Journal of Universal Language 7
September 2006, 121-145

Significant Lexical Similarities between a Language of Brazil and Some Languages of Southeast Asia and Oceania: From Typolocial Perspective

Vladimir Pericliev
Bulgarian Academy of Sciences

Abstract

The paper examines computationally the similarities in 100-word lists of basic vocabulary between Xokleng (a language of southeastern Brazil, classified as Macro-Ge) and Tagalog and Malay (languages of Southeast Asia) and Fijian, Samoan, and Hawaiian (languages of Oceania). It is found that in all five pair-wise comparisons the resemblances found are statistically highly significant (i.e., are greater-than-chance). A plausible explanation of these results is a possible historical (i.e., genetic or diffusional) relationship between these languages, a conjecture which is in accord with our previous studies, as well as with some contemporary genetic investigations indicating the existence of genetic affinities between Brazilian Indians and Southeast Asian and Oceanic populations. The hypothesis suggested, however,

* This work was partly supported by grant MI-1511/2005 from the Bulgarian Ministry of Education and Science.

requires a thorough historical linguistic test, including also other relevant languages. One of the basic goals of the paper is to stimulate such test.

Keywords: Macro-Ge, Austronesian, language classification, application of computational methods

1. Introduction

Languages may resemble each other in their phonology, grammar or lexicon. Observed similarities between languages may, in principle, be due to different factors. First, the languages in question may be "historically related", in the sense that they are either genetically affiliated and have common descent or, alternatively, have been in contact and one language has borrowed from the other, possibly through a mediation of yet another language. A second reason to account for noted similarities between languages is their obeying universal principles and tendencies. And last but not least, languages can be similar in certain respects due to mere chance.

In this paper, I look at lexical similarities between languages that are guaranteed to be greater-than-chance. Insomuch as there can be no universal principles or tendencies requiring any lexical resemblances, or what is the same, that the same form in different languages should be associated with the same meaning, languages exhibiting such lexical similarities would probably be historically related in some sense. In particular, a computer program which inspects wordlists of a pair of languages and estimates the statistical significance of the form-meaning (=lexical) resemblances found in the examined languages, is used in comparing geographically distant languages. The languages are Xokleng of Southeastern Brazil (and currently classified as Macro-Ge) and five major Austronesian languages, of Southeast Asia (Tagalog and Malay) and of Oceania

(Fijian, Samoan and Hawaiian). Most surprisingly, the similarities found in pair-wise comparisons were significantly higher than could be expected by mere chance. This result suggests some type of historical relationship of Xokleng with Austronesian, a conjecture corroborated both by our previous investigations revealing affinity of Xokleng to Austronesian in its kinship semantics and some phonological and grammatical features, and some recent research in population genetics.

The paper is organized as follows. Section 2 examines a dozen lexemes in Xokleng as as compared to Tagalog, Malay, Fijian, Samoan and Hawaiian by Greenberg's method of mass comparison, which is a useful heuristic for noticing a potential relationship between languages. This preliminary stage of investigation is shown to yield quite suggestive, but statistically undetermined, and hence not fully reliable results. Then, in Section 3, I briefly introduce the computer program that estimates the statistical significance and discuss its outputs from examining 100-item wordlists of the investigated languages. In Section 4, some lexical similarities are shown between Xokleng and other Austronesian languages. Section 5 makes some concluding remarks by placing these findings in the context of population genetic and other results, including our own linguistic evidence, which, taken collectively, further corroborate the plausibility of potential affinity of Xokleng to Austronesian.

2. Applying the Method of Mass Comparison

Xokleng, commonly known also as Aweikoma esp. among anthropologists (with alternate names Shokleng, Kaingang, Bugre, Botocudos), is spoken by about 780 people in south-eastern Brazil, viz., in Santa Catarina, along the tributary of the Itajaí River. Its current classification according to *Ethnologue* (Gordon 2005) and Ruhlen (1987) is Macro-Ge/Ge-Kaingang/Kaingang/Northern. A

basic ethnographical source is Jules Henry (1941). The language has been studied by Gensch (1908), Henry (1935, 1948), and Urban (1985) (cf., Wiesemann 1986, Rodrigues 1999). In what follows, our data will come basically from the studies by Gensch and Henry.[1]

Tagalog, Malay, Fijian, Samoan and Hawaiian are basic languages of the much better studied Austronesian language family and hence do not require any special introduction. We will only mention that the languages are representatives of major branches of the very widely spread Austronesian language family, viz., the Western Malayo-Polynesian branch (Tagalog, Malay) and the Oceanic branch (Fijian, Samoan and Hawaiian). In the discussion below, the data for Austronesian comes from the authoritative *Austronesian Basic Vocabulary Database* (at URL *http://language.psy.auckland.ac.nz*; citation is made with the permission of the authors). Currently, the database contains information on 200 basic vocabulary items of around 350 Austronesian languages.

The method known as "mass comparison" is mostly familiar from the work of Joseph Greenberg (1957), though as he acknowledges, it is the oldest method used by linguists for arriving at plausible conjectures as to whether or not languages are historically related. The method comprises of compilation and inspection of wordlists of basic vocabularies (one of the most stable aspects of language) of the languages to be compared. Greenberg views the procedure as "the swiftest and surest" heuristic (Greenberg 1957: 42) for accomplishing the task, as it makes conspicuous various aspects of the similarities observed.

Table 1 presents a comparative wordlist of 14 lexical items from Xokleng and the five Austronesian languages it is compared to (the

[1] The notation of the older source, Gensch (1908), is slightly altered wherever possible to be compatible with the broad transcription used by Henry. Also, some apparent affixation is segmented off words in the same source, as e.g., the predicating suffixes *-ma*, *-mu*, *-ko*, etc., which natives always pronounce together with nouns and verbs (but are absent in Henry's description).

apostrophe in Hawaiian orthographically renders the glottal plosive, denoted by 'ʔ' in Samoan).

Table 1. A Comparative Wordlist of Xokleng and Five Austronesian Languages

	Gloss	Xokleng	Tagalog	Malay	Fijian	Samoan	Hawaiian
1	come	*katéŋ*	*datíŋ*	*dating*	*lakomai*	*o mai*	*mai*
2	ear	*niŋná*	*taqiŋa*	*teliŋa*	*daliŋa-*	*taliŋa*	*pepeiao*
3	fear	*ŋai-*	*tákut*	*takut*	*rere-*	*ma-taʔu*	*ma-ka'u*
4	feathers	*kulá*	*buhog*	*bulu*	*vuti-na*	*fulu*	*hulu*
5	five	*pélemo*	*lima*	*lima*	*lima-na*	*lima*	*lima*
6	four	*mpét*	*apat*	*empat*	*evā*	*efa*	*ha*
7	leg	*pa*	*pa(qa)*	*kaki*	*yavana*	*wae*	*wawae*
8	manure	*kaé-wé*	*tae*	*tahi*	*ndena*	*tae*	*kae*
9	rain	*úgua*	*ulan*	*hujan*	*uca*	*timu*	*ua*
10	shoot	*pænṹ*	*barilin*	*menembak*	*vana*	*fana*	*pana*
11	stab/kill	*patɨ*	*patáy*	*membunuh*	*mokuta*	*kisioti*	*ho'omake*
12	sun	*laɨ*	*araw*	*mata-hari*	*siga*	*la*	*la*
13	three	*kél*	*tatlo*	*tiga*	*e tolu*	*e tolu*	*kolu*
14	we(incl/excl)	*aŋ háma*	*kami*	*kami*	*keda*	*tatou*	*kami*

Even a cursory glance at the data makes obvious some association of the Brazilian language to the Austronesian languages.

In the first place, there is a notable resemblance in *form* between the Xokleng words and the corresponding words in two or more of the other languages in most of the inspected lexemes. One example is the formal overlaps in the words for 'five' in all languages (which would be no surprise, recalling that the reconstructed Proto-Malayo-Polyneisan form is **(qa)lima*). As further illustrations, consider e.g.,

the words for ‘four’ in Xokleng and Tagalog and Malay (the Proto-Malayo-Polynesian form being **epat*), the identical words for ‘sun’ in Xokleng, Hawaiian and Samoan, the words for ‘come’ in Xokleng, Tagalog and Malay, etc.

In the second place, some (indeed only very tentative) *sound correspondences* emerge between Xokleng and the other languages. To give an example, Xo *k* = Haw *k* (as a reflex of Proto-Malayo-Polynesian (PMP) **t*), as suggested by PMP **telu* ~ Xo *kél* ~ Haw *kelo* ‘three’, PMP **taqi?* ~ Xo *kaé-wé* ~ Haw *kae* ‘shit, manure’ or PMP **ma-takut* ~ Xo *ŋai-kaúg*^{n} ~ Haw *ma-ka’o* ‘fear’. The other compared languages (except Fijian), in contrast, will be seen in Table 1 to preserve the original PMP sound **t* in these cases.

Finally, one may notice in Table 1 a *morphological similarity* between Xokleng and some of the other languages, which is also highly valued in assessing relationships, viz., the one observable in the words for ‘fear’. Thus, Xokleng *ŋai-kaúg*^{n} ‘fear, be frightened’, quite analogously to Hawaiian *ma-ka’o* and Samoan *ma-ta?u* (cf., also PMP **ma-takut*) consists of a prefix (viz., *ŋai-*, also used as a reflexive ‘self’) and a stem (viz., *kaúg*^{n}). The prefixes and stems of the languages are formally similar. In addition, it turns out that the prefixes are similar in function as well. The Austronesian (Proto-Oceanic) *ma-* is known to have a valency-decreasing function (e.g., Evans & Ross 2001), i.e., prefixing it to a verbal stem reduces the number of arguments this verb may have. Xokleng’s prefix *ŋai-* behaves in exactly the same way. Henry (1935, 1948) draws the attention to the latter fact of Xokleng grammar. He writes that “Certain verbs that begin with *ŋai* omit it when they have direct objects” (Henry 1935: 204), giving among others examples with the verb ‘fear/be frightened’ (curiously, this verb is commonly given as an illustration of the same phenomenon in Austronesian).

One can go on enumerating similarities, but we needn’t do that here. How can one interpret these results from mass comparison? Most linguists would probably find them suggestive, but far from

conclusive, and would additionally require, undoubtedly, further evidence indicating that the similarities found are not due to sheer chance. In mass comparison, languages are not compared binarily. Thus, a word in one language is compared to a *set of words* in other languages, which, quite obviously, increases the probability of chance matches, and one may wish to argue that since there are so many Austronesian languages (possibly over 1000) we can easily find a set of Xokleng words matching words in one or more different languages. Both statistical science and linguistic practice have taught us the lesson that coincidental resemblances do occur in language comparison and therefore we should be able to assess the statistical significance of the results obtained. Mass comparison, as usually practiced without estimation of statistical significance, is thus a useful heuristic, but an untrustworthy proof of a relationship.

In the next section, I describe a computer program that does the job, and then run it on 100-item comparative lists of basic vocabulary of these languages. Rather than look for a match between Xokleng and some of the other languages, Xokleng will be compared pair-wise with each of the Austronesian languages, and the significance of the resemblances found estimated.

3. The Machine Lexical Comparison

3.1. The Program

To be able to compare the lexicons of languages and evaluate the statistical significance of the similarities found, we first need to compile comparative wordlists of pairs of these languages, in our case Xokleng-Hawaiian, Xokleng-Tagalog, Xokleng-Malay, Xokleng-Fijian and Xokleng-Samoan.

We selected 100 basic lexical meanings (≈a Swadesh list), as follows: and, arm/hand, ashes, bad, beat, belly/stomach, big, bird,

black/dark-brown, bone, breast, burn, child, cloud, cold, come, cut (wood), die/dead, dig, dream, dry, ear, eat, egg, eye, fall, far, fat, father, fear, feather, fire, fish, five, four, fruit, grass, green/blue, hair, head, hear, hit, strike, hunt, husband, I, in, inside, kill, leaf, leg/foot, lie down (sleep), lips, long, lot (a lot), louse, man, manure/shit, mother, mouth, no/not, nose, old, person/human being, pierce/stab, plant, rain, red, road/path, rope/cord, sand, say, see, sharp(en), shoot, sick, skin, sky, small, spider, split, stick (wood), stone, sun, tail, three, throw, thunder, tongue, tooth, understand, water, we (incl/excl), wet, white, wife, wind, wing, woman, woods, forest, yellow, you (sg/pl).

Then we filled the above meaning slots with the Austronesian words, as they are given in *The Austronesian Basic Vocabulary Database*.[2] As far as Xokleng is concerned, the list was compiled basically from Gensch (1908) and Henry (1935, 1948). In the cases when more than one word was available to fill a slot (either because the two sources supply distinct forms or one source gives synonymous forms), one form was randomly selected in order to maintain the impartiality of the method.

A computer program was built that handles comparative wordlists of two languages. It allows the user to flexibly define various criteria of phonetic similarity for a pair of forms. The similarity criteria to be used here will be described shortly.

To estimate the statistical significance of the phonetic similarities between a pair of languages, one needs to find the number of matching pairs, according to the chosen criteria for phonetic similarity, and compare them to the number of matches that could be expected to occur within the range of chance. We used a standard statistical procedure, viz., so-called "permutation test", to make the estimation (cf., Good 1994 for a general discussion).

[2] In a couple of cases, other sources are used, as for instance when the database does not include a word from our 100-word list.

Seemingly the first applications of the idea to the task at issue are Oswalt (1970) and Oswalt (1991), and I follow his method here; a more recent work is Kessler (2001); another application of the permutation method, to finding universals, is described in Valdés-Pérez & Pericliev (1999).

In a permutation test, the number of matches found in the original, or non-permuted, data (called *actual score*) is compared to those found in many (usually 1000 or more) random permutations of the original data. In our case, if we imagine a comparative wordlist of two languages as comprising two columns, with words with identical meanings set against each other, a random permutation will scramble one of the columns, so that forms will be compared that no longer have identical meaning. Our data was permuted 1000 times, and the *random mean* was calculated, which is the average of the number of matches obtained in the 1000 random permutations. The *actual deviation* is computed (=actual score minus random mean), which is a figure, indicating the number of greater-than-chance similarities.

The distribution of random scores is sufficiently close to a normal (bell-shaped) curve, which allows the computation of statistical significance, which is a better measure of the strength of the relationship than the actual deviation, as follows. The *standard deviation*, a measure of the dispersal of random scores, is computed as the square root of the mean of the squares of the deviations of each random score from the random mean. The *standard score* (or *z-score*) is computed by dividing the actual deviation by the standard deviation. The standard score is an important figure in these computations, as it allows us to determine the probability (or significance) of the similarities of the examined languages. The higher the standard score, the lower the probability of getting the result by chance and hence the higher the significance. Tables of areas under the standard normal curve (to be found in most statistical textbooks) show the probability (=significance) of finding

such a score or one higher.

Now, in what follows, we give all these figures in the actual language comparisons, as they might be meaningful to linguists familiar with statistics. At the same time, our results from these comparisons should be fully comprehensible also to the uninitiated in statistics, since all one needs to know to fully assess whether the resemblances found between a pair of languages are greater than chance is that the commonly used level of significance is 0.05, but here we prefer the more conservative and reliable level of 0.01, usually referred to as a level of "high significance" (corresponding to a standard score of 2.3 or higher). It could also be mentioned that one unit of standard score means a lot in terms of changing significance (e.g., the change from 2.5 to 3.5 yields a probability change from 0.0062 to 0.0002) and normally statistical tables stop at a standard score of 4, giving a significance of 0.0000 to four decimal places (or practically, a certainty).

3.2. The Comparisons

We have now to define the similarity criteria, or what counts as a match, in comparing two word forms. One can define these criteria more strictly or more loosely, the latter choice generally resulting in increasing the number of matches found, but at the cost of deteriorated formal resemblances and lower significance level. In the statistical approach, in contrast to non-statistical studies usually aiming at unearthing more matches (or putative cognates), how one chooses to define these criteria is not really crucial insofar as the statistical test is in control of the process. What really matters in language comparisons is not the number of matches, but the strength of the relationships, which is reflected in the statistical significance. For the purposes of this paper, we compare all consonant-vowel-consonant (CVC) sequences in a word pair, following so-called "extended criteria method" in a familiar paper by Bender (1969).

The criteria, then, are:

(i) $C^1V^1C^2$ in one language is accepted as corresponding to (=matching) $C^3V^2C^4$ in another if the vowels are identical (disregarding differences in tone or length) and if one or both pairs of consonants are identical while the other pair differs by only one feature (occasionally a difference of two features is accepted; see point (v)).

(ii) $C^1V^1C^2$ in one language is accepted as corresponding to $C^3V^2C^4$ in another if C^1 is identical to C^3 and C^2 is identical to C^4, disregarding of the intervening vowels. This criterion makes explicit use of the general assumption that consonants count for more in correspondences than vowels.

(iii) An item consisting of a CV alone is counted in comparison with an identical CV standing alone or occurring in a larger item.

(iv) An item comprising any sequence of three sounds is counted in comparison with an identical three-sound sequence standing alone or occurring in a larger item.[3]

(v) The following consonantal pairs with more than one feature difference match: v=p, t=l, t=n, h=k, h=g.

There follow the results from running our program on pairs of the inspected languages, with the similarity criteria above.

[3] This criterion does not figure in Bender's extended criteria method, but is included here since found useful in many language comparisons we made outside those reported in this paper.

3.2.1. Xokleng and Hawaiian

Gloss	*Xokleng*	*Hawaiian*	*Results*
burn	pun	kuni	Actual Score: 13
cut	ki	ʻoki	Random Mean: 2.49
feather	kulá	hulu	Actual Deviation: 10.51
five	pélemo	lima	Standard Deviation: 1.56
man	kon-gang	kāne	Standard Score: 6.74
manure, shit	kaé-wé	kae	Probability: 0.0000
rain	úgua	ua	
red	kulu-kutʃúg	ʻula	
shoot	pænũ´	pana	
sun	la	la	
three	kél	kolu	
understand/ think	maŋ	manaʾo	
we (incl/excl)	aŋ háma	kami	

As seen above, the program has found 13 matching words (=the actual score) in Xokleng and Hawaiian. If the association between the two languages were random, one could expect something like 2.49 matches (=the random mean), so the difference between actually observed matches and those to be expected by chance (=the actual deviation) is large, viz., 10.51. The standard deviation is low, viz., 1.56, which is again an advantageous situation. Finally, the standard score is very large, viz., 6.74, recollecting that a standard score of 4.00 already yields a probability of 0.0000 up to the fourth decimal place. Thus, it is practically certain that the noted

similarities are non-chance, and therefore very significant.

One may notice that the above list of similar word pairs does not include e.g., the word pair for 'fear/be afraid', Xo *ŋaikaúg*^{n} and Haw *maka'u*, which was shown in Section 2 to be a good candidate for true cognacy. The reason for this exclusion (and actually the exclusion of a number of similar pairs in our 100-wordlist) is that the pair(s) do(es) not pass our criteria for similarity, implying that the similarity criteria may be defined too strictly and so exclude some reasonable matches. The converse is also true, and the criteria may be defined too loosely and so include some wild matches; so defining criteria is always a trade-off. The number of found matches, however, as mentioned earlier, is not the most essential part in looking for historical relationships, and what really matters is the strength of the proof for relationship, manifested in the significance level (and very impressive in this specific case).

3.2.2. Xokleng and Tagalog

Gloss	*Xokleng*	*Tagalog*	*Results*
come	katɛ́ŋ	datíŋ	Actual Score: 10
fat, grease	ta	tabáq	Random Mean: 3.85
five	pélemo	lima	Actual Deviation: 6.15
four	mpét	ápat	Standard Deviation: 1.91
leg, foot	pa	paqá	Standard Score: 3.23
red	kulu-kutʃúg	pula	Probability: 0.0006
stab/kill	patɨ	patáy	
tail	bu	buntót	
we (incl/excl)	aŋ háma	kami	
you (sg/pl)	a háma	kayoŋ	
		lahat	

The program found 10 matching word pairs, one of which is clearly spurious: by similarity criterion (iv) above, the sequence

-aha- of Xokleng's *a háma* 'you (sg/pl)' is matched with the identical sequence *-aha-* of Tagalog's *lahat*. But recollect the remark in the previous paragraph. The deviation in the number of actually observed matches (10) in comparison to the number that is most likely to occur by chance (3.85) is 6.15, a smaller figure than that in Hawaiian, but still large enough. The standard deviation is, again, low and the standard score sufficiently high, both favourable circumstances in unearthing historical relationships. Looking up in a statistical table of areas under the standard normal curve gives for this value of the standard score a probability (=significance) of 0.0006, i.e., a very significant result.

3.2.3. Xokleng and Malay

Gloss	*Xokleng*	*Malay*	*Results*
come	katéŋ	dating	Actual Score: 11
die	tɨ	mati	Random Mean: 3.92
ear	niŋná	telinga	Actual Deviation: 7.08
five	pélemo	lima	Standard Deviation: 1.85
four	mpét	empat	Standard Score: 3.82
skin	kut	kulit	Probability: 0.0001
small	kaitʃig^{n}	kecil	
we (incl/excl)	aŋ háma	kami	
woods	kuté	hutan	
yellow	kulu-klã	kuning	
you (sg/pl)	a háma	kamu sekalian	

Xokleng exhibits with Malay 11 matches (notice that though we keep the orthography of Malay as in *The Austronesian Basic Vocabulary Database*, the sounds are appropriately represented, so that e.g., Malay *ng* matches Xokleng *ŋ* in comparisons like Xokleng *katéŋ* ~ Malay *dating*).[4] The actual deviation is high and the standard

deviation low (both favourable circumstances), and the standard score is also high, viz., 3.83, leading to the high level of significance of 0.0001.

3.2.4. Xokleng and Fijian

Gloss	*Xokleng*	*Fijian*	*Results*
black	(kuro) loa	loaloa	Actual Score: 7
ear	niŋná	daliga-na	Random Mean: 2.98
five	pélemo	lima-na	Actual Deviation: 4.02
lips/mouth	ɲat-kusó	gusu-na	Standard Deviation: 1.60
man	kon-gang	tagane	Standard Score: 2.51
skin	kut	kuli-na	Probability: 0.0060
tail	bu	bui-na	

Xokleng shows 7 matches with Fijian in the inspected 100-item list. The standard score is lower in comparison with those in the previous tests, but still high enough to give a significance of 0.0060 (recall that social science usually considers a level of 0.01 highly significant).

3.2.5. Xokleng and Samoan

Gloss	*Xokleng*	*Samoan*	*Results*
die	tɨ	oti	Actual Score: 6
ear	niŋná	taliŋa	Random Mean: 2.18
five	pélemo	lima	Actual Deviation: 3.82
red	kulu-kutʃúg	ʔula	Standard Deviation: 1.40
sun	la	la	Standard Score: 2.72
thunder	totol	faaititili	Probability: 0.0033

[4] This, of course, applies to all Austronesian languages, not only Malayan. Xokleng, as described by Jules Henry, is represented appropriately in broad transcription.

Finally, Xokleng and Samoan share 6 matches, with standard score of 2.72 and probability of 0.0033, showing, though somewhat less strongly than in the cases with Hawaiian and Malay, a highly significant relationship.

4. Some Lexical Similarities between Xokleng and Other Austronesian Languages

If Xokleng bears statistically significant lexical similarities to the inspected five languages, which are widely distributed and belong to different branches of Austronesian, it would be expected that the Brazilian language would also show similarities in lexicon to other Austronesian languages. Below I give some examples of such resemblances in the style of mass comparison. As will be clear from the previous discussion, these examples will lack the strength of proof in the previous section, involving estimation of significance, but will nevertheless be suggestive (abbreviations: Xo=Xokleng, PAN=Proto-Austronesian).

1. THREE: Xo *kel*; PAN **telu*, Teor *tel*, Ainaro *tel*, Tami *tol*, Masela *wokkel*, Kisar *wo'kelu*, Lóvaia *okelo*.
2. FOUR: Xo *mpét*; PAN **Sepat*, N Sama *mpat*, Jama Mapun *mpat*, Bajau *mpat*, Lom *mpat*, Iban *mpat*, Kembayan *mpat*, Ribun *mpat*, Sanggau *mpat*.
3. FIVE/HAND: Xo (*pé)lemo*; PAN (*qa*)*lima*, Minangkabau *limo*, Tarangan *lém*, Lamaholot *léma*, Alor *lémma*, Kedang *lémé*, Tiang *patlima*.
4. EYE: Xo *kona, kuna*; Mengen *kana*, Tarpia *kani-*, Marquesan *konohi*, Rarotongan *kano`i*.
5. TOOTH: Xo *nona*; PAN **nipen*, Ngaibor (S.Aru) *nin*, Ujir (N.Aru) *néin ~ neyn*, Kambera *ŋandu*, Elat, Kei Besar *ni|no*,

Dehu *ñö*, Marshallese (E. Dialect) *ŋi*, Mengen *ŋiŋina*, Misima *nini|na*, Numfor *na*, Toba Batak *ŋiŋi*.

6. HEAD: Xo *klẽ*, *krẽ*; PAN **qulu*, Paiwan *qulu*, Watubela *kulu*, Maleu *kuri*.
7. HAIR: Xo (*klẽ*) *kula*; Proto Central Malayo Polynesian **qulu*, Proto Oceanic *(*nraun ni*) *qulu*, *(*daun ni*) *qulu*, Erai *kuru*, Marshallese (E. Dialect) *kooḷ*, Bwaidoga *kulua*, Kilivila *kulu*.
8. HAND: Xo *nɛŋga*; Maori *ringa(ringa)*, Hoava (New Georgia) *reŋgu*-, Dayak Ngaju *lenge'*.
9. BREAST: Xo *-kumbé* 'woman's breast'; Tuamotu *koouma*, Tahitian (18th Century) *'oouma*, Rurutuan *ʔoouma.*
10. NOSE: Xo *neyá*; PAN **mujiŋ*, Muna (Katobu-Tongkuno Dialect) *nee*, Wuna *nē*, Kilokaka (Ysabel) *nehu*-, Kia (Zabana) *nehu*-, Waropen *niha*.
11. EAR: Xo *niŋná*; PAN **Caliŋa*, Dehu *ineŋeñë*, *ñaŋeñyë*, Lio, Flores Tongah *hiŋa*, Ngadha *siŋa*, Geser *tiliŋa*, Watubela *telŋa*-, Maori, Rarotongan *taringa*, Kapingimarangi *taringa*, *dalinga*, Puluwatese *taninga*, Kwara'ae (Solomon Islands') *alinga*, Banoni *taŋina*-, Tunjung *neneŋ*, Melanau (Mukah) *liŋa*.
12. SKIN: Xo *kut*; Proto Malayo-Polynesian **kulit*, Canala *kɬ̃*, Lau *ʔuya*, Sengseng *ho*-, Yakan *kuit*, Rejang Rejang *kaʔ*.
13. TONGUE: Xo *numa*; PAN **Sema*, Letinese *nama*, Lenakel *nam*-, Tarpia *nama-k*, Mor *néma*.
14. RAIN: Xo *úgua*; PAN **quzaN*, Ngaibor (S.Aru) *goyan*, Lio, Flores Tongah *uja*, Marquesan *ua*, Mangareva, Rurutuan, Maori, Tahitian (18th Century), Tahitian (Modern), Rarotongan *ua*, Easter Island *uua*, *ua*, Uvea, East *ua*, Luangiua *ua*, Futuna, East *ua-ina*, Rennellese *ua*, Tunjung *ucan*, Dayak Ngaju *ujam*, Katingan *učam*, Bali, Sasak *ujan*.
15. FIRE: Xo *pẽ*; PAN **Sapuy*, Proto Oceanic **api*, Puyuma *ápui*, *apui*, *apoi*, Lio, Flores Tongah *api*, Kambera *api*, *epi*, Manggarai, Ngadha *api*, Lamaholot Ile Mandiri (Flores Timur) *ape*.

16. SUN: Xo *la*; Tongan *la'a*, Maori *ra*, Rapa Nui *ra*.
17. SKY: Xo *tae*; Peterara (Maewo) *tae laŋi*, Bontok (Guinaang) *dáya*; Geser *aŋin tái* 'cloud', Watubela *laŋit ni tái* 'cloud', Bobot *lak tain* 'cloud', Peterara (Maewo) *taelaŋi* 'cloud', Kasira (Irahutu), unclassified, *taye* 'cloud'.
18. CLOUD: Xo *gaikava*; Ujir (N.Aru) *kafkafal*, Wuvulu *u?ukafu*, Futuna, East *ko/kofu*.
19. TO SHOOT: Xo *pãnú*; PAN **panaq*, Manggarai *pana*, Buru (Namrole Bay) *pana*, Manam *pana*, Popalia *pana*.
20. TO STAB, TO PIERCE: Xo *pati*; Rurutuan *paatia*, Tahitian (18th Century) *paatia* [*tia*], Tahitian (Modern) *paatia*.
21. TO EAT: Xo *ko*; PAN **kaen*, Lio, Flores Tongah *ka*, Ngadha *ka*, Buru (Namrole Bay) *ka*.
22. MANURE/SHIT: Xo *kaé(-wé)*; Tagalog *tae*, Cebuano *tae*, Tongan *ta'e*.
23. GREEN/BLUE: Xo *taig*; PAN *(*ma*)*taq*, Manggarai *taʔa*, Kédang *taŋe*, Sika *daäŋ*.
24. YELLOW: Xo *klã*; Proto Malayo Polynesian **ma-kunij*, Savu *kalara*, Soboyo *kuni*, Nelemwa *kari*, Jawe *kari*, Tanna, Southwest *akwlha.*
25. YOU: Xo *a háma*, *ma*; PAN **i-kamu*, Wuvulu *ama*, Nelemwa *mo*, Nauru *ami*|*ai*, Lamogai (Mulakaino) *mu*, Mouk *umu*.
26. THOU: Xo *a háma*, *ma*; PAN **i-kaSu*, Toba Batak *hamu*, Tontemboan *kamu*, Old Javanese *kamu,* Ma'anyan *hañuʔ*.
27. I: Xo *eŋ háma*, *nũ*; PAN **i-aku*, Old Javanese *kami*, Toba Batak *ahu*, Favorlang *ina*, Bima *nahu*, Kambera *ñuŋga*, Luangiua *n*|*au*, Takuu *anau*, *nau*, Teanu *ene*, Nengone *inu*, Dehu *eni*, Nelemwa *na*, Mota *[i]nau*, Paamese (South) *inau*, Mono *ma*|*ha*, As *anɛ*.
28. WE (incl/excl): Xo *aŋ háma*; PAN **kami*, Ci' uli Atayal *cámi*, Ngaibor (S.Aru) *kama*, Ujir (N.Aru) *kama*, Cebuano *kami*, Old Javanese *kami*.
29. HE-SHE: Xo *ti* (*háma/ta), ði (háma/ði)*; PAN **si-ia*, Babatana

ɤœi, *sa*, Patpatar *ie, aie*, Popalia *ía, iːa*, Rejang Rejang *si*, Melayu Ambon *dia*.

30. WHAT: Xo *ne*; PAN **n-anu*, Bontok (Guinaang) *nə,* Ida'an *nu*, Yakan *ine*, Canala *anɛ̃,* Biga (Misool) *ane.*
31. AND: Xo *kũ*; PAN **ka*, Popalia *kɛ*, Bonerate *kɛnɛ*.

5. Discussion and Conclusion

The lexical similarities of Xokleng to the Austronesian languages Tagalog, Malay, Fijian, Samoan and Hawaiian were shown in Section 3 to be non-chance beyond any reasonable doubt. Also, as we saw in Section 4, the Brazilian language exhibits some suggestive lexical resemblances to other languages of the Austronesian family. These linguistic facts need to be explained.

One natural explanation, anticipated from the previous discussion, would be that Xokleng and these Austronesian languages are "historically related" (i.e., their relationship is either genetic or diffusional). On the surface, this might seem an implausible conjecture, potentially objectionable for the following two reasons. First, the Brazilian language, as far as current wisdom goes, is not an isolate language, but is classified in the Kaingang subfamily (with the language Kaingang), which in turn is classified as a part of the Ge-Kaingang family, itself a part of the Macro-Ge stock. Secondly, Xokleng and the Austronesian languages seem much too geographically distant to be historically related (in any of the senses above).

In response to the second potential objection, the great geographical distance between the examined languages, we may recall that Thor Heyerdahl (1950) has experimentally demonstrated with his primitive raft Kon-Tiki that see voyages *can* be achieved between South America and Polyneisa. (Indeed, he suggested that Polynesia is colonized from a people of Pre-Inca Peru, an idea that

has not met general approval, but is still not excluded from consideration by authoritative genetic investigations such as Cavalli-Sforza et al. 1994).

Regarding the first potential objection, viz., that Xokleng is not a language isolate, we should say that our suggestion of a possible connection with Austronesian does not necessarily contradict its present genealogical classification. Thus, Xokleng can very well be a genuine Macro-Ge language and at the same time be related to Austronesian, only a logical consequence of this situation would be that all languages validly classified as Macro-Ge would also be related to Austronesian. However, insomuch as I do not at present have sufficient data, I refrain from making such a strong supposition, limiting it only to Xokleng. At the same time, Xokleng is known to be very strongly related to Kaingang and less strongly, but still noticeably, to some other Ge/Macro-Ge languages, so it can be noted that even a version of the stronger supposition may not look out of place. Here, I will confine myself to giving only some lexical similarities between the inspected five Austronesian languages and the Macro-Ge Xokleng, Kaingang, and Maxakali (a Macro-Ge language spoken in Minas Gerais and Espirito Santo). Table 2, based on data for Kaingang and Maxakali from Wiesemann (2002) and Popovich & Popovich (2005) respectively, summarizes such similarities from our 100-item wordlist. It is important to emphasize that, while I have not systematically compared Maxakali, the pair-wise comparisons of Kaingang to the five Austronesian languages, which cannot be discussed here, all turned out to be highly statistically significant, similarly to the comparisons with Xokleng.

Below I return to the question of the possible historical relationship of Xokleng to Austronesian, providing some further support for this conjecture.

First, South America is both linguistically and genetically the most diverse part of the world. Now, if South America was exclusively colonized from the north, as it is predominantly believed

today, then it should be expected that both South American languages and South American native populations would, in a significant way, resemble languages and populations of North America.

Table 2. Lexical Similarities between Some Macro-Ge Languages and Austronesian Languages

(Abbreviations: X=Xokleng, K=Kaingang, Mx=Maxakali; P=Proto-Austronesian (*), T=Tagalog, M=Malay, F=Fijian, S=Samoan, H=Hawaiian)

No.	Gloss	Macro-Ge languages	Austronesian languages
1	come	singular: *katéŋ* X, *kãtĩg* K, *mã* Mx plural: *kãmũ*, *mũ* X, *kãmũ* K	*datíŋ* T, *dating* M *lakomai* F, *o mai* S, *mai* H
2	ear	*niŋná* X, *nĩgrẽg* K	*teliŋa* M, *daliŋa-na* F, *taliŋa* S
3	fear	*ɪɲai-kaúg*^{n} X, *kamég* K, *katuk* Mx	**ma-takut* P, *ma-ka'u* H, *ma-taʔu* S, *tákut* T, *takut* M
4	five	*pélemo* X	*lima* T, M, S, H, **(qa)lima* P
5	four	*mpét* X	*empat* M, *apat* T, **Sepat* P
6	kill	*patɨ* 'stab' X, *putui* Mx	*patáy* T, **patay* P
7	leg/foot	*pa* X, *pẽn* K, *pata* Mx *kaka* 'at foot of' Mx	*paqa* T **qaqay P, kaki* M
8	shoot	*pænũ´* X, *pénũ* K, *mãn* Mx	*pana* H, **panaq* P, *vana* F, *fana* S
9	sun	*la* X, *ra* K	*la* H, S, *araw* T
10	three	*kél* X *tãgtũ* K	*kolu* H *tiga* M, *tatlo* T

However, South American languages are well known to be very different from North American ones. The same applies to population genetics. Regarding the Macro-Ge people in particular, it was found, in drawing a phylogenetic tree of 23 American tribes, grouped according to linguistic criteria (Cavalli-Sforza et al. 1994: 323-4), that they are the worst outliers. It can therefore be concluded that linguistic and genetic connections with other parts of the world cannot be a priori excluded.

Secondly, some recent genetic investigations give direct evidence of a possible link between Brazilian native populations and populations in the Pacific. Ribeiro et al. (2003: 59), analyzing the Macro-Ge-speaking Xikrin and the Tupi-speaking Parakanã (note that Tupi is believed to be related to Macro-Ge), found them to be genetically similar to Indonesians and South-East Asian populations, concluding that "These results corroborate the existence of genetic affinities between Brazilian Indians and South-east Asian and Oceanic populations", their investigation being intended to "further contribute to the theory of a predominantly Asiatic origin of the American natives". Ribeiro et al. (2003) cite other genetic work to the same effect.

And last but not least, our own previous linguistic investigations also link Xokleng to Austronesian. Suffice it to only summarize some of our basic results here. We studied computationally (with a version of the program also used here) the kinship semantic patterns of 566 societies, based on the data set contributed by Murdock (1970). The Murdock data set focuses on eight sets of kin: grandparents, grandchildren, uncles, aunts, nephews and nieces (male speaker), siblings, cross-cousins, and siblings-in-law. Every type of kin is described in terms of "kin term patterns", showing the number of kin terms used for that kin as well as their range of reference (e.g., for the kin "grandparents" we may have a *Bisexual Pattern* (has two terms, distinguished by sex, which can be glossed as "grandfather" and "grandmother"), *Merging Pattern* (has a single

undifferentiated term, which can be glossed as "grandparent"). (Murdock gives overall 20 patterns for this kin.). The computer program was used to discover statistically significant similarities in patterns between languages, belonging to *different* language families according to the classification of *Ethnologue*. It was found that Xokleng bears very strong relationship to each of the Austronesian languages Amis, Ulithian and Trukese (in fact the most statistically significant results in comparison to all investigated language pairs). Malay and Samoan, studied here, also showed substantial similarities in kinship semantics with Xokleng. This result in fact initiated further investigations, some of which are reported here. Some other of our findings were that Xokleng is very significantly statistically tied to the original Proto-Austronesian language in regard to its basic vocabulary and some sound correspondences were suggested to account for the similarities. Also, further evidence was adduced as to the resemblances of Xokleng to Austronesian, pertaining to both their phonological and grammatical structure, the latter being quite suggestive of genetic rather than diffusional relationship.

Summarizing, we may say that the conjecture of a possible historical (most probably genetic) relationship between Xokleng and languages like Tagalog, Malay, Fijian, Samoan and Hawaiian (and Austronesian more generally) is far from being so implausible as it might look at a first glance. Quite the contrary, it becomes quite reasonable in the context of all that was said in the previous paragraphs. However, at present, it can be viewed as no more than a hypothesis, which requires further and much more detailed historical linguistic test, as well as tests from the other related fields like genetics, to be corroborated or, alternatively, falsified by proposing other reason(s) to explain the diverse non-chance similarities that have been found to exist among these languages. A basic goal of this paper is to invite such tests.[5]

References

Bender, M. 1969. Chance CVC Correspondences in Unrelated Languages. *Language* 45, 519-531.

Cavalli-Sforza, L., A. Menozzi, & A. Piazza. 1994. *The History and Geography of Human Genes*. Princeton, NJ: Princeton University Press.

Evans, B. & M. Ross. 2001. The History of Proto-Oceanic *ma-. *Oceanic Linguistics* 40, 269-290.

Gensch, H. 1908. Wörterverzeichnis der Bugres von Santa Catharina. *Zeitschrift für Ethnologie* 40, 744-759.

Good, P. 1994. *Permutation Tests*. New York: Springer Verlag.

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

Greenberg, J. 1957. *Essays in Linguistics*. Chicago, IL: University of Chicago Press.

Henry, J. 1935. A Kaingang Text. *International Journal of American Linguistics* 8, 172-218.

_____. 1948. The Kaingang Language. *International Journal of American Linguistics* 14, 194-204.

_____. 1941. *Jungle People: A Kaingang Tribe of the Highlands of Brazil*. New York: Vintage Books.

Heyerdahl, T. 1950. *The Kon-Tiki Expedition*. London: Allen and Unwin.

Kessler, B. 2001. *The Significance of Word Lists*. Stanford, CA: CSLI Publications.

Murdock, G. 1970. Kin Term Patterns and Their Distribution. *Ethnology* 9, 165-207.

Oswalt, R. 1970. The Detection of Remote Linguistic Relationships. *Computer Studies in the Humanities and Verbal Behavior* 3, 117-129.

_____. 1991. A Method for Assessing Distant Linguistic Relationships.

[5] At the moment this paper goes to print, I dispose of further results supporting eventual affinity of the Kaingang family with Austronesian. More than 50 quite convincing cognate sets, with sound correspondences, have been worked out between Xokleng and Kaingang and Polynesian languages.

In S. Lamb & E. Mitchell (eds.), *Sprung from Some Common Source: Investigations into the Prehistory of Languages* 389-404. Stanford, CA: Stanford University Press.

Popovich, H. & F. Popovich. 2005. *Maxakali-English Dictionary*. Cuiabá, MT: Sociedade Internacional de Lingüística.

Ribeiro, D., M. Figueiredo, F. Costa, & M. Sonati. 2003. Haplotypes of α-globin Gene Regulatory Element in Two Brazilian Native Populations. *American Journal of Physical Anthropology* 121, 58-62.

Rodrigues, A. 1999. Macro-Jê. In R. Dixon & A. Aikhenvald (eds.), *The Amazonian languages* 164-206. Cambridge: Cambridge University Press.

Ruhlen, M. 1987. *A Guide to the World's Languages: Classification*. Stanford, CA: Stanford University Press.

Urban, G. 1985. Ergativity and Accusativity in Shokleng (Gê). *International Journal of American Linguistics* 51, 164-87.

Valdés-Pérez, R. & V. Pericliev. 1999. Computer Enumeration of Significant Universals of Kinship Terminology. *Cross-Cultural Research* 33, 162-174.

Wiesemann, U. 1986. The Pronoun Systems of Some Je and Macro-Je Languages. In U. Wiesemann (ed.), *Pronominal Systems* 359-380, Tübingen: Gunter Narr.

_____. 2002. *Dicionário Bilingüe Kaingang-Português*. Curitiba: Editora Evangélica Esperança.