Levenshtein distance is always quite high. A simple way to compensate for this is to scale the values by dividing the raw Levenshtein distance by the mean length of the terms (Ellison & Kirby 2006). Thus, in the example above of the distance between Touo *zavaŋa* and Lavukaleve *savun*, where the basic edit distance is 4, the scaled Levenshtein distance would be $4/5.5$, i.e. approximately 0.72. The utility of normalization for word length has, however, been called into question in a recent study (Heeringa et al. 2006). The relevance of this result to our study is unclear: Heeringa et al. are using distance measures to measure dialect similarity/difference, whereas we are using them to calculate an index of the probability that two forms are lexical cognates. Borderline cases, where the members of the lexical pair under comparison differ greatly in length, seem to correlate with greater (geographic) distance between dialects, but such cases do not make any obvious prediction about likelihood of cognacy. We treat this question empirically, and test the performance of both scaled and unscaled measures against our lexical database.

Heeringa et al. (2006) also report that measuring Levenshtein distance using other units, such as bigrams (pairs of phonemes) and trigrams (triplets), also performs better for dialectometric purposes than using individual phonemes,

presumably because some of the phonological context is taken into account. A bigram representation of *zavaŋa* would be [#z za av va aŋ ŋa a#] and the minimum number of changes required to ‘become’ *savun* as follows:

Table 4. Bigram changes.

Touo	#z	za	av	va	aŋ	ŋa	a#	
	↓	↓		↓	↓	↓	↓	number of changes = 6
Lavukaleve	#s	sa	av	vu	un	n#	ø	

We have tested bigrams of the raw lexical strings, as well as bigrams of consonants only. No improvement in the matching of aligned versus unaligned data has been found, presumably because Greenberg’s matches do not depend on phonological context.

Other refinements of Levenshtein distance include weighting systems so that changes between some phonemes cost less than changes between others. There are many ways of doing this. One approach would be to use some kind of feature representation, so that, for example, a change from /p/ to /f/ costs only 1 (a change in the feature STRIDENT), whereas /p/ to /ŋ/ would cost 5 (it would vary according to the features used, but this might be construed as changes in OBSTRUENT, ANTERIOR, CORONAL, BACK, and NASAL). This binary feature system was devised following Lass (1984). However, with the relatively small set of canonical matches given in Greenberg (1971) this kind of system is unnecessarily elaborate.

We test two simple refinements of the measure. As well as calculating Levenshtein distance on the entire string, we test Levenshtein distance on consonants only (Greenberg data seems to allow matches of each vowel with almost any other vowel) and ‘archiphonemic’ distance, where classes of consonants are treated as equivalent. The values of these archiphonemes have been determined by inspection of Greenberg’s Central Solomons cognate tables:

Table 5. Greenberg’s cognate segments. Boxed segments are treated as identical for the purposes of matching.

p	b	t	d	k	h	i	u
f	v	s	l	y		e	o
	m	z	r				a
			n	ŋ			

A consonant-only Levenshtein distance calculation is displayed in Table 6:

Table 6. Consonant-only Levenshtein distance.

Touo	<i>z</i>	<i>a</i>	<i>v</i>	<i>a</i>	<i>ŋ</i>	<i>a</i>
	<i>z</i>		<i>v</i>		<i>ŋ</i>	
	↓				↓	
	<i>s</i>		<i>v</i>		<i>n</i>	
Lavukaleve	<i>s</i>	<i>a</i>	<i>v</i>	<i>u</i>	<i>n</i>	<i>ø</i>

number of changes = 2

So the Levenshtein distance here is 2, scaled Levenshtein distance $2/3 = 0.66$.

Table 7. Archiphonemic distance.

Touo	<i>z</i>	<i>a</i>	<i>v</i>	<i>a</i>	<i>ŋ</i>	<i>a</i>
	<i>T</i>		<i>P</i>		<i>N</i>	
	<i>T</i>		<i>P</i>		<i>N</i>	
Lavukaleve	<i>s</i>	<i>a</i>	<i>v</i>	<i>u</i>	<i>n</i>	<i>ø</i>

number of changes = 0

In this table, archiphonemes are represented by the capitalized form of the top-most/leftmost item in the boxes in Table 5. According to the archiphonemic measure, the Levenshtein distance is equal to the scaled Levenshtein distance which in turn equals 0. In other words these forms are the same.

Our archiphonemic measure is similar to Oswalt's methodology for determining matches (described in Oswalt 1998:201), but we use a distance measure rather than categorical matching. Where we do use a threshold distance as being equivalent to a match, this threshold is determined empirically from Greenberg's data, rather than setting an arbitrary threshold number of features. Conceptually, a distance measure is more appealing than a binary match, since we expect the nuances that it captures to correlate with the probability that two forms are cognates.

With a larger set of training data even more elaborate distance measures would be possible. Kondrak & Sherif (2006) test a series of sophisticated lexical distance measures, and they conclude that machine learning methods (bioinformatic exotica such as Paired Hidden Markov Models and Dynamic Bayesian Nets) outperform systems with manually constructed parameters. Our archiphonemic distance measure looks very shabby compared to the state of the art, but its simplicity is probably appropriate given the small amount of training data available. More sophisticated methods could be calculated to replicate very precisely the set of matches given by Greenberg, but could not be expected to generalize well: more complex models are not necessarily better. One recent example is Kessler (2007), which tests a multilateral comparison algorithm using a series of very simple metrics (including one which considers only the voicing of the first sound in

the word!) and concludes that “the basic methodology ... performs adequately and stably with any reasonable metric” (2007: 13). Likewise, Wieling et al. (2009) show that very simple alignment techniques (with, however, an empirical weighting applied to segment distances) are competitive with quite sophisticated machine learning techniques.

Implementation. First we establish parameters for the lexical matching algorithm. For a given algorithm at a given threshold, we measure the rate of detection of true matches, and compare it to the rate of false detection, i.e. the rate of detection of matches in the randomized lists. We have excluded from comparison forms with fewer than two segments (i.e. fewer than two consonants for consonant distance, etc.), since we can’t formalize Greenberg’s criteria for matching. Note also that Greenberg is taking into account similarities with Indo-Pacific languages beyond the Central Solomons languages — as noted, he is carrying out multilateral comparison, rather than pairwise, bilateral comparisons. Distance measures on such short forms are insensitive: there is no scale of similarity, as forms are either identical or completely different. In Figure 1 we test each of eight different measures: we take the raw Levenshtein distance, the bigram distance, the consonant distance and the archiphonemic distance, and for each calculate a scaled and unscaled value.

The solid lines in Figure 1 show the performance of the Levenshtein matching algorithm against Greenberg’s table of cognates for each possible threshold value. The dashed lines show the rate of matches made against the randomized lists. At any given threshold value (on the x-axis), the vertical distance between the solid line and the dashed line is the difference between the true detection rate and the false detection rate. We are seeking an algorithm and a threshold value which give a reasonably high true detection rate, and a reasonable low false detection rate. While a number of different algorithms/thresholds give a reasonable high difference between true and false detection rates, the best trade-off between true and false detection rates while minimizing false detection and maximizing true detection is given by the scaled archiphoneme distance with a threshold of around 0.75.

Figure 1 shows that 98% of the aligned pairs have a distance less than 0.75. The dotted line shows the same test done for a scrambled version of the data, i.e. randomly selected (non-cognate) pairs of words. For the randomized data, only 28% of the pairs are closer than the 0.75 threshold.

Figure 2 shows a comparison of the performance of the scaled archiphoneme distance algorithm on Greenberg’s Lavukaleve and Bilua lexical data, which the threshold parameter set to the empirically determined optimum of 0.75. Under this model, 98% of the putative cognates are identified as matching, compared to 28% of the non-cognates, as shown in Figure 1. This model is used to count the apparent cognate rate in a sample of randomized wordlists, giving a distribution

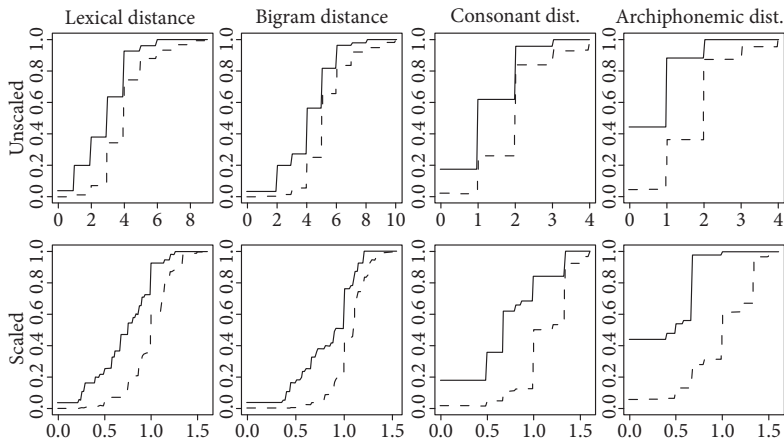


Figure 1. Cognate match rates for eight match algorithms: Scaled and unscaled versions of Levenshtein distance calculated over raw lexemes, bigrams, consonants-only, and ‘archiphonemes’. The x-axis shows the measured distance and the y-axis shows the match rate using that distance as the threshold. The solid line shows the rate of matches using semantically and formally aligned wordlists, and the dashed line shows the rate of matches using randomized wordlists (the false detection rate).

of expected rates of apparent matches between semantically non-aligned lists. Figure 2 shows where the observed rate of apparent matches sits on this distribution: the p values show the proportion of randomized word lists which would be expected to have a similarity score greater than or equal to the value on the horizontal axis.

The intercepts of two dotted lines with the curve (the points marked $p=0.05$ and $p=0.01$) show the similarity scores which encompass, respectively, 95% and 99% of the randomized similarity scores. This diagram shows that the observed similarity score (0.98) is very unlikely to have occurred by chance ($p=0.00008$). The same calculation carried out on the other five pairs of Papuan languages shows p-values ≤ 0.0001 for all pairs except Lavukaleve-Touo, which has $p=0.49$, but note that there are only six semantically matched words shared by the Lavukaleve and Touo lists.

How legitimate is our model? Baxter (1998) criticised Oswalt’s work for ignoring the historical linguistic details of the hypothesis posed by the original linguist, charging that Oswalt instead generated his own hypothesis for the relationship between languages and tested that. This is clearly a serious criticism, but how serious its implications are for this kind of study is an empirical question. In our own study we produce (and test) an explicit computational model of the hypothesis we are investigating. The parameters of the model are set to optimize the fit of the

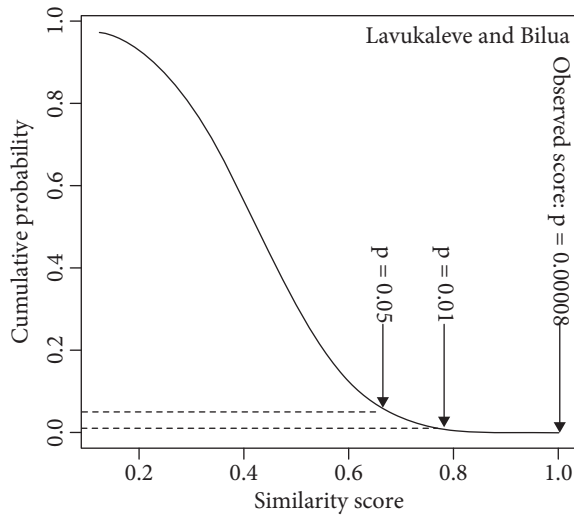


Figure 2. The cumulative probability of obtaining a range of match rates (using the scaled archiphonemic distance measure with a threshold of 0.75) with randomized Lavukaleve and Bilua wordlists. The observed match rate (0.98) has a probability of occurring by chance of only 0.00008.

model to the canonical output of the hypothesis. The computational simulation of Greenberg’s matching criteria given here is clearly imperfect: while our algorithm detects most of the matches that Greenberg makes it also matches a reasonably high number of semantically unaligned forms (the output of randomizing the order of the word lists). The error tendency in the model is towards over-matching. Over-matching is however not as much of a problem for the test as under-matching would be. Under-matching would tend to obscure any signal of lexical similarity that might be present in the data, increasing the likelihood of a null result. The effect of over-matching is to add additional noise to the data. Randomized and aligned lists are equally subject to the effect of an overly liberal matching algorithm, and so if (as occurs with Greenberg’s canonical data) there is nevertheless a significantly greater degree of similarity in the aligned lists, we can conclude that the algorithm performs adequately. We conclude that we have established a simple algorithm which correlates with Greenberg’s own criteria for matching. The comparison of the aligned and randomized lists of Greenberg’s cognates shows that this algorithm detects the same signal of similarity that Greenberg found.

2.3 Austronesian loans

Greenberg, well aware of the problem of borrowing, uses a number of criteria to separate loans from genuine cognates. The criteria include the following: cultural vocabulary is more likely to be borrowed than core vocabulary; loans often look more similar than cognates would; and borrowing most likely would affect a single language (“Massive borrowing that is multilateral is improbable,” Croft 2005: xix); if the form-meaning pairs cluster in a single semantic domain loans are more probable; special sound correspondences (i.e. not found in other pairings) may indicate borrowing; and if a form is grammatically analyzable in one language, and not in the other, borrowing is likely (Greenberg 2005b).

As we have seen, the Papuan languages of the Central Solomons have long been known for having some degree of intermixing of Austronesian (AN) vocabulary. In fact, Greenberg was conscious of the problem of Austronesian loans into the Papuan languages of the Solomon Islands, and excluded most of those he identified. The exceptions are instances where “... forms have been included when on balance it seemed that a reasonable case could be made for a non-AN origin in spite of similarities with AN languages” (Greenberg 1971: 816). He also acknowledges that his data was insufficient to exclude Austronesian loans entirely, and that more research is needed (Greenberg 1971: 809). However Greenberg indeed missed many Austronesian loans, and even retained some words which he identified as loans as Papuan cognate sets, which is difficult to reconcile with his stated methodology.

3. Results

We have carried out a series of Oswalt Monte Carlo tests using three different datasets, the first being Greenberg’s (1971). In these tests we assumed that the aligned lists were more similar than would be predicted by chance, to identify a suitable model and parameters for simulation of Greenberg’s matching criteria. To produce a more rigorous test of the Central Solomons subgroup hypothesis, we introduce a better dataset, larger and not specifically selected for apparent cognacy. This is an extended Swadesh list published as Tryon & Hackman (1983). From the same source we also take a sample of Oceanic languages from the same region, so we have Oceanic-Oceanic pairs, Papuan-Oceanic pairs, and Papuan-Papuan pairs. There are more than 90 Oceanic languages in the Solomons; we sampled the nearest Oceanic language to the north of each Papuan language since the Papuan languages tend to be distributed along the southern edges of the archipelago. Finally, we used a list of lexemes from Papuan languages in which probable Oceanic loan words have been removed.

Figure 3 shows the 28 pairwise comparisons of the Tryon & Hackman lexical data for the four Papuan and four Oceanic languages, indicating the probability that the observed match score falls within the random distribution for the 28 possible pairs of four Papuan and four Oceanic languages. The very low p-values for the Oceanic-Oceanic pairs are unsurprising, showing that these languages are lexically very similar. Of the Papuan-Papuan pairs, five out of six have a probability of $p < 0.05$, showing that at least some of these languages are more lexically similar than pure chance would predict. However, of the mixed Oceanic-Papuan pairs, eleven out of the sixteen also have a probability < 0.05 . In fact, the two most definitely similar languages are Papuan Savosavo and Oceanic Nggela. The mixed pairs of languages are uncontroversially unrelated, which means that their greater

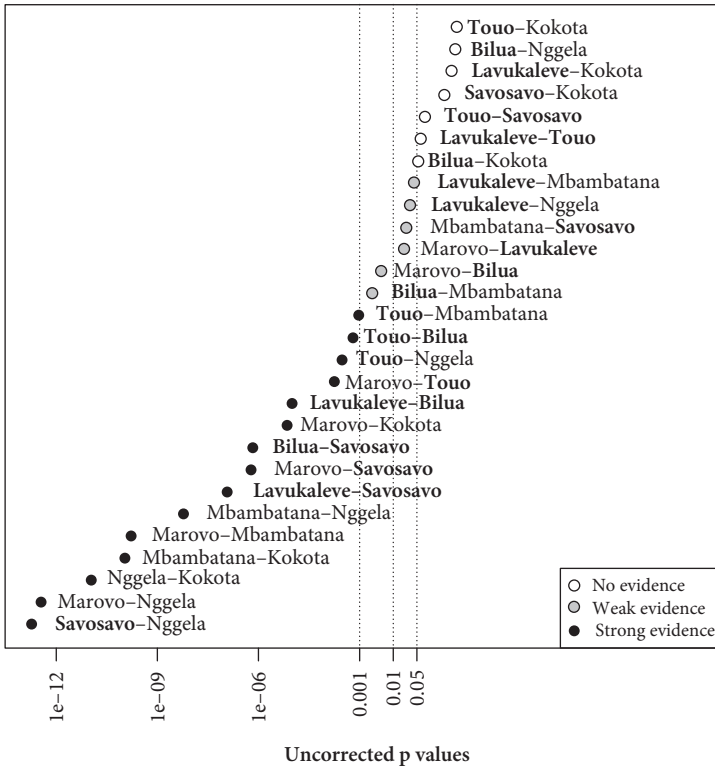


Figure 3. Probability that wordlists from pairs of languages are more similar than would be predicted by chance. The names of Papuan languages are indicated by bold text; the other languages are Oceanic/Austronesian. The p-value for statistical significance is somewhere between 0.05 and 0.0023 (the Bonferroni correction for multiple testing, $0.05/22$); values within this range are indicated by gray dots, which we interpret as weak evidence of greater-than-chance similarity. White dots indicate no evidence of greater-than-chance levels of similarity, and black dots show strong evidence.

than chance levels of lexical similarity must be due to borrowing. Note that the conventional definition of statistical significance — conventionally defined as a 1 in 20 probability of occurring by chance — is inappropriate when carrying out 22 tests. There are a number of statistical techniques to compensate for multiple testing, but we use the simplest, the Bonferroni correction, which simply divides the p-value for significance by the number of tests. The Bonferroni correction tends to be overly strict, so the true p-value for significance (assuming we are satisfied by the statistical significance of $p=0.05$ for a single test) should be somewhere between 0.05 and 0.0023 (i.e. $0.05/22$); this range we call ‘weak evidence’.

Borrowing has definitely played a role in the resemblances between the Oceanic-Papuan languages in the Solomon Islands, although mass borrowing has not occurred (Terrill 2003b), for an examination of the very small number of loans between Oceanic languages and Lavukaleve). The previous figure showed that borrowing has been a factor in the lexical similarities between each of the Papuan languages and its Oceanic neighbours, since most of the Papuan-Oceanic pairs show greater than chance similarity. It seems probable that Oceanic loans into Papuan have also played a role in the similarities between the lexical items in the Papuan languages.

To test the extent to which the lexical similarities between Papuan languages are due to shared Oceanic loans, it is necessary to remove the loanwords from these lists. It proved impossible to reliably and consistently remove all Oceanic loanwords from the Tryon & Hackman lists from published reconstructions of Oceanic or its subgroups. Many apparent loanwords in Papuan languages are regionally delimited to parts of the Solomon Islands; they are present in low level Oceanic subgroups, and are not part of Proto-Oceanic, so do not appear in published references. We instead devised a procedure to identify probable loanwords using lexical distance (cf. van der Ark et al. 2007). The distance was measured between each Papuan term and every semantically corresponding form in the Tryon & Hackman comparative word list. We report two conditions:

Condition 1: Possible loanwords into Papuan are identified where there are terms closer than a specified threshold in at least two Oceanic subgroups. Oceanic subgroups in the Solomon Islands are strongly associated with archipelagos, geographic areas within which interaction is easy. If a term exists in two different subgroups then we judge it less likely that the term is a Papuan loanword into Oceanic.

Condition 2: Possible loanwords are identified where there exist any Oceanic terms closer than the threshold. This is intended to test the tendency in the data: Bilua and Touo are both geographic neighbours of the New Georgia subgroup; Lavukaleve and Savosavo are both neighbours of Guadalcanal, and so

it is possible that Condition 1 is too strict, and that there may be many shared loans from Oceanic terms originating from a single Oceanic subgroup.

The archiphonemic distance measure was not used for detecting possible Oceanic loanwords, since it was impractical to render all 111 languages and dialects in Tryon & Hackman into archiphonemic equivalence tables. Instead we used scaled Levenshtein distance on consonants, with a threshold determined empirically at 0.5, the distance under which most Papuan terms seemed to be plausible cognates with their Oceanic comparands. Consonants-only scaled Levenshtein was the second best performing algorithm from Figure 2.

Figure 4 shows plots of the probability of finding the observed degree of lexical similarity between language pairs before and after probable Oceanic loanwords have been excluded from the Papuan wordlists. As expected, the similarities between mixed pairs of languages (one Oceanic, one Papuan) decrease markedly, but

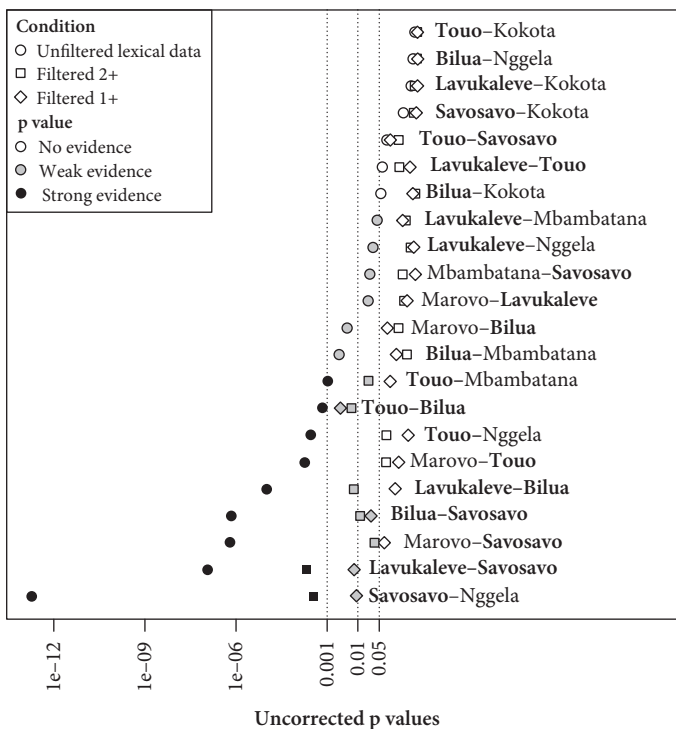


Figure 4. Probability that wordlists from pairs of languages are more similar than would be predicted by chance, corrected for probable loanwords. The circles show the unfiltered results from Figure 3; squares show the results once forms which match words in two or more Oceanic subgroups are removed; diamonds show results once forms are filtered to remove all forms which match words in at least one other Oceanic language.

there is also a dramatic decrease in the similarity detected between pairs of Papuan languages. Note that no terms were filtered from Oceanic wordlists, so the variation in the p-values for the Oceanic pairs gives a sense of the inherent variation in a measure determined by a randomization procedure.

Under condition 1, with the constraint that possible Oceanic loanwords have to be attested in at least two different Oceanic subgroups before they are excluded, the highly similar Savosavo-Nggela wordlists are still similar, but no longer extremely so. Under this condition all the Oceanic-Oceanic pairs are more certain than any other pairing, which is what would be expected from our prior knowledge of language relationships in the area. Under condition 2, only weak evidence for the Savosavo-Nggela connection remains.

The strongest Papuan-Papuan pairing is Lavukaleve-Savosavo. The change in the p-values from the comparison of the unfiltered lists to the lists with Oceanic loanwords removed shows that a large part of the lexical similarity between these languages is due to shared Oceanic content. The relatively large difference between condition 1 and condition 2 is interesting: Lavukaleve and Savosavo share a lot of lexical items otherwise only found in the Oceanic languages of the local region. This could plausibly be a Papuan substrate in Oceanic (either from some ancestor of Lavukaleve and/or Savosavo, or from some third now-extinct Papuan language of pre-Austronesian Guadalcanal), but the data offers no reason to prefer this hypothesis over the hypothesis of Oceanic loans into Papuan.

All four Papuan languages of the Solomons can be linked by a chain of weak signals of similarity: Lavukaleve-Savosavo, Savosavo-Bilua, and Bilua-Touo. It is notable that like Lavukaleve-Savosavo, Touo-Bilua are both neighbours to the same Oceanic subgroup. Touo-Savosavo and Lavukaleve-Touo are no more similar than chance under any condition; Lavukaleve-Bilua is weakly supported under condition 1, but unsupported under condition 2.

There still must be undetected loans in these lists, since some of the Papuan-Oceanic pairs still show (weakly) greater than chance levels of similarity, despite being genealogically unrelated. It may be that this is a trace of Papuan loanwords in the Oceanic languages. In either case this suggests that the signals of relatedness that we have detected between the Papuan languages are somewhat inflated due to further undetected loanwords.

4. Conclusions

The Oswalt Monte Carlo test fails to find a convincing signal of greater than chance lexical similarity between the Papuan languages of the Solomon Islands. This contradicts the evidence from structural phylogenetics, which suggests on the basis of

typological diversity that there has been some kind of intensive historical interaction between these languages. One possible conclusion to draw from this would be that in this specific instance, language structure seems to have been more stable than the lexicon. In general we judge that a false negative in the Oswalt Monte Carlo test is more likely than a false positive in the spatial autocorrelation test on structural features of language, especially given the archaeological and genetic evidence for long term, intensive interaction in the pre-Austronesian period.

A subsidiary aim was to examine Greenberg's methods and evidence for his claim for the relatedness of the Central Solomons Papuan languages, a claim based largely on lexical evidence. We cannot disprove that these languages are related, but we have proven that Greenberg has not shown these languages to be related. Typological comparison has shown that the genealogical unity of Solomons Papuan is a plausible hypothesis, and it is perhaps surprising that the lexical signal of similarity is so weak.

Greenberg was aware of the dangers of positing putative relationships among languages that have undergone mass borrowings. However the present case has been instructive in that there have not been mass borrowings in basic vocabulary at all between Papuan and Austronesian languages: the highest percentages of shared lexicon occur with Bilua, and are in the region of 16%, but for all the other languages the highest percentage is much lower than this. For languages that must have been in contact for thousands of years this is an extremely small amount of lexical borrowing.

But even at relatively low levels, shared Austronesian vocabulary has been enough to inflate the level of apparent cognacy between the Papuan languages. This has swamped any possible signal of true genealogical relationships between the Papuan languages, since the level of possible true cognates is lower than the level of shared loans.

This study highlights the importance of triangulation: the possibility of obtaining a far clearer picture of linguistic relationships if there is language-external evidence of the histories of the languages in question. Thus, we know that there is a huge amount of time between earliest settlement of the area and the arrival of the Lapita peoples. This means that a clear statement on the language relationships between the Papuan languages may remain beyond reach. On the other hand, relationships between the Papuan and Oceanic languages are of a much shorter time-scale and we may therefore expect them to be more amenable to exploration by our current linguistic tools.

References

- van der Ark, Rene, Philippe Mennecier, John Nerbonne & Franz Manni. 2007. "Preliminary Identification of Language Groups and Loan Words in Central Asia". *Proceedings of the RANLP Workshop on Computational Phonology* ed. by Petya Osenova, 13–20.
- Baxter, William H. 1995. "A Stronger Affinity ... Than Could Have Been Produced by Accident: A probabilistic comparison of Old Chinese and Tibeto-Burman". *The Ancestry of the Chinese Language* ed. by William Shi-Yuan, 1–39. Berkeley: Project on Linguistic Analysis.
- Baxter, William H. 1998. "Response to Oswalt and Ringe". *Nostratic: Sifting the evidence* ed. by Joseph C. Salmons & Brian D. Joseph, 217–236. Amsterdam: John Benjamins.
- Capell, Arthur. 1969. "Non-Austronesian Languages of the British Solomons". *Papers in the Linguistics of Melanesia* 2:1–16.
- Codrington, Robert Henry. 1885. *The Melanesian Languages*. Oxford: Clarendon Press.
- Cox, Murray P. & Martha Mirazón Lahr. 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.
- Croft, William. 2005. "Editor's Introduction". *Genetic Linguistics: Essays in theory and method* ed. by William Croft, xi–xxxvi. Oxford: Oxford University Press.
- Crowley, Terry & R. M. W. Dixon. 1981. "Tasmanian". *Handbook of Australian Languages Vol 2* ed. by R. M. W. Dixon & Barry Blake, 395–421. Canberra: Australian National University Press.
- Dunn, Michael. 2009. "Contact and Phylogeny in Island Melanesia". *Lingua* 119:11.1174–1678.
- Dunn, Michael & Malcolm Ross. 2007. "Is Kazukuru Really Austronesian?" *Oceanic Linguistics* 46:1.210–231.
- Dunn, Michael, Ger Reesink & Angela Terrill. 2002. "The East Papuan Languages: A preliminary typological appraisal". *Oceanic Linguistics* 41:1.28–62.
- Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley & Stephen C. Levinson. 2005. "Structural Phylogenetics and the Reconstruction of Ancient Language History". [Science](#) 309.2072–2075.
- Dunn, Michael, Stephen C. Levinson, Eva Lindström, Ger Reesink & Angela Terrill. 2008. "Structural Phylogeny in Historical Linguistics: Methodological explorations applied in Island Melanesia". *Language* 84:4.710–759.
- Ellison, T. Mark & Simon Kirby. 2006. "Measuring Language Divergence by Intra-lexical Comparison". *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia*. Association for Computational Linguistics.
- Friedlaender, Jonathon S., ed. 1987. *The Solomon Islands Project: A long-term study of health, human biology, and culture change*. Oxford: Oxford University Press.
- Good, Philip. 2006. *Resampling Methods: A practical guide to data analysis*. 3rd ed. Boston: Birkhäuser.
- Greenberg, Joseph H. 1971. "The Indo-Pacific Hypothesis". *Current Trends in Linguistics Vol. 8: Linguistics in Oceania* ed. by Thomas A. Sebeok, 807–871. The Hague: Mouton de Gruyter.
- Greenberg, Joseph H. 1987. *Language in the Americas*. Stanford CA: Stanford University Press.
- Greenberg, Joseph H. 2001. "The Methods and Purposes of Linguistic Genetic Classification". *Language and Linguistics* 2:2.111–135.

- Greenberg, Joseph H. 2005a. *Genetic Linguistics: Essays on theory and method*. Oxford: Oxford University Press.
- Greenberg, Joseph H. 2005b. "Historical Linguistics and Unwritten Languages". *Genetic Linguistics: Essays on theory and method* ed. by William Croft, 3–31. Oxford: Oxford University Press.
- Heeringa, Wilbert Jan. 2004. *Measuring Dialect Pronunciation Differences Using Levenshtein Distance*. PhD dissertation, University of Groningen.
- Heeringa, Wilbert Jan, Peter Kleiweg, Charlotte Gooskens & John Nerbonne. 2006. "Evaluation of String Distance Algorithms for Dialectology". *Proceedings of the Workshop on Linguistic Distances*, 51–62. Morristown, NJ: Association for Computational Linguistics.
- Hunley, Keith, Michael Dunn, Eva Lindström, Ger Reesink, Angela Terrill, Meghan E. Healey, George Koki, Françoise Friedlaender & Jonathon S. Friedlaender. 2008. "Genetic and Linguistic Coevolution in Northern Island Melanesia". *PLoS Genetics* 4:10.e1000239.
- Justeson, John S. & Laurence D. Stephens. 1980. "Chance Cognation: A probabilistic model and decision procedure for historical inference". *Papers from the 4th International Conference on Historical Linguistics* ed. by Elizabeth C. Traugott, Rebecca Labrum & Susan Shepherd, 37–45.
- Kessler, Brett. 1995. "Computational Dialectology in Irish Gaelic". *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 60–67. Dublin: European Association for Computational Linguistics.
- Kessler, Brett. 2007. "Word Similarity Metrics and Multilateral Comparison". *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, 6–14. Stroudsburg PA: Association for Computational Linguistics.
- Kessler, Brett & Annukka Lehtonen. 2006. "Multilateral Comparison and Significance Testing of the Indo-Uralic Question". *Phylogenetic Methods and the Prehistory of Languages* ed. by Peter Forster & Colin Renfrew, 33–42. Cambridge: McDonald Institute for Archaeological Research.
- Kirch, Patrick. 1997. *The Lapita Peoples*. London: Blackwell.
- Kondrak, Grzegorz & Tarek Sherif. 2006. "Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification". *Proceedings of the Workshop on Linguistic Distances*, 43–50. Sydney: Association for Computational Linguistics.
- Kruskal, Joseph B. 1983. "An Overview of Sequence Comparison: Time warps, string edits, and macromolecules". *Siam Review* 25:2.201–237.
- Lanyon-Orgill, Peter A. 1953. "The Papuan Languages of the New Georgian Archipelago, Solomon Islands". *Journal of Austronesian Studies* 1:1.122–138.
- Lass, Roger. 1984. *Phonology: An introduction to basic concepts*. Cambridge: Cambridge University Press.
- Levenshtein, Vladimir. 1966. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals". *Cybernetics and Control Theory* 10:8.707–710.
- Lincoln, Peter C. 1978. "Reef-Santa Cruz as Austronesian". *Second International Conference on Austronesian Linguistics: Proceedings* ed. by Stephen A. Wurm & Lois Carrington, 929–967. Canberra: Pacific Linguistics C-61.
- McMahon, April & Robert McMahon. 2005. *Language Classification by Numbers*. Oxford: Oxford University Press.
- Metropolis, Nicholas. 1987. "The Beginning of the Monte Carlo Method" *Los Alamos Science (1987 Special Issue dedicated to Stanislaw Ulam)* 15.125–130.

- Metropolis, Nicholas & S. Ulam. 1949. "The Monte Carlo Method". *Journal of the American Statistical Association* 44:247.335–341.
- Næss, Åshild & Brenda H. Boerger. 2008. "Reefs-Santa Cruz as Oceanic: Evidence from the verb complex". *Oceanic Linguistics* 47:1.185–212.
- Nerbonne, John, Wilbert Jan Heeringa & Peter Kleiweg. 1999. "Edit Distance and Dialect Proximity". *Time Warps, String Edits and Macromolecules: The theory and practice of sequence comparison* ed. by David Sankoff & Joseph B. Kruskal, v–xv. Stanford CA: CSLI.
- Obata, Kazuko. 2003. *A Grammar of Bilua: A Papuan language of the Solomon Islands*. Canberra: Pacific Linguistics 540.
- Oswald, Robert L. 1970. "The Detection of Remote Linguistic Relationships". *Computer Studies in the Humanities and Verbal Behavior* 3.117–129.
- Oswald, Robert L. 1998. "A Probabilistic Evaluation of North Eurasian Nostratic". *Nostratic: Sifting the evidence* ed. by Joseph C. Salmons & Brian D. Joseph, 199–216. Amsterdam: John Benjamins.
- Pawley, Andrew K. 2007. "Recent Research on the Historical Relationships of the Papuan Languages, or, What does Linguistics say about the Prehistory of Melanesia?" *Genes, Language and Culture History in the Southwest Pacific* ed. by Jonathon S. Friedlaender, 36–58. Oxford: Oxford University Press.
- Ray, Sidney Herbert. 1926. *A Comparative Study of the Melanesian Island Languages*. London: Cambridge University Press.
- Ray, Sidney Herbert. 1928. "The Non-Melanesian Languages of the Solomon Islands". *Festschrift publication d'hommage offerte au P. W. Schmidt* ed. by Wilhelm Koppers, 123–126. Vienna: Mechtharisten-Congregations-Buchdruckerei.
- Ross, Malcolm. 1988. *Proto Oceanic and the Austronesian Languages of Western Melanesia*. Canberra: Pacific Linguistics C-98.
- Ross, Malcolm. 2001. "Is There an East Papuan Phylum? Evidence from pronouns". *The Boy from Bundaberg: Studies in Melanesian linguistics in honour of Tom Dutton* ed. by Andrew Pawley, Malcolm Ross & Darrell Tryon, 301–321. Canberra: Pacific Linguistics.
- Ross, Malcolm & Åshild Næss. 2007. "An Oceanic Origin for Aiwoo, the Language of Reef Islands". *Oceanic Linguistics* 46:2.456–498.
- Sankoff, David & Joseph B. Kruskal. 1983. *Time Warps, String Edits, and Macromolecules: The theory and practice of sequence comparison*. Reading MA: Addison-Wesley.
- Schmidt, Pater Wilhelm. 1902. "Die sprachlichen Verhältnisse von Deutsch-Neuguinea". *Zeitschrift für afrikanische, ozeanische und ostasiatische Sprachen* V.354–384, VI.1–99.
- Spriggs, Matthew. 1997. *The Island Melanesians*. London: Blackwell.
- Terrill, Angela. 2002. "Systems of Nominal Classification in East Papuan Languages". *Oceanic Linguistics* 41:1.63–88.
- Terrill, Angela. 2003a. *A Grammar of Lavukaleve*. Berlin: Mouton de Gruyter.
- Terrill, Angela. 2003b. "Linguistic Stratigraphy in the Central Solomon Islands: Lexical evidence of early Papuan/Austronesian interaction". *Journal of the Polynesian Society* 112:4.369–401.
- Todd, Evelyn. 1975. "The Solomon Language Family". *Papuan Languages and the New Guinea Linguistic Scene* ed. by Stephen A. Wurm, 805–846. Canberra: Pacific Linguistics C-38.
- Trask, Robert Lawrence. 1996. *Historical Linguistics*. London: Arnold.
- Tryon, Darrell T. & B. D. Hackman. 1983. *Solomon Island Languages: An internal classification*. Canberra: Pacific Linguistics C-72.
- Usher, Timothy. 2002. "The Origin of the Tasmanian Language". *Mother Tongue* 7.65–84.

- Wegener, Claudia. 2008. *A Grammar of Savosavo: A Papuan language of the Solomon Islands*. Phd dissertation, Max Planck Institute series in Psycholinguistics no. 51, Nijmegen.
- Wieling, Martin, Jelena Prokić & John Nerbonne. 2009. "Evaluating the Pairwise String Alignments of Pronunciations". *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities and Education/(LaTeCH — SHELTeR) Workshop at the 12th Meeting of the European Chapter of the Association for Computational Linguistics* ed. by Lars Borin. & Piroška Lendvai, 26–34.
- Wurm, Stephen A. 1971. "The Papuan Linguistic Situation". *Current Trends in Linguistics Vol. 8: Linguistics in Oceania* ed. by Thomas A. Sebeok, 541–657. The Hague: Mouton de Gruyter.
- Wurm, Stephen A. 1972. "The Classification of Papuan Languages and Its Problems". *Linguistic Communications* 6.118–178.
- Wurm, Stephen A. 1975. "The East Papuan Phylum in General". *Papuan Languages and the New Guinea Linguistic Scene* ed. by Stephen A. Wurm, 783–804. Canberra: Pacific Linguistics C-38.
- Wurm, Stephen A. 1978. "Reefs-Santa Cruz: Austronesian, but...!" *Second International Conference on Austronesian Linguistics: Proceedings* ed. by Stephen A. Wurm & Lois Carrington, 969–1010. Canberra: Pacific Linguistics C-61.

Résumé

En l'absence de reconstruction par les méthodes de la grammaire comparée, on se sert souvent du pourcentage de lexique partagé comme argument afin de démontrer que deux langues sont apparentées. Le présent article se sert du test "Monte Carlo Shift" d'Oswalt (d'après Oswalt 1970) afin d'évaluer les fondements statistiques de l'hypothèse selon laquelle les quatre langues papoues des îles Salomon ont des taux de similarité lexicale qui dépassent le hasard. Les premiers résultats de ce test paraissent indiquer que les ressemblances lexicales entre les langues papoues du Centre des Salomon sont significatifs, d'un point de vue statistique: toutefois, l'effet disparaît dès lors que l'on supprime les emprunts lexicaux provenant de langues océaniques. Ainsi, le test Monte Carlo d'Oswalt est une technique efficace lorsqu'il s'agit de vérifier un cas de ressemblance entre deux listes de mots, et tester ladite ressemblance dépasse une simple ressemblance due au hasard. Cependant, les emprunts non identifiés risquent toujours de déformer les résultats.

Zusammenfassung

Wenn keine Rekonstruktionen durch die Anwendung der vergleichenden Methode zur Verfügung stehen, dann wird oft auf eine Vielzahl von potentiell verwandten lexikalischen Wörtern als Beweis für die Verwandtschaft von Sprachen zurückgegriffen. In diesem Artikel wird der Oswalt Monte-Carlo-Shift-Test (eine Variante von Oswalt 1970) verwendet, um die statistische Basis der Hypothese zu überprüfen, die besagt, dass der Grad der lexikalischen Ähnlichkeit zwischen den vier Papuasprachen der Salomon Inseln nicht zufällig sei. Erste Ergebnisse dieses Tests zeigen in der Tat, dass die lexikalische Ähnlichkeit zwischen den Papuasprachen der Zentral-Salomonen statistisch signifikant ist, aber dieser Effekt verschwindet, wenn bekannte

ozeanische Lehnwörter bei der statistischen Analyse nicht berücksichtigt werden. Das bedeutet, dass der Oswald Monte-Carlo-Test durchaus eine nützliche Methode ist, um Hypothesen zu testen, die besagen, dass die Ähnlichkeit zwischen zwei beliebigen Wortlisten nicht zufällig sei — allerdings mit dem Vorbehalt, dass nicht erkannte Lehnwörter die Wahrscheinlichkeit einer Pseudo-Identifikation stark erhöhen.

Authors' addresses

Michael Dunn
Max Planck Institute for Psycholinguistics
PB 310
6500 AH NIJMEGEN
The Netherlands
michael.dunn@mpi.nl

Angela Terrill
Radboud University Nijmegen
Max Planck Institute for Psycholinguistics
PB 310
6500 AH NIJMEGEN
The Netherlands
angela.terrill@mpi.nl