The higher phylogeny of Austronesian and the position of Tai-Kadai

Laurent Sagart

THE HIGHER PHYLOGENY OF AUSTRONESIAN AND THE POSITION OF TAI-KADAI[1]

Laurent Sagart
CNRS, Paris

[1] This is a modified version of a paper presented at the workshop on "Les premiers austronésiens: langues, gènes, systèmes de parenté", Paris, May 5, 2004. Thanks go to Sander Adelaar, Peter Bellwood, Bob Blust, Isabelle Bril, Alexandre François, Jeff Marck, Estella Poloni, Lawrence Reid, Malcolm Ross, Alicia Sanchez-Mazas and John Wolff for useful discussion.

Abstract

This paper presents a new higher phylogeny for the Austronesian family, based on three independent lines of evidence: the observation of a hierarchy of implications among the numerals from 5 to 10 in the languages of Formosa and in PMP; the finding that the numerals *pitu '7', *walu '8' and *Siwa '9' can be derived from longer additive expressions meaning 5+2, 5+3 and 5+4, preserved in Pazeh, using only six sound changes; and the observation that the phylogeny which can be extracted from these and other innovations -mostly changes in the basic vocabulary- evinces a coherent spatial pattern, whereby an initial Austronesian settlement in NW Taiwan expanded unidirectionally counterclockwise along the coastal plain, circling the island in a millennium or so. In the proposed phylogeny, Malayo-Polynesian is a branch of Muic, a taxon which also includes NE Formosan (Kavalan plus Ketagalan). The ancestor language: Muish, is deemed to have been spoken in or near NE Formosan. Further evidence that the The Tai-Kadai languages, contrary to common sense, are a subgroup of Austronesian (specifically: a branch of Muic, coordinate with PMP and NE Formosan) is presented.

This paper presents a new higher phylogeny of Austronesian based on strictly cladistic principles: each node will be supported by linguistic innovations. This is not a new approach: the phylogeny in Starosta (1995) was based in morphological innovations, and those in Blust (1999) and Ho (1998) were based primarily or entirely on phonological innovations, mostly mergers. Here I will use innovations drawn almost exclusively from changes in the basic vocabulary. Using this methodology I will construct a tree-like phylogeny for the higher (non-MP) part of Austronesian phylogeny, and give further evidence for the claim, made earlier (Sagart 2001; in press, a) that the Tai-Kadai languages are a subgroup of Austronesian. Before proceeding with the main issue, I need to make some methodological remarks.

1. Methodological remarks

Phonological mergers are convenient features in subgrouping, because one can be sure that they are innovations. In that respect they fit the basic neo-grammarian requirement that subgrouping be effected on the ground of shared innovations rather than shared retentions. At the same time, phonological mergers are consequences of regular sound changes; regular sound changes in turn are known to spread along social networks which routinely cross-cut dialect boundaries, and even, through bilingual speakers, language boundaries: witness the spread of Parisian /r/ in parts of 17th- and 18th century Europe (Trudgill 1974:162); in Taiwan, the merger of the PAn phonemes *d and *z which has affected Basai and Kavalan but not Trobiawan, Rukai but not Tsouic, and all other Formosan languages save Taokas, Siraya and Favorlang: that collection of languages is not a taxon by anyone's subgrouping: spreading by contact must have played a major role. A long list of sound changes having spread across language boundaries, sometimes over very large expanses of land, such as the spread of tonal contrasts in East Asia, could be presented.

The propensity to spread over dialect or language boundaries is not a curious idiosyncrasy of certain sound changes: it is the way regular sound change habitually works. That is why phonological isoglosses *normally* overlap in dialect maps, and why phylogenies constructed from phonological mergers tend to be ambiguous and inconclusive, except of course in situations where spreading in made impossible by the geography, as in the eastern Pacific island world. In such regions, phonological mergers are useful evidence in

constructing linguistic phylogenies.[2] In the case of language taxa having evolved on relatively extensive land masses, like Taiwan, it is preferable to construct phylogenies on those types of innovative characters that are least likely to be transferred through contact. The most useful and readily available are morphological changes and especially lexical replacements in notions belonging to the core vocabulary: personal pronouns, numerals, body part terms and the like.[3]

The main difficulty with morphological and lexical changes is how to be sure that one is dealing with innovations. For morphology, Starosta (1995) used the principle that innovation can be equated with complexification, but this does not appear to be a reliable principle, as it appears that morphological processes can be lost without leaving any lexicalized traces. With the lexicon, it is sometimes claimed that lexical innovations cannot be identified prior to subgrouping. This, if true, would render them useless in subgrouping. Yet principles allowing the identification of lexical innovations prior to, rather than as a result of, subgrouping, exist. Here are examples of situations where this is possible:

- When a word can be shown to be phonologically reduced from an expression consisting of several words in a related language, it is a natural inference that the reduced form is an innovation.
- It is often the case that when two etyma compete for a certain meaning, the more etymologically transparent of the two is the innovation.
- when a root occurs with two meanings in different languages, and one of the meanings is clearly the older one. See the discussion of 'moon' in section 4.1.
- In the case of closed lexical systems, it is a fair bet that an analogically leveled form is innovative and that its non-leveled counterpart is a preservation.

[2] I am grateful to Malcolm Ross for this important caveat.

[3] I am *not* claiming that lexical innovations in the basic vocabulary do not spread at all: I am only claiming that when we select lexical innovations from the basic part of the vocabulary, we minimize the risk of selecting characters that have spread by contact, without eliminating that risk.

- At times linguistic geography allows us to tell which of two competing forms for a certain meaning is old, and which is innovative (see fn. 14).

In what follows I will make use of these principles to investigate the higher phylogeny of Austronesian.

2. The distribution of the numerals 5-10

Nearly consensual reconstructions for the PAn numerals are *isa or *esa '1', *duSa '2', *telu '3', *Sepat '4', *lima '5', *enem '6', *pitu '7', *walu '8', *Siwa '9', *puluq '10'. An interesting situation can be observed with the numerals from '5' to '10' (Table 1): throughout Taiwan, a reflex of *puluq '10' implies the presence of a reflex of *Siwa '9',[4] which implies the presence of *walu '8', which implies the presence of *enem '6', which implies the presence of *lima '5', which implies the presence of *pitu '7', while the reverse implications do not hold. PMP has reflexes of all numerals from 5 to 10, in conformity with the Formosan implicational hierarchy. In diagrammatic form:

puluq >> Siwa >> walu>> enem >> lima >> pitu

where '>>' means 'implies the presence of'. This is shown in Table 1.

<Table 1 >

The numerals 5-10 can be used as lexical criteria (or 'characters') to classify Austronesian languages. The data in Table 1 show that they are are mutually compatible in the sense of Meacham and Estabrook (1985): that is, they are all compatible with the same phylogenetic tree. A natural explanation for a distribution of the kind shown in Table 1 is that we are dealing with a sequence of nested innovations: specifically, that PAn had neither of these numerals, and that they arose one after the other, in succession, with *pitu arising first, then *lima, in a language that already had *pitu; then *enem, in a language that already had *pitu and *lima; and so forth.

[4] In some varieties of Rukai, such as Oponohu (as cited in Ferrell 1969), there is a reflex of *puluq but none of *Siwa: presumably a reflex of *Siwa was displaced by the Rukai-speciic form vaŋatə 'nine'.

The alternative is to suppose that the PAn numerals for 5-10 were lost in various Formosan languages, especially on the West coast and in the center, in such a way that the languages which lost *pitu '7' were a subset of those which lost *lima '5'; that those which lost *lima were a subset of those which lost *enem 'six': and so forth. It is difficult to think of a reason why this would happen. The odds of it occurring by accident do not seem high either.

The idea that the numerals under discussion are not PAn is certainly paradoxical, considering that the consensus opinion favors the opposite view. Yet one should keep in mind that assignment of *lima, *enem, *pitu, *walu, *Siwa and *puluq to the PAn level crucially relies on the fact that all are reflected in PMP, combined with the assumption that MP is a primary branch of An (or even two primary branches, in Dyen's view). If however, as argued by Harvey (1979, 1982), Reid (1982), Starosta (1985, 1994, 1995, 1996, 2001), Ross (1995), Benedict (1995), Bellwood (1997), Ho (1998), Ho and Yang (1999) and Sagart (2002), the MP languages are part of a taxon which also includes languages of the Formosan east coast, then the testimony of PMP counts for little, and assignment to the PAn level should be decided primarily on an etymon's distribution in the languages of Taiwan. If so, there is no particular reason why the six numerals under consideration should be PAn words.

Assuming that the consensus forms of the numerals 5-10 are post-PAn innovations immediately raises the question of their etymology. There is a chance that cues to the origin of these words can still be found in the numeral systems of the languages where the innovative forms are not reflected.

3. New etymologies for *pitu*, *walu* and *Siwa.*

What forms of the numerals 5-10 do we find in the Formosan languages that do not have the familiar forms ? Most common above '5' are analytic forms. In Pazeh we find additive forms (discussed in detail below): 6=5+1, 7=5+2, 8=5+3, 9=5+4; additive forms also in Saitaoyak and Tarumyan, two varieties of Saisiat recorded by Ino Yoshinori (Ino 1998), the word for '7': *saivuseaha* is made up of *saivusa* '6' and *aha* '1'. Multiplicative forms of the kind of 2x3, 2x4 are common for '6' and '8': Sediq *materu*, Thao *katuru* are based on *telu '3' and Sediq *maspat*, Thao *kashpat* '8' are based on Sepat '4'. In addition to

katuru and *kashpat*, Thao has longer forms *makalh-turu-turu* '6' and *maka(lh)-shpa-shpat* '8'. In Saisiat too the form for '8' is based on '4': Tarumyan *kaspat*, Saitaoyak *makaspat*. Favorlang *maaspat* < *ma[k]aSpat '8' agrees with Saitaoyak. Taokas *ma-hal-pat*, Siraya *kuixpa* similarly appear to be based on '4'.[5] Various languages have subtractive forms for '9': Sediq *maŋali* 'nine', imperative of *maŋal* 'take', apparently through 'take [one out of ten]', as noted by Pecoraro (1977); perhaps also Saisiat-Saitaoyak *ra:ha* 'nine' (Ino 1998) which contains *aha* 'one'. An interesting set consists of Thao *tanacu*, Favorlang *tannacho*, Taokas *tanaso* '9', which point to an earlier *[st]a[nŋ]aCu. The first syllable might reflect *sa- 'one', in which case we are perhaps dealing with a subtractive form.[6]

In Taiwan, the prevalence of analytic forms for numerals above five is essentially a west coast phenomenon. This is striking, because the west coast faces the continent and can be suspected of being the area of the earliest An settlement and the place where the first An diversification took place.

It is not in principle impossible that analytic forms could have arisen secondarily on the west coast, displacing earlier short forms. The development of analytic forms based on 'five', displacing older reflexes of the numerals 5-10, occurs in various languages of the Philippines, New Caledonia and north Vanuatu, such as Ilongot (Reid 1971), Nêlêmwa (Bril 2002: 381-82) and Mwotlap (François 2001: 344), for instance.

Another interpretation of the facts is possible: PAn had a numeration system with stable words for numerals up to '5', and no stable words for '6', '7', '8',

5 The alternation in these multiplicative forms of ma- and ka- prefixed forms is interesting. Where both occur (Thao, Saisiat-Saitaoyak, Favorlang), it is always in the order ma+ka. The final consonant in Thao makalh- unambiguously reflects *R: this is perhaps the same formative as in Rukai ma- < *maR- 'dual' (Li 1975: 14, 74, 261), south Paiwan mag- 'dual, plural' (Elizabeth Zeitoun, p.c., 2002). In Thao (Blust 2003:113, 115) prefix makalh- is attested only with numerals (including makalh-tanacu a maqcin '90') but maka- derives verbs meaning 'to resemble X, produce X, from X, in X, to X' out of nouns. It is possible that the PAn prototypes of the multiplicative forms for '6' and '8' are retained in the long Thao forms makalh-turu-turu '6' and maka(lh)-shpa-shpat '8'. If so, the corresponding PAn forms would be *makaR-telu-telu '6' and *makaR-Sepat-Sepat '8'. They would be verbal in origin, perhaps something like 'to be from N (maka-) doubled (-R)'.

6 This form could contain a verb of taking, perhaps the same etymon as in Pazeh *asu* 'bring' < *aCu. The intervening nasal could be the segment that sometimes attaches to the end of the numeral 'one', as in Pazeh *adang* '1', Bunun (ishbukun: Imbault-Huart 1893:256) *tashang* '1': the prototype would then be *sa-ŋ-aCu 'one brought (towards ten)'.

'9'. Expressions for the corresponding notions were made up on the spot using additive, multiplicative and subtractive strategies. The analytic forms in the west coast languages are their fossilized descendants. The familiar disyllabic forms of the numerals arose one after the other in successive daughter languages of PAn, gradually displacing all the old PAn analytic expressions.

I will argue that the latter explanation is true, by proposing new etymologies for '7', '8' and '9'. I will take as my starting point the Pazeh numerals for 6–9, from Li and Tsuchida (2001):

xaseb-uza '6'
xaseb-i-dusa '7'
xaseb-a-turu, *xaseb-i-turu* '8'
xaseb-i-supat '9'

The Pazeh word xasep means 'five' and the forms from 6 to 9 are additive: 5+1, 5+2, 5+3, 5+4. Li and Tsuchida (2001) give two alternating forms for 'eight': one with linker -a- and another with linker -i-. They give *xasep-a-turu* a separate entry and list *xaseb-i-turu* under *xasep* 'five'. In the Pazeh texts edited by the same authors, only *xaseb-a-turu* occurs (Li and Tsuchida 2002: 49, 53). I regard *xaseb-a-turu* as the primary spoken form and *xaseb-i-turu* as its analogically levelled variant. This will explain why the -i- form tends to appear when numerals are elicited as part of a word list (Ferrell 1969; Lin 2000:159). Only 'eight' shows this variation. The numerals for '7' and '9' attest only -i-. This idiosyncratic fact will prove significant.

Change of final -p to -b in *xasep* is regular (Blust 1999:326, rule I).[7] Pazeh *xasep* has cognates in other West coast languages: Favorlang *achab* (drawn from Ferrell 1969), Saisiat *a:seb* (Yeh 2000); these two reflect *RaCep, if we suppose that, as in Pazeh, final voicing is secondary in Favorlang and Saisiat. Taokas *hasap* (drawn from Ferrell 1960) appears to reflect *qaCep. The numerals added to *RaCep in the Pazeh words for 6, 7, 8, 9 are the PAn words for 1 (*esa), 2 (*duSa), 3 (*telu), 4 (*Sepat), all with regular developments.

[7] "A rule of intervocalic voicing that affects voiceless stops before a morpheme boundary but not within a morpheme".

The full additive forms found in Pazeh are not observed elsewhere, but shorter forms in some west coast languages are indicative of their former existence in other dialects/branches of An:

- The final three syllables of Pazeh *xase**baturu***, the additive form for '8', are paralleled in Luilang '8' ***patulu****-nai* (where *-nai* is detachable, compare '7' *in-nai* and '9' *satulu-nai*, with *sa-* = '1' in Luilang).
- One dialect of Siraya (Tsuchida, Yamada and Moriguchi 1991, point M2) has *sipat* 'nine': [8] given the phonetic proximity with PAn *Sepat 'four', this is almost certainly based on an additive 5+4 form.

Let us now return to the Pazeh paradigm, leaving the word for '6' aside:

xasebidusa '7'
xasebaturu '8'
xasebisupat '9'

Under the standard explanation, any resemblances between the analytic Pazeh expressions for '7', '8', '9' and the familiar words *pitu, *walu and *Siwa must be fortuitous. My suspicion that the standard explanation does not account for the facts was raised by the observation that the Pazeh word for '7' contains the sequence *-bidu-*, close to *pitu, the familiar word for '7':

x	a	s	e	b	i	d	u	s	a
				p	i	t	u		

That -b- in the Pazeh form is phonologically a /p/ makes the resemblance more specific.

[8] This form *sipat* '9' was obtained by Ogawa from a nonnative policeman in Kalapo. Siwa-type forms for '9' are not common in Siraya. From another policeman in the same locality, Ogawa recorded *ra-siwa* '9' (dialect M3 in the same book). Probably the -p- was in the process of being lenited and final -t was weak.

Looking now at the Pazeh word for '8', we notice that it contains consonants of the same point of articulation, and the same vowels, as *walu '8', all in the correct sequential order, even though a syllable intervenes between them:

x	a	s	e	b	a	t	u	r	u
				w	a			l	u

The -r- in Pazeh *turu* 'three' reflects PAn -l-, which improves the goodness of fit with *walu. Note also that the intervening syllable -tu- had schwa (orthographically 'e') in proto-Austronesian: PAn *telu '3'. Suppose that schwa fell, a -tl- cluster would result, which could then easily simplify to -l-, given the hostility of An phonology to consonant clusters.

With '9' we also find in the Pazeh words several of the phonetic ingredients that enter in the familiar reconstruction *Siwa, again in the right sequential order:

x	a	s	e	b	i	s	u	p	a	t
						S	i	w	a	

Pazeh s- reflects proto-Austronesian S-, which again improves the comparison. Moreover, the correspondence between Pazeh /pa/ and 'consensus PAn' /wa/ is the same as in '8' (remember that in the Pazeh word for 'eight' xasebaturu, 'b' is phonologically a p). Least satisfactory is the match between Pazeh u and PAn i in the first syllable, but note that Pazeh u reflects PAn schwa in this case (PAn *Sepat 'four'), and that of all PAn vowels, schwa is the most sensitive to contextual influence.

These facts suggest that the 'consensus PAn' forms for '7', '8' and '9' arose as reduced forms of earlier additive expressions. I will now describe a simple evolution path leading from the latter to the former. Let us first go back to the list of Pazeh numerals and reconstruct the PAn forms that they would derive from, based on the accepted sound correspondences between Pazeh and PAn (Blust 1999; Li and Tsuchida 2001), supposing that the additive forms existed in PAn:

*RaCep-i-duSa '7'
*RaCep-a-telu '8'
*RaCep-i-Sepat '9'

I first introduce two arbitrary modifications to this paradigm:

1. I change RaCep-i-duSa to RaCep-i-tuSa. There is independent evidence for an old variant *tuSa of '2' (Amis *tusa*, Puyuma *towa*, Thao *tusha*) with initial *t- instead of *d- perhaps on the analogy of *telu '3'. The arbitrariness is in stipulating that the *tuSa variant was used in the PAn form which underlies *pitu '7'.
2. I arbitrarily[9] assign stress to the penultimate syllable of '7' and to the final syllable in each of '8' and '9'.

In the resulting paradigm I write stressed vowels in bold type:

*RaCep-i-t**u**Sa '7'
*RaCep-a-tel**u** '8'
*RaCep-i-Sep**a**t '9'

Now I submit this paradigm to six sound changes (Table 2).

<Table 2>

There are several possible variants of this derivation: in particular, changes 4, 5 and 6 are not crucially ordered with respect to one another.

Although they are arbitrary, the changes supposed here

[9] I assume PAn generally had final stress. A reason why '7' could have its penultimate vowel stressed would be if -a in *tuSa were the ligature, which became attached to the original word for '2', which ended in -S. This in turn could explain why in Pazeh /a/ is optional (indeed, mostly absent) between *dusa* '2' and a following noun: *dusa daali* '2 days', *dusa rakihan* 'two children', *dusa saw* 'two persons', *dusa ilas* 'two months', *dusa isit* 'twenty' (two tens); compare *adang a daali* 'one day', *turu a rakihan* 'three children', *supad a saw* 'four people', *supaz a isit* 'forty', *xaseb a saw* 'five people', *isid a ilas* 'ten months', etc. This is apparently not because of a constraint on sequences of like vowels, as we find noun phrases like *tula a daran* 'path of an eel' (Li and Tsuchida 2002: 82).

- are natural changes: assimilations, cluster simplifications, schwa deletions, lenitions, stress-conditioned prunings, not outrageous changes like p > r, or l > m, or i > q;
- affect at least two forms (except for changes 1 and 6): two changes (# 3, 4) affect the entire paradigm (three forms). By definition, *ad hoc* changes would affect only one form. The relatively marked lenition -pa- > -wa- affects two forms;
- do not change the vowels, except for schwas: this is a general tendency of later An phonetic evolution;
- do not affect the points of articulation of the consonants.

The changes in Table 2 did not apply to the entire vocabulary. There is ample evidence that, for instance, PAn /pa/ normally remains /pa/ in PMP, and that the schwas of PAn are usually retained in PMP. I am arguing that these changes took place when expressions of four or more syllables were reduced to disyllables as a result of the 'drive to disyllabism' which has been at work throughout early An history. Such reduction could have taken place when long forms came to be treated prosodically as prosodic feet, rhythmically equivalent to canonical disyllabic feet. Compression of the phonic material into the narrow temporal confines of a rhythmic foot would have provided the basis for the lenitions, schwa-deletions and procrustean prunings in Table 2.

I am well aware that the sound changes I am hypothesizing lack the support of recurring sound correspondences. That is unavoidable, since the drive to disyllabism could only have affected a small number of expressions simultaneously: even supposing that the changes happened before our very eyes, the number of examples for each of the changes involved would be too small for sound correspondences to be established anyway. What Table 2 establishes is that phonetic evolution from the long to the short forms is possible and that it only requires the application of a small number of natural sound changes.

At the same time, lack of support from sound correspondences leads to the question as to whether the resemblance between long and short forms is not accidental, especially considering that I have made use of two *ad hoc* stipulations and six arbitrary sound changes. I have not conducted a proper

probabilistic study of this question,[10] but I believe that the resemblances are meaningful, for the following reasons.

First, there exists independent evidence for change #4 Prune left:

- Thompson (1873) records some words from a now-extinct variety of Pazeh. The numerals from 6 to 9, cited here as reproduced by Imbault-Huart (1893:319) are *boudah* '6', *bidousut* '7', *bitouro* '8', *bissoupat* '9': leaving aside the curious word for '7', these are clearly left-pruned outcomes of the long forms in Li and Tsuchida (2001).
- Luilang *patulu-nai* '8' (Luilang *-nai* is detachable, as already noted) appears to be the left-pruned outcome of *RaCep-a-telu.
- Siraya-Makatao *sipat* '9' (point M2 in Tsuchida, Yamada and Moriguchi 1991), which contains *Sepat 'four' in a slightly altered form must be the left-pruned outcome of an earlier additive expression, although in this case there is no evidence that the pruned-off part was *RaCep.

Second, the irregularity in Amis *falu* '8', earlier *balu, instead of expected walu, can now be explained: the lenition of -pa- in *RaCepatelu went through βa-, which was reinterpreted word-initially as ba- in pre-Amis. At the same time, -pa- in '9' was fully lenified to -wa (Amis *siwa* '9'), perhaps because it was in intervocalic position, a facilitating context for a lenition.

Third, the present interpretation explains an interesting detail in the distribution of forms across Formosan languages: no language has at the same time a reflex of *pitu and a transparent additive from based on *RaCep '5' for another numeral. This is because a language that has a reflex of *pitu is at least a stage-5 language in terms of Table 2: at stage 5, additive forms based on *RaCep have already been reduced to disyllables.

Above all, the etymologies proposed here relate the familiar *pitu, *walu, *Siwa paradigm to another paradigm: the Pazeh additive forms for '7, '8' and

---

[10] I have counted the number of changes required to convert the same three putative PAn analytic forms into the same three familiar numerals, only *paired differently* (7>8, 8>9, 9>7): I have found that at least 10 changes, some strange or unmotivated, are required, most applying to only one form.

'9'; rather than to forms unrelated to each other, reflected in different languages.

The alternative requires one to accept that:

- the resemblances between the long and short forms for 7, 8 and 9 are accidental;
- /f-/ in Amis *falu* '8' is irregular;
- the fact that An languages reflecting *pitu can have multiplicative and subtractive forms, but not additive forms based on *RaCep, is accidental;
- the fact that the three etymologies proposed for '7', '8' and '9' form a paradigm in Pazeh is accidental.

Table 1 shows that (except for some varieties of Rukai, see fn. 4) *walu and Siwa occur in the same languages. Before we examine the question in more detail, let us first pause to consider what the PAn numeral system may have been. The numerals for '1' to '4' were *esa, *duSa~tuSa, *telu, *Sepat; '5' was *RaCep (the 'standard PAn' form *lima is transparently from 'hand', while *RaCep is opaque; only *RaCep occurs in the additive forms: so *lima must have displaced *RaCep in early post-PAn times; see below). There is the possibility that final -a in *esa and *duSa~tuSa is a captured ligature. There was also a word for '10': Pazeh *isit*, Luilang *isit*, Favorlang *tsxiet*, Taokas *(ta)isid*, Papora *metsi*, Hoanya *(miata)isi*. An approximation of the PAn form is #(sa)-iCit, where sa- = '1'. That word was later displaced by *puluq, as we will see. Between '5' and '10' there were no stable numerals. Table 3 describes the analytic forms found in Formosan languages for numerals 6-9.

<Table 3>

Additive and multiplicative strategies generally appear to have been favored over subtractive ones for '6' to '8', but subtractive formations are more common for '9': all numbers were open to additive formation, and both even numbers to multiplicative formation. With additive expressions, only those involving the highest lexicalized numeral (*RaCep 'five') were in use (to the exception of Saisiat *saivuseaha*, from saivusa '6' plus *aha* '1', where the origin of *saivusa* is problematic).

Let us now return to the question of why *walu and *Siwa appear only in Formosan languages that have *pitu. The shift from the PAn numeration system to the post-PAn system was effected, I have proposed, through the reduction to disyllables of full PAn additive forms, as shown in Table 2: but while the old additive forms disappeared even as they were being phonetically reduced, this did not eliminate multiplicative and subtractive competitors for each numeral. Since *pitu appears at stage 5, the common ancestor of all the languages which show *pitu must be the stage-5 language. By Table 2, in that language, one had innovative forms *pitu, *watlu and *Siwa for 7-8-9, but competing multiplicative and subtractive forms were still available for '8' and '9'. In contrast, for '7', there were no multiplicative and subtractive competitors. For this reason, all the daughters of the stage-5 language have a reflex of *pitu, but only some of them show *walu and *Siwa. Those which do not (Atayal, Sediq, Thao, Favorlang, Taokas and Siraya) all have a multiplicative form for '8', and either a subtractive form or a form of unknown origin for '9'. One last consideration is that the combination in one language of a multiplicative form for 'eight' and an additive form for 'nine' is never observed, probably because both forms would then end in a reflex of *Sepat 'four'. On the other hand, the absence of languages showing simultaneously the additive form for 'eight' and a subtractive form for 'nine' is regarded as coincidental.

The three innovations discussed above have qualities that make them ideal subgrouping criteria:

- their etymology is known and direction of the innovation is certain (in contrast, the etymologies of *enem '6' and *puluq '10' are not known);
- the risk that they might have spread by contact is minimized by the fact that they belong to the (relatively) basic vocabulary;[11]

[11] Although borrowing of numerals is rare in Indo-European languages, numerals, especially the higher numerals, are not immune to borrowing. This is true especially in east Asia, where sets of Chinese numerals have been borrowed by Thai, Be, Miao-Yao, Bai, Baonan among others; where Chamorro has borrowed the entire set of Spanish numerals (Topping 1973: 166); where Qau Gelao has borrowed its numerals from 2 to 10 from Yi (Edmondson and Thurgood 1992) etc. In many (though not in all) cases, the lending language is a state language and the borrowing language has been subjected to the pressure of the lending language within the confines of that state, in the context of a monetary economy. That the power of a state is often behind the transmission of numerals by contact suggests that

- the risk that each of these innovations was made several times independently is almost nonexistent, given the complexity of the six-stage process in Table 2 and the fact that the resulting cognate sets obey the habitual An sound correspondences (in contrast, the *lima '5' < 'hand' innovation consists of a mere semantic shift: it could have occurred several times independently).

I will therefore use the *pitu, *walu and *Siwa innovations as the backbone of a new higher Austronesian phylogeny, shown in Figure 1.

<Figure 1>

I call PAn the language spoken by the first neolithic settlers of Taiwan, from the moment they set foot on the island to the moment when their language broke up into two or more dialects some generations later. I call 'Pituish' the hypothetical daughter language of PAn in which *pitu was first innovated and, having no multiplicative or subtractive competitors, became the sole word for 'seven'. In Table 2 this corresponds to the stage-5 language. Pituish is ancestral to all An languages except Pazeh, Saisiat and Luilang. I assume Pituish already had *Siwa for '9' and a form which I take to have been [watlu] for '8'; but these two forms are not expressed by Pituish's descendants Atayalic, Favorlang, Thao, Taokas and Siraya because of competing multiplicative forms for 'eight' and subtractive forms for 'nine'. In another of its descendants, however, watlu had become walu, and it and Siwa eliminated their multiplicative and subtractive competitors, thereby becoming the only words for '8' and '9'. I call the language where this occurred 'Walu-Siwaish'. I regard this language as ancestral to all An languages except Pazeh, Saisiat, Luilang, Atayalic, Favorlang, Thao, Taokas, Siraya, as well as Papora and Hoanya. Although these two appear to have reflexes of *walu and *Siwa, I suspect they are loans from southern Tsouic, as they show the change w > zero, a regular development in Kanakanabu and Saaroa (compare Kanakanabu *(h)a:ru* 'eight', *siya* 'nine'). Note also initial h- in the Kanakanabu word for '8', reflected in Papora *ma-hal*. Taokas *mahalpat* looks like a blend of a multiplicative form and Papora *ma-hal*.

---

borrowing of numerals would perhaps not have been very frequent in the early neolithic communities which spoke PAn.

Mention must be made of the situation in Rukai. Rukai has a reflex of *walu eight' but has *baŋatə* for 'nine', an isolated innovative form of unknown origin. Since Rukai has been shown to belong to Rukai-Tsouic by Tsuchida (1976) and since all Tsouic languages have reflexes of *Siwa, it is likely that *baŋatə* has displaced a reflex of *Siwa as a Rukai-specific innovation.

That PMP shares the *pitu, *walu and *Siwa innovations with other *Walu-Siwaic languages, as shown in Figure 1, indicates that PMP is not a primary branch of PAn, but part of Walu-Siwaic, a branch of Pituic. This is in agreement with Harvey (1979, 1982), Reid (1982), Starosta (1985, 1994, 1995, 1996, 2001), Ross (1995), Benedict (1995), Bellwood (1997), Ho (1998), Ho and Yang (1999) and Sagart (2002).

4. Enriching the phylogeny

I will now enrich the phylogeny in Figure 1 with additional lexical and morphological characters that are compatible with it. I will also show how the principal sound changes that have affected Formosan languages are to be understood in the present framework. Before I proceed, I need to state that I accept Tsuchida's proposal that Rukai and the three Tsouic languages (Tsou, Kanakanabu, Saaroa) subgroup together on the basis of his documentation, which includes several uniquely shared lexical innovations for basic meanings, notably 'leg', 'nose', hand', shoulder', 'star' and 'river' (Tsuchida 1976:11-12). Atayalic, a taxon consisting of Atayal and Sediq, is self-evident and has been silently accepted in Figure 1. Likewise, NE Formosan, a taxon consisting of Kavalan and Ketagalan is accepted on the basis of the documentation in Blust (1999), who cites a uniquely shared irregular dissimilation in *susu 'breast' > sisu, and in Li (2001) who cites uniquely shared innovations for the meanings 'tooth', 'eyelash', 'spider' and 'unhusked grain/cereal'.

4.1. Fitting more lexical characters into the phylogeny

I now turn to some well-known cognate sets which are often assigned to the PAn level. I will claim that they are post-PAn innovations. Although their etymologies are not all known, they all have an 'opposite number' (by this I mean another Formosan etymon of the same meaning) which I take to be the PAn word they have displaced. This opposite number is represented only in the higher regions of my phylogeny, while the innovative word is

represented in the lower regions (including, in most cases, PMP), with a cut-off point somewhere in between. This distribution forms the basic ground on which the innovation is recognized. Some vagueness in the cut-of point, or overlap in the distributions of the competing forms (apparent resurgence of the old form below the cut-off point) is tolerated: it is recognized that displacement of a lexical item by another is normally effected through a phase during which both words are competing in the language, so that daughters of that language may randomly reflect one or the other, and a degree of overlap between the distributions of the two etyma in a phylogenetic tree may result.

A first group of additional compatible innovative characters is furnished by the other numerals in Table 1, that is, by '5', '6' and '10'.

- *lima 'five'. The 'consensus PAn' word for '5' is *lima, which is also the PAn word for 'hand'. Here the older meaning is presumably 'hand'. This in itself is an indication that *lima is innovative as '5'. A second indication of the innovative character of *lima is the fact that *RaCep, never *lima, is used in the old additive expressions for '6', '7', '8' and '9'; moreover, in those languages that have additive expressions based on '5' for any of '6', '7', '8' and '9', the word for '5' is never *lima. This suggests that *lima did not mean '5' at the time these expressions were created. That the *lima innovation took place later than PAn is shown by the fact that its opposite number *RaCep is (1) etymologically opaque, as befits an inherited word, and (2) is distributed only in the upper region of my phylogeny: specifically in Pazeh, Saisiat and Taokas (Luilang has an obscure form: (na)lup). To the exception of Taokas and Favorlang, the innovative form *lima is universal in Pituic as '5'. One explanation is that the descendants of Pituish included on the one hand the ancestor(s) of Taokas and Favorlang, and on the other hand Limaish, where *lima was innovated: Limaish then broke up into Atayalic, Thao, Siraya, Hoanya, Papora and Walu-Siwaish. Alternatively, *lima already had the meaning '5' in Pituish; but Pituish had not eliminated *RaCep, so that there were competing forms for 'five'; of its daughter languages, Taokas and Favorlang eliminated *lima, while the others eliminated *RaCep. Either explanation is compatible with the tree in Figure 1.

- *enem 'six'. This is the PAn form habitually reconstructed in this meaning. Benedict (1995:400) has gathered intriguing evidence that this etymon was phonetically more complex. He reconstructed *ʔəmləm (based on forms like Makatao *ulum* and Bunun-Ishbukun *ʔabnum* 'six'. Benedict's *l- is equivalent to PAn *N). The etymology of this word is not known but there are good reasons to suppose that it too is a post-PAn innovation: as shown above, some of the highest languages in our phylogeny have additive or multiplicative forms for 'six': Pazeh *xaseb-uza* (5+1), Sediq *ma-teru*, Thao *ka-turu*, etc.: the rest have forms of obscure origin: Luilang *na-tsulup*, Favorlang *nataap* etc. All other Formosan languages, including Siraya, together with all east coast languages and PMP, show reflexes of *enem. In our phylogeny this translates as an innovation from an unknown source, taking place in a language ('Enemish') ancestral to Siraya and Walu-Siwaish. A variant story has *enem arising in Pituish and coexisting with older additive or multiplicative forms; being eliminated in most daughters of Pituish, but eliminating the older forms in Siraya and Walu-Siwaish. However, as we shall see below ('year'), another lexical innovation supports an Enemish node, and I shall here tentatively accept the Enemish node story.

- *puluq 'ten'. This is the 'consensus PAn' word for '10'. In Formosan languages, it occurs exclusively in Rukai (some varieties), Paiwan, Puyuma and Amis, all of which are Walu-Siwaic languages, and outside of Formosa in PMP. Its opposite number is the entirely pre-Walu-Siwaic set given above as #sa-iCit, which must reflect the PAn form. Saisiat has *ranpon*, an isolated form and apparently a local innovation. However the situation is more complex than this simplified account would suggest, as a third cognate set is reflected in Sediq *mahal*, Tsou *máskə*, Kanakanabu *ma:nə*, Saaroa *ma:ɬə* and Bunun (Ishbukun) *masʔan*. An approximation for this set is #masehaN.[12] This etymon appears to occupy an intermediate position between the other two in our phylogeny. I propose the following account: the PAn word was #(sa-)iCit; Pituish innovated #masehaN which competed with #(sa-)iCit, and finally displaced it in Atayalic and in Walu-Siwaish; all other Pituish languages which have not otherwise innovated retain *(sa-)iCit. Walu-Siwaish then innovated *puluq which competed with #masehaN, with the result that Bunun and the Tsouic languages reflect

[12] Tsuchida (1976) reconstructed 'Proto-South-Formosan' *masʔaL '10'

#masehaN, while the rest of Walu-Siwaic reflects *puluq —barring ulterior innovations of the kind of Kavalan *betin* and Ketagalan *labatan*, of course—.

The other lexical innovation which supports an Enemic taxon is the following:

- *CawiN 'year'. The PAn word for 'year' was given as *kawaS by Blust (1999). This is undoubtedly correct, as reflexes of *kawaS occur exclusively in the higher regions of our phylogeny: Pazeh, Saisiat, Atayalic, Thao. Another form for the same meaning: *CawiN has reflexes in Rukai, Kanakanabu, Saaroa, Paiwan and Bunun (Tsuchida (1976:145), to which Sander Adelaar (p.c. 1999) adds Siraya *tawil* 'agricultural season, year'. In terms of my phylogeny *CawiN displaced *kawaS as 'year' in a language ancestral to Siraya and Walu-Siwaish, in other words, in Enemish. An Enemish *CawiN would normally give *tawin in PMP, and this appears to be the form reflected in some Central Cordilleran languages and in Ilokano: *tawen* (Lawrence Reid, p.c., August 2, 2004), although a competing form *taqun is more widespread as 'year' in Malayo-Polynesian.

I will now add other lexical innovations from other areas of the basic lexicon. First I will discuss two lexical innovations supporting a sub-taxon of Walu-Siwaic which I call 'Muic', based on one of its innovations. Muic consists of NE Formosan (Ketagalan, Kavalan), PMP, and a language I call FATK (discussed in section 5). The sharing of these items by Ketagalan (though not by Kavalan) and PMP has been noticed by Paul Li (1995) who misunderstood their innovative character and took them as evidence that Ketagalan migrated to Taiwan around 2000 years ago.[13]

- *-mu '2sg-genitive'. Blust (1977) has argued that a politeness shift replacing the PAn 2sg-genitive pronoun *-Su with the former 2pl-genitive *-mu is a characteristic innovation of PMP. In a more recent paper (1995) he acknowledged that -Su did not disappear as 'your' (sg.) in MP languages, but maintained that -mu is a MP innovation. The coexistence in MP languages of reflexes of -Su and -mu as 2sg-genitive pronouns

[13] In a more recent paper, Li (2001) apparently abandoned the idea that the Ketagalan migrated to Taiwan in a separate migration. His new views are close to those in Blust (1999), although he does not say so himself.

probably means that both existed side by side in PMP: presumably -mu was a polite form. Paul Li (1995) citing an unpublished text recorded by Asai, has pointed out that Trobiawan, a Ketagalan language of north Taiwan, shows that very innovation (*tama imu* 'your father'). It is unclear from Li's account whether Trobiawan *imu* is a polite form and whether Trobiawan also reflects -Su as 2sg-genitive, but Basai (the other Ketagalan language) has *isu* for 'your' (sg.) (Tsuchida, Yamada and Moriguchi 1991:257). This indicates that proto-Ketagalan had variation between *isu* and *imu* for 'your' (sg.), with *imu* presumably the polite variant. The appearance of -mu as polite 2sg-genitive must be regarded as an innovation of the common ancestor of Ketagalan and MP. Since, as mentioned earlier, Kavalan and Ketagalan form a taxon ('North-East Formosan'), and since that taxon must be part of Muic, the question arises as to why Kavalan shows a reflex of -Su, not -mu, as 2sg-genitive. It is possible that Kavalan eliminated the polite pronoun in favor of the non-polite form.

- *manuk 'bird' is reflected in PMP *manuk and Ketagalan *manuk(ə)* (Basai), *manukka* (Trobiawan), but in no other language of Taiwan; its opposite number *qayam 'bird' is widespread in Formosa (including in Kavalan where 'bird' = *alam*). Yet *manuk did not displace *qayam, which is reflected in MP languages in some cases still as 'bird', but more often as 'domesticated animal'. My understanding is that *qayam was the PAn word for 'bird', including the meanings 'wild bird' and 'fowl, domesticated bird'; that *manuk first arose in Muish, from an unknown source, as a hyponym of *qayam meaning specifically 'wild bird': *manuk and *qayam then coexisted in Muish and PMP as 'wild bird' and 'domesticated bird' respectively. Kavalan, a Muic language, abandoned *manuk, keeping *qayam as the only word for 'bird'. Later, in some WMP languages, *qayam expanded its meaning to 'domesticated animal' in general, leaving *manuk free to shift to 'domesticated fowl, chicken', or not.

Together, these two items ('2sg-genitive', 'bird') argue for the existence of a language ancestral to Ketagalan and MP, which I call Muish. Note that neither of these innovations displaced an earlier form of the exact same meaning: the -mu innovation resulted in the addition of a polite pronoun where none existed, and the manuk innovation in the creation of a hyponym of the general term for 'bird'. Both *-Su and *qayam continued existing side by side

with -mu and manuk in the Muish proto-language. Unlike displacing innovations, which are irreversible, these two could be reversed. Judging from the meager evidence at hand, Kavalan seems to have abandoned both innovative forms, keeping only reflexes of *-Su and *qayam, thereby reversing the -mu and manuk innovations. But we should remember that absence of evidence is not evidence of absence. More detailed descriptions of moribund Kavalan are needed to establish this point.

The main lexical innovations and the phylogeny they support are shown in Table 9.

I will now describe another lexical innovation whose discontinuous geographical distribution across Formosan languages apparently challenges my phylogeny.

- *bulaN 'moon'. There are two competing items in Formosa for 'moon': *qiNaS and *bulaN. *qiNaS is limited to Formosa but *bulaN is regularly reflected in PMP *bulan 'moon'. Of the two, *qiNaS is clearly the older, being the only word reflected above Pituish: Pazeh *ilas* 'moon, month', Saisiat *ʔilaŝ* 'moon'. In Pituic, it is also reflected in Sediq *idas* 'moon, month', and in Favorlang *idas* 'moon': but interestingly, it also occurs in the Walu-Siwaic language Paiwan: *qilas* 'moon' (Puyuma/Katipul *qilas* cited by Ferrell 1969:94 is clearly a loan from Paiwan). The other etymon: *bulaN, is reflected in some Pituic languages: Thao *furaz* (Blust 2003), Papora (Hajyovan and Vudol) *voda* (Ino 1998), perhaps also Hoanya (Taorakmun) *pu:loa* (Ino 1998), and in Siraya, Tsouic, Bunun, Amis, Kavalan and Ketagalan. The peripheral and discontinuous geographical distribution of *qiNaS in the northwest and south of the island, in contrast to the compact and more central distribution of bulaN is interesting because on the one hand it forcefully argues that *qiNaS is the old form,[14]

[14] J. Gilliéron was to my knowledge the first scholar to make use of the principle that a feature found in at least two distinct zones at the periphery of a linguistic area is older than the feature which occurs between them. The principle was explicitly formulated by Dauzat (1922), overgeneralized by Bartoli (1925), endorsed by Bloomfield (1935) and more recently by Chambers and Trudgill (1998:94) In Bloomfield's words:

> Especially when a feature appears in detached districts that are separated by a compact area in which a competing feature is spoken, the map can usually be

while on the other hand the discontinuity itself calls for an explanation. I propose the following. The PAn word for 'moon' was *qiNaS. It also meant 'month'. In Pituish *bulaN was innovated as 'moon'. It coexisted with *qiNaS, the basis for their coexistence being that *bulaN meant 'moon' not 'month'. This coexistence lasted through Enemish and Walu-Siwaish, until Muish where *bulaN finally eliminated *qiNaS. Later on, among those pituic, enemic and walu-Siwaic languages which had both, there was a tendency for *bulaN 'moon' to generalize its meaning to 'month' and thus compete with *qiNaS, displacing it in central Taiwan, and separating Paiwan from the western and northwestern languages.

We have collateral evidence to reinforce the presumption that *bulaN is the innovation. In his dictionary, Buck (1949:54) showed that those Indo-European languages which do not show reflexes of PIE *mēnes 'moon' in the meaning 'moon' have independently replaced the inherited PIE word with other words, "most of them from the notion of 'brightness'": thus Gk *σελήνη* from *σέλας* 'light, brightness', Irish *gealach* 'moon', also 'brightness', from *geal* 'bright, white' etc. Further on (1949:1054) he states that "most of the words for 'white' come from the notion of 'bright'". Thus 'moon' and 'white' are normal semantic outcomes for words meaning 'bright'. Reversing Buck's observations, we may state that when a language has identical forms for 'moon' and 'white', there is a presumption that they have a common source in an older word meaning 'bright/brightness'. PMP had *bulan 'moon' and *bulan 'white' (Blust, ACD). We may therefore presume that these two words have a common source in an older word for 'bright(ness)', even though such a word has not been reconstructed.[15] This, then, supports the view that *bulaN 'moon' is a post-PAn innovation, and that *qiNaS is the PAn word for 'moon'.

In this section I have enriched the simple phylogeny in Figure 1 with additional lexical characters and dealt with an apparent piece of lexical

---

interpreted to mean that the detached [districts] were once part of a solid area. In this way dialect geography may show us the stratification of linguistic features. (Bloomfield 1935: 340)

[15] Note Saisiat *bolalas* 'white' (Ferrell 1969).

counterevidence. I will now examine whether any morphological innovations can be fitted into the model.

4.2. Fitting morphological innovations in the model

Six of the eight morphological characters discussed in Starosta (2001) are found both in Saisiat and/or Pazeh and elsewhere in Taiwan. They must be PAn features by our phylogeny:

- Ca- verbal reduplication for deriving instrumental nouns (examples see Blust 1998), seen in Pazeh, Saisiat, Thao, Siraya, Paiwan, Puyuma, Amis, and MP. Starosta assumes it is a post-PAn innovation because of its absence in Rukai and in the Tsouic languages.
- Ca- verbal reduplication marking non-completive aspect (Pazeh, Thao, Atayalic, Bunun, Tsouic, Rukai, Puyuma, Amis, MP). To these, add Siraya (Sander Adelaar, p.c. August 12, 2004), Kavalan (Chang Yung-li 2000:59). Starosta points out that under Blust's flat tree, this distribution must be regarded as the result of independent innovations or of independent losses. In the present phylogeny it is simply a PAn process, lost in Paiwan.
- CV- verbal reduplication marking future or imperfect (Pazeh, Tsou, MP).
- Sa- prefixation marking 'instrumental focus' (Pazeh, Rukai, Amis).
- Sa- prefixation deriving instrumental nouns out of verbs (Pazeh, Rukai, Amis).
- Si- prefixation marking 'instrumental focus' (Pazeh, Saisiat, Bunun, Paiwan, MP). In Starosta's phylogeny this is a post-PAn innovation, and (unlike me) he does not have to explain the absence of this process in Rukai-Tsouic languages.

Fitting these onto the tree in Table 4 is only a matter of assuming the requisite extinctions. For instance with the first item in the list (Ca- verbal reduplication for deriving instrumental nouns), I have to suppose that this PAn process was lost once in proto-Rukai-Tsouic and once in proto-Atayalic, and that any lexical traces it might have left there were later displaced by other forms.

Only two of the processes discussed by Starosta can be post-PAn innovations in my phylogeny, and are therefore potentially informative for early An subgrouping:

- a-prefixation marking future in verbs (Thao, Tsouic, Rukai, Amis).
- paŋ-prefixation deriving instrumental nouns out of verbs (Amis, MP).

The first of these two is not seen in Saisiat or Pazeh, though its absence in these languages could be the result of a loss or of incomplete description. If neither possibility applies, the process can be an innovation of Pituish. In any case we need to suppose several independent events of loss of this feature at later times.

The second process is most likely a post-PAn innovation, as its absence in all of Pazeh, Saisiat, Thao, Atayalic, Favorlang, Siraya, Paiwan and Rukai-Tsouic can hardly be fortuitous. In Starosta's phylogeny as in mine, it is meaningful that Amis, an east coast language, is the only Formosan language to show this feature. In Blust's, it is coincidental. In the phylogeny in Table 4, the paŋ- deverbal instrumental derivation has to be an innovation of Walu-Siwaish; unless independent evidence for a lower-level taxon including Amis and Muic appears.

4.3. Fitting phonological innovations into the model

In the present proposal, for reasons explained above, the suspicion that the sound changes which have formed the basis of several previous attempts at classifying An languages might have spread by contact has relegated them to a secondary place for classification purposes. Their contribution to our understanding of early An phylogeny is limited, because of the risk that they might have spread to already individualized languages. In this section I will show that the principal sound changes having affected Formosan languages are better explained as areal events than as phylogeny-defining events. I will discuss five important mergers:

1. the merger of PAn *C into *t in Ketagalan, Kavalan, Amis, Siraya, Bunun and PMP.
2. the merger of PAn *j into *n in Ketagalan, Kavalan, Amis and Siraya;
3. the merger of PAn *N into *n in Ketagalan, Kavalan, Kanakanabu, Bunun and PMP

4. the merger of PAn *ŋ into *n in Favorlang, Papora, Taokas and Thao.
5. the merger of the PAn phoneme called S_2 in the tradition of Dahl and Ho with S_1 in Amis, Bunun, Puyuma, Kavalan and PMP.

The merger of *C and *t is seen in Siraya, Amis, Bunun, Kavalan, Ketagalan and PMP. The NE languages: Kavalan and Ketagalan, were not in contact with the rest in recent historical times, but this appears to be due to the intrusion of Atayalic on the east coast in the Yilan region a few hundred years ago (Mabuchi 1954). If so, the precursors of all of Siraya, Bunun, Amis, Kavalan and Ketagalan were in contact a few hundred years ago, and may well have been contiguous at the time the merger occurred. That PMP also underwent the change indicates that its Formosan precursor was located within the zone where the change occurred. This means that the change occurred before 4000 BP. However, PMP and Kavalan-Ketagalan did not inherit the change from Muish, because one muic language: FATK (see section 5) did not merge *C and *t. The conclusion must be that the change spread to Kavalan-Ketagalan and to the Formosan precursor of PMP between the breakup of Muish and 4000 BP. This is more parsimonious and realistic than supposing three separate occurrences of *C => *t, one in Bunun, one in PMP and one on the East coast (so Blust 1999: 46, 52).

Blust (1999:46) regards the merger of *j and *n as the defining innovation of his 'East Formosan', a construct comprising Ketagalan, Kavalan, Amis and Siraya. This, again, cannot be a monophyletic taxon in my phylogeny. In order to account for the facts under my phylogenetic assumptions, a similar scenario as for *C => *t would have to be supposed, with the innovation arising on one coast and spreading to the other, this time leaving out PMP. In this case however, the hypothesis of a spread from one coast to the other is less easy to maintain than with *C => *t, because geographically intermediate Bunun did not undergo the change, and it is not clear that Siraya was ever in direct contact with the east coast languages that have undergone *j => *n.[16] Supposing that such direct contact existed earlier on between the precursors of Amis and Siraya would be a leap of faith. Another

[16] In their study of Ogawa's material on Siraya varieties, Tsuchida and Yamada (in Tsuchida, Yamada and Moriguchi 1991, for instance on pp. 55, 57, 60, 65, 89, 107 etc.) identified a small number of Amis loans into a variety of Taivoan -a dialect of Siraya- spoken in the village of Dazhuang 大庄. This is a southern outlier of Siraya, however. No loans to other varieties of Siraya were identified.

troubling element with this change is that it affects simultaneously the place and mode of articulation of its target phoneme, and involves the highly unusual process of spontaneous nasalization: in a word it is a highly unnatural change and I strongly doubt that such a merger ever occurred in Taiwan. Suffice it to say here that another interpretation of the facts is possible: the correspondence identified as PAn *j was a palatal nasal in PAn, not a stop. The main innovation with this phoneme was its shift, under systemic pressure, to a voiced palatal stop, in all the languages of Taiwan (including the ancestor of PMP) with the exception of two conservatives zones: Amis and Kavalan-Ketagalan on the east coast and Siraya on west coast, where the phoneme preserved its nasal character: in these two areas independently, again under systemic pressure, the palatal nasal merged with *n, as part of the general process of loss of palatal sounds. It is this merger which gives the appearance of a palatal stop merging with *n. A full discussion of the phonetic value of the PAn phoneme identified as *j requires a reconsideration of the PAn consonant system, however, and I must reserve it for another occasion.

The merger of *N into *n affected Bunun, Ketagalan, Kavalan and PMP, but not Amis, and Kanakanabu but not the rest of Tsouic. Lack of contact between these languages in historical times again seems due to the late intrusion of Atayalic, and a single event followed by geographical spread can explain the membership of this change.

The fourth change is the merger of n and ŋ. This is one of the innovations characterizing Blust's Western Plains group (Blust 1999), which comprises Papora, Hoanya, Favorlang, Taokas and Thao. Here the position of Hoanya is uncertain. On p. 44 Blust does not list n/ŋ as a merger in Hoanya, yet he places Hoanya next to Papora as a Central Western Plains language on p. 45. However this may be, all these languages were in contact in historical times and the basic requirement is met for a sound change to spread by contact. There is indeed some evidence to show that contact played a role in the distribution of this feature.[17] Note however, that there is nothing in the present phylogeny to contradict Blust's Western Plains group.

[17] There seems to be a difference between southern Hoanya (self-designation Lloa, represented by Taorakmun in Ino's data) and northern Hoanya (self-designation Arrikun: represented by Savava in Ino's data): thus 'ear', PAn *Cariŋa, Savava *sangera*, Taorakmun

The fifth change is the merger of the sounds identified as S_1 and S_2 in the tradition of Dyen, Tsuchida, Dahl and Ho Dah-an in Amis, Bunun, Puyuma, Kavalan and PMP. An example of a word with S_2 is *kaS_2uy 'wood, tree'. Again, a full discussion of the PAn consonant system must be reserved for another occasion, but some initial remarks on S_2 are in order here, as the existence of S_2 as a separate phoneme is controversial: Blust does not regard it as distinct from S_1. The reflexes of S_1 and S_2 are however markedly different (Dahl 1981:35; Ho 1998:165). On the other hand, while S_2 is typically viewed as a sibilant with a different point of articulation from S_1, it is curious that S_1 and S_2 are never reflected as different sibilants in the same language. In fact, with some exceptions in Bunun, the only languages which reflect S_2 as a sibilant are Amis, Bunun, Kavalan and PMP, precisely those which merge it with S_1.[18] In all the languages where S_2 is not merged with S_1, its reflexes are identical with those of PAn *h (a.k.a. *H_1). This suggests that the correspondence called S_2 defines a subset of PAn *h- (probably *h- before a high front vowel) which palatalized to ɕ- after PAn, thereby merging with *S_1 in an unbroken contact area on the east coast covering Amis, Bunun, Kavalan and the Formosan ancestor of PMP. Due to the small number of forms including S_2, it is unclear whether Ketagalan participated in this merger. At any rate the membership of this sound change, and of the other sound changes discussed in this section as well, is adequately explained by geographical contact.

Overall, the contribution of sound changes to the subgrouping of Formosan languages is not great. The most useful information they provide is geographical: what languages were in contact at the time when a change took place. WE can use this information to probe the location of the Formosan precursor of PMP: that language must have been spoken near the precursors of Bunun and Kavalan-Ketagalan, since these are the languages with which it shares the most sound changes. This probably indicates an east-coast location.

---

sa:rinna (Ino 1998). This suggests that the change spread to Hoanya without affecting all its dialects.

[18] Puyuma also merges the two but the result is zero. It is not clear whether S_1 and S_2 converged towards zero in Puyuma or first merged as a sibilant.

The enriched phylogeny in Table 4 summarizes the discussion of lexical and morphological innovations in this section.

<Table 4>

5. The position of Tai-Kadai

The observant reader will have noticed that Table 4 includes two languages named 'FATK' and 'FAMP'. These acronyms means 'Formosan Ancestor of Tai-Kadai'[19] and 'Formosan Ancestor of Malayo-Polynesian'[20] respectively. With respect to the former, the claim is being made that, contrary to common sense, the Tai-Kadai languages are descended from an East Formosan language: in effect, that they are a branch of Austronesian, and specifically, a subgroup of Muic.

The modern Tai-Kadai languages are spoken in parts of south China, Vietnam, Laos, Thailand, Burma and Assam. Ostapirat (2000)[21] distinguishes three branches: Hlai in Hainan, and on the mainland Kam-Tai and Kra, but in a more recent paper (in press), he suggests a different classification, with a northern branch (Kra and Kam-Sui) and a southern branch (Hlai and Tai). The Tai-Kadai languages outside of south-east China and adjacent areas of Vietnam are all within the Tai subgroup, and are very homogeneous. This is due to the historically well-attested expansion of Tai speakers in the late first and early second millennia CE. The area of highest diversity is in the north-east part of the Tai-Kadai domain: in Hainan Island, in northern Vietnam, and in the Chinese provinces of Guangxi and Guizhou. This is presumably where the TK homeland was located (although some of the original diversity must have been lost to Chinese in Guangdong province).

On the ground that they share a remarkable set of very basic vocabulary items (personal pronouns, numerals, body part terms, basic verbs), Benedict (1942) proposed that the An and TK families are coordinate within an

[19] In Sagart (2001; in press, a), FATK was called 'AAK'. I will from now on use the synonymous, but more specific term FATK.

[20] I introduce the acronym FAMP to distinguish the pre-migration language from PMP proper, which I believe was spoken in the Philippines.

[21] Ostapirat (2000) uses the name 'Kra-dai' for Tai-Kadai

'Austro-Tai' macrophylum. Yet, because of his exuberant methodology, his proposal did not meet with the full approval he expected, in particular from Austronesianists. Yet, as argued in Sagart (2001; in press, a) and Ostapirat (in press), sets in Benedict's basic vocabulary comparisons can be isolated which exhibit strong phonetic regularities,[22] as shown in Table 5. This kind of evidence virtually eliminates the possibility of chance resemblances.

<Table 5>

For the three words in Table 5 Benedict reconstructed Proto-Austro-Tai *maplay 'die', *mapla 'eye' and *mamlok 'bird': the medial clusters then evolving to *C and *N in PAn, and either to pl-, ml- or to t-, n-, in Tai-Kadai. To him, this was proof that a language ancestral to both PAn and PTK was needed to explain the sound correspondences between them. However, both Sagart (in press, a) and Ostapirat (in press) reject this interpretation. Sagart argues that the TK forms are better accounted for on the ground of cluster-less forms like PAn maCa, maCay, or PMP manuk, based on an explanation originally proposed by Haudricourt (1956). Thus, in the case of 'eye':

maCa > mCa > pCa > pta > pla ~ ta

Sagart observed that in general, most An-related TK forms can be adequately explained on the ground of PAn, or even PMP.

To remove any lingering misconceptions that the resemblances between Tai-Kadai and Austronesian are due to chance, I give below comparisons involving Buyang, a recently-described Tai-Kadai language from the Kra branch, spoken near the China-Vietnam border (data from Li Jinfang 1999). While TK languages generally reflect An disyllables as monosyllables (either by losing the first syllable, or by collapsing the two), Buyang is remarkable in that it preserves several An disyllables as disyllables (Table 6). Observe that in these words, the first syllable is reduced: the vowel is always /a/, the syllable is toneless ('tone zero'), and the inventory of initial consonants is

[22] Ostapirat (in press) has greatly clarified the sound correspondences between An-TK, although he cautiously refrains from characterizing the An-TK relationship as genetic as opposed to contact-induced.

limited to a few: m- (for An m- and w-), q- (for An q- and k-), t- (for An C- and t-). Yet Buyang shows that An words in Proto-TK were still disyllabic.

<Table 6>

It is noteworthy that among the best An-TK comparisons are personal pronouns and numerals, as shown in Table 7 and Table 8:

<Table 7>

<Table 8>

What is remarkable about the vocabulary shared by Tai-Kadai and Austronesian is not so much the very basic nature of the shared elements as the paucity of credible comparisons in the cultural vocabulary, notably the vocabulary of agriculture (Sagart 2003), the names of domestic animals, and the vocabulary of house-building. I have argued (Sagart 2001; in press, a) that this is not compatible with Thurgood's explanation in terms of An loans to Tai-Kadai (Thurgood 1994): Tai-Kadai could not plausibly borrow principally basic vocabulary from An. The only realistic explanation left is genetic, as Benedict thought, although the explanation I have proposed is different from his.

Benedict regarded TK as a very old taxon, with an ancestral language of a comparable age to PAn, or even older. Misled by exaggeratedly early archaeological dates for bronze in North Thailand (Solheim 1971), he characterized proto-Austro-Tai as the bearer of the high culture in early east Asia, and the source of many loanwords to Chinese, in particular in the domains of metallurgy and agriculture. It is now crystal-clear that the loans went the other way (Sagart 1999 for metal names). Benedict located the Austro-Tai homeland in south-eastern China. However, Ostapirat's recent reexamination of the family, independently supported by Peiros's glottochronological study, shows the date to be considerably more recent, not earlier than 2000 BCE (Ostapirat, in press) or 1800 BCE (Peiros 1998:15).[23]

[23] Peiros did not take the Kra languages into account in his calculations, however.

Inspection of Table 6, Table 7 and Table 8 shows that where PAn and PMP disagree, Buyang sides with PMP. On the ground that they share the *-mu '2sg-genitive' and *manuk 'bird' innovations with PMP, I have argued (Sagart 2001; in press, a) that the Tai-Kadai languages are a subgroup of Austronesian, closely related to PMP, rather than a related but separate language family. Table 7 now shows that Buyang also sides with PMP against PAn in the matter of numerals. The inescapable conclusion, then, must be that TK is a subgroup within An. The interesting question with the An-TK relationship, in my view, is not so much 'why are there so many An words in TK ?' as 'why are there so few ?'. [24] My tentative answer (Sagart 2001, in press, a) is that TK evolved on the mainland out of the Formosan Austronesian language I call FATK; and that, once on the mainland, it underwent intimate contact with, and extensive relexification from, a local language having not left any other descendant (although macrophylic connections to AA or MY are a distinct possibility). In the course of that period of contact, a large part of the original An vocabulary of TK was lost, with only the most basic part of the vocabulary resisting relexification.

Under the present theory of An subgrouping, sharing the *-mu and *manuk innovations with PMP and with Ketagalan quite definitely makes FATK a muic language. It is interesting to consider the predictions made by this claim. If Tai-Kadai truly goes back to a muic language, and barring, of course, ulterior innovations, then we should expect it to have all the post-PAn innovations discussed earlier in this paper, such as the short forms of the numerals '7', '8' and '9'. It would be devastating if Tai-Kadai reflected PAn etyma which I have claimed had already been displaced by newer words in muic: for instance if Tai-Kadai reflected *RaCep-i-tuSa as 'seven', *RaCep as 'five', *(sa-)iCit or #masehaN as 'ten', *kawaS as 'year', *qiNaS for 'moon' etc. On the contrary it would support my theory if Tai-Kadai reflected *pitu as 'seven', lima as 'five', *puluq as 'ten', etc.

The predictions on Tai-Kadai of the present theory of An phylogeny are verified, a shown in Table 9. The original Tai-Kadai numerals are preserved

[24] This is the question asked by Peiros (1998:103): while he recognized the basicness of the list of the list of An-TK lexical comparisons, and noted it was indicative of a genetic relationship, he was puzzled by its brevity, and by the failure of attempts to enlarge it.

only in the Hlai and Kra branches: in the Kam-Tai branch they have been replaced with Chinese numerals. In Table 8 we have already seen the numerals of Buyang, a Kra language: the resemblance between the Buyang and MP numerals is obvious. The Proto-Kra numerals from five to ten, as well as the words for 'your' (sg.) and 'bird' as reconstructed by Ostapirat (2000), are listed in Table 9.

<Table 9>

The sound correspondences with PAn are only beginning to be elucidated (Ostapirat in press), and a full demonstration that all the segments of P-Kra words regularly correspond to An words cannot yet be made. Even so, one can make some preliminary observations.

The PK word for 'five': *r-ma matches the last syllable of *lima and probably the first consonant too, as there are parallels for An l :: PK r ('eight'). There are also good parallels for An m :: PK m and An a :: PK a. Ostapirat reconstructs PK *x-nəm 'six', where 'x' is an element aiming at accounting for the unexplained alternation between high and low tones in this set. It may have been a glottal stop. Benedict (1975: 212) similarly reconstructs *nəm 'six' for proto-Hlai in Hainan. This is close to An *enem. Ostapirat (2004) gives parallels for An *n- :: TK *n- and An *-m :: TK *-m. Ostapirat's reconstruction for 'seven': *t-ru, is problematic. The r- in it is nowhere reflected as a r-type sound, and Hlai indicates a simple voiceless t-. I believe an alternative PK reconstruction for 'seven' is *(C-)tu (where C is a voiceless stop), meaning that PK had an alternation between *C-tu and *tu, with a majority of languages reflecting the latter, but Paha *C-tu > ʔd- > ʔr- > *ðhuu*$_{A1}$. If so, *(C-)tu is a likely match for *pitu: Ostapirat (2004) gives other examples of *t :: *t and *u :: *u.[25] In Ostapirat's PK reconstruction for 'eight', -ru matches the last syllable of *walu, with the same An l :: PK r correspondence as in 'five'. Whether or not Kra initial m- regularly matches An *w- is unclear. We have seen that in Buyang, m- in the first syllable of '8' matches An m- and w-, but it is unclear whether this is true of the other Kra languages. The Kra language Laqua prefixes mə- to all of 'seven', 'eight' and

[25] Weera Ostapirat (p.c., January 2004) indicates that the reconstruction *C-tu for '7' had been proposed in his PhD dissertation. Although he abandoned it in his book (Ostapirat 2000), he now considers it preferable on Kra-internal grounds.

'nine' (Benedict 1975: 212), perhaps as a result of leveling. For Proto-Gelao (a subgroup within Kra) Ostapirat (2000:122) reconstructs initial *wr- in 'eight', which appears to agree well with An *walu. Proto-Gelao *wr- is not the regular outcome of PK *m-r-: it could represent the original Kra onset of this word. Finally, Proto-Kra *s-ɣwa 'nine' is an attractive match for *Siwa, especially since the correspondence An *S- :: PK s- has parallels (notably *duSa 'two' : PK *sa $_A$ 'two'). The correspondence An *-w- :: PK ɣw- has a parallel in PAn *duwa 'come, go' (Pazeh *dua* 'go', Puyuma *Zowa* 'come'): PK *ɣwa C 'go'. With 'ten' the final consonant correspondence An -q :: PK -t seems off but Ostapirat (in press) treats it as regular following /u/.

Tai-Kadai, then, has very plausible candidate reflexes for the post-PAn innovative forms in Table 9. I know of no case of a PAn word being displaced by an innovation after PAn and before Muish, and at the same time having an attractive TK comparison. This supports the view that TK is a daughter language of PAn, and more specifically a muic language. FATK cannot be a MP language because, as discovered by Ostapirat (in press), TK preserves the distinction between PAn *C and *t, and, at least in some words ('two'), has a sibilant reflex of PAn *S.

Ostapirat further argues that, contrary to my claim, TK cannot have its origin in an early east Formosan language, because all east coast languages merge *C and *t. He considers that, if the relationship between Tai-Kadai and An really is a genetic one, Tai-Kadai must be coordinate with the whole of Austronesian, as Benedict thought, or, at least, with a higher-order taxon than East Formosan-PMP. I maintain, based on the evidence of lexical innovations shown in Table 9, that Tai-Kadai is part of Muic. As indicated earlier, I believe that the merger of *C into *t spread to Ketagalan and to the Formosan ancestor of PMP but failed to reach FATK, after the breakup of Muish. In fact there is a tiny bit of evidence to suggest that FATK underwent a merger which cannot be traced back earlier than Walu-Siwaish, and which no language of the west coast exhibits, to wit, the merger of S_1 and S_2: Buyang has *ɕa* 'two', corresponding to S_1 in PAn duS_1a 'two', and *ɕui* 'fuel', corresponding to S_2 (my *h) in PAn *kaS_2uy (my *kahiuy) 'wood'.

Recognition that TK is a subgroup of An opens new perspectives for the reconstruction of PAn, notably in the area of final laryngeals, which are reflected in TK tones.

I am claiming, therefore, that the tree in Table 4 is an approximation of the higher phylogeny of the An family, including its newly-recognized subgroup Tai-Kadai.

6. The An phylogeny in space

Supposing that the implicational hierarchy in Table 1 was accidental, and that the etymologies proposed for *pitu, *walu and *Siwa in section 3 were fanciful, we should not expect the phylogeny proposed in Table 4 to result in any kind of recognizable geographical pattern. Yet a clear pattern emerges. Approximate geographical locations for PAn, Pituish, Enemish, Walu-Siwaish and Muish can be determined based on the location of their direct descendants in historical times. The most likely location of PAn is in the region of Luilang, Saisiat and Pazeh, in the north-west of Taiwan. Pituish must have been spoken in the western plains somewhat to the south of PAn, being ancestral to Atayalic, Favorlang, Taokas, Thao, Papora and Hoanya. There is a tradition that the present-day northern location of Atayal and Sediq was reached following migrations from west-central Taiwan (Mabuchi 1954): Blust (2003:6) reports that Thao was previously spoken near Jiayi in the western lowlands, reaching Sun-Moon Lake only 300-350 years ago. Enemish must have been spoken more to the south on the West coast, towards the area occupied by Siraya in recent times. Walu-Siwaish may have been spoken near the southern tip of the island, or on the south-east coast. Muish was probably further north along the east coast, as suggested by the location of its modern descendants Kavalan and Ketagalan in the NE of the island, and the observation made in section 4.3. that FAMP must have been spoken in contact with the precursors of Kavalan-Ketagalan and of Bunun.

This model shows a consistent geographical pattern: early Austronesian speakers settling Taiwan progressively in a counter-clockwise movement, starting from the north-west, then expanding southward along the west coast, and reaching the southern tip of the island before finally settling the east coast from south to north, as shown in Figure 2.

<Figure 2>

The progression of the early ANs in Taiwan is an illustration of the "wave of advance" model (Cavalli-Sforza and Ammerman 1984), which describes the progression of neolithic settlers in areas not yet touched by agriculture. In Taiwan, the cumulative character of the linguistic innovations at each stage leaves no other explanation than the progression of such a wave of advance, gradually encircling the island (although minor population movements, leaving no linguistic traces, must certainly have occurred).

One may wonder why there was no clockwise progression from north-western Taiwan around the northern tip of the island towards the east coast. I conjecture that the large and possibly malarial freshwater lake or swamp which at the time occupied the Taipei basin rendered movement in that direction unattractive.

Although the location of Pituish, Enemish, Walu-Siwaish and Muish cannot be established with a high degree of precision, the general pattern is clear: a gradual, unidirectional encirclement of the island by Austronesian speakers. Apparently the main direction of movement was along the coastal plains. This implies that, given a choice, the early Austronesians preferred to expand into the coastal plains. This pattern is consistent with what archaeology and linguistics tell us about their mode of subsistence, which combined exploitation of marine resources, including fishing, with hunting and gathering and cultivation of rice and millet. We may suppose that population movements into the mountains, as with the Saisiats, Atayalics, Thaos, Tsouics and Bununs, were generally late, and made under pressure. Such indeed is the pattern observed in the rest of the Austronesian world (Blust 1999:53). The pattern of progression from the west to the east coast is moreover consistent with archaeological dates for Ta-Pen-K'eng sites, which are older on the west coast than on the east coast. It is tempting to imagine that the Nan-kuan-li people, who were active near Tainan c. 5000-4500 BCE (Tsang, in press) spoke a form of Enemish, while the Yuan-Shan people, who were active north of Taipei from around 4500 BP (Bellwood 1997:215), spoke a form of Muish.

The geographical stability over time of the initial settlement pattern is striking. Most modern languages are still spoken or were still spoken until recently in the area of the meso-language they are descended from. A major factor in this is the geography of Taiwan, where the central mountain range

very effectively prevents contact and migration between the east and west coasts.

Finally, under the present interpretation, FAMP and FATK, the two Muic languages whose speakers left Taiwan to settle other regions, were probably located in the north-east or north of the island, where the last available agricultural lands had been. The MP and TK migrations out of Taiwan thus appear motivated by the need to find new agricultural lands. It is probably no coincidence that the site of Yuan-Shan near Taipei, in the region where Ketagalan was spoken until the early 20th century, has significant connections to the earliest neolithic of the Philippines (Bellwood 1997: 215).

7. The time scale of the early An settlement of Taiwan

The primary evidence for the time scale of the Austronesian settlement of Taiwan comes from archaeology. Bellwood's estimates for the date of the initial Austronesian settlement of Taiwan, inferred from the earliest Ta-Pen-K'eng radiocarbon dates (plus a few hundred years for good measure), and from the earliest neolithic sites in the Cagayan Valley in the northern Philippines, are c. 5500 BP for the former and ca. 4000 BP for the latter (Bellwood 2004). During that period, the daughter languages of PAn were presumably confined to Taiwan. While these are provisional dates, they provide approximate external limits between which the full settlement of Taiwan must have taken place: in the present framework, the initial settlement of the Philippines, presumably by PMP speakers, could not have occurred until Muish had already broken up into its three components. Likewise, the initial Tai-Kadai settlement on the mainland, by FATK speakers, could not have occurred before the breakup of Muish. It is relevant here to recall that Ostapirat estimates the date of PTK to be no older than 4000 BP, simultaneous with the Cagayan dates.

That two sound changes: merger of *C and *t and merger of *n and *N, spread to FAMP but not to FATK (Ostapirat, in press, claims these four initials were distinguished in proto-TK) indicates that a significant amount of time elapsed between the break-up of Muish and the MP migration out of Taiwan, even though it is impossible to say precisely how much. That FATK failed to undergo these changes means that either it was located further north along the east coast than FAMP and the changes, spreading from the south,

stopped at FAMP; or alternatively that the TK migration was earlier than the MP migration.

8. Archaeology and language: some conjectures on pre-Austronesian times

I have argued elsewhere (Sagart, in press, b) that the pre-Austronesians spoke a language related to Sino-Tibetan, and that they reached Taiwan from a location in NE China where millet and rice were cultivated, and where ritual evulsion of the upper lateral incisors in boys and girls was practiced. The eastern China seaboard region north of the Yangzi estuary, from north Jiangsu to north Shandong, is the one area in East Asia where the distribution of these three traits overlaps in the period before the arrival of the Austronesians in Taiwan: thus both rice and millet were cultivated in Xihe in north Shandong (Wright 2004) c. 8000 BP and in Longqiuzhuang in the lower Huai basin c. 7000-5000 BP. Tooth evulsion is attested from 6500 BP on in Shandong and north Jiangsu (Han and Nakahashi 1996). We may surmise that before they reached Taiwan, the pre-Austronesians were expanding southward along the coastal plains of central-eastern China in Jiangsu, Zhejiang and north Fujian. We can expect that archaeological sites with rice, Setaria, tooth evulsion, and a technology intermediate between the Dawenkou culture of north-east China and Ta-Pen-K'eng of Taiwan will eventually appear there.

If this scenario is correct, it is likely that the passage to Taiwan did not exhaust the pre-An population of the Fujian coast. More likely, this population continued expanding along the coast in a south-westerly direction towards the Pearl River delta, even after a group of them had crossed to Taiwan. Their archaeological traces SW of Fujian are perhaps seen in the Pearl river delta, although direct evidence of agriculture there has so far not appeared; Hedang in the Pearl River delta, with tooth evulsion (Higham 1996:84), c. 3000-2000 BCE, may be one such site. In Taiwan, Tsang (in press) describes the newly excavated site of Nan-kuan-li near Tainan in south-west Taiwan, where a team led by him recently discovered a neolithic culture having rice, millet, and practicing ritual tooth ablation around 5000-4500 BP. In the same paper he argues that the Ta-Pen-K'eng culture, as seen in Nan-kuan-li near Tainan, "has close affinities with the Neolithic cultures of Hong Kong and the Pearl River Delta". I disagree with Tsang when he concludes that "The Pearl River Delta of Kuangtung is most

probably the source area of the Tapenkeng Culture in Taiwan". I think it more likely that both cultures are descended from a common precursor on the Fujian coast. Pearl River delta sites having affinities to Taiwan TPK like Hedang are also probably too early and too far east to be ancestral to the Tai-Kadai-speaking cultures.

9. conclusion

I have presented an explicit account of the early phylogeny of the Austronesian family. The new phylogeny is tree-like. A salient characteristic is that out of a majority of nodes, only one branch leads to further branching (Table 4). This makes Formosan phylogeny similar to Malayo-Polynesian phylogeny. Non-branching nodes can be associated with stay-at-homes, and branching ones with out-migrating groups. PMP has been shown to be part of a taxon that also includes languages of the NE Formosan Coast, as well as Tai-Kadai (as proposed in Sagart 2001; in press, a). That taxon itself is part of a larger taxon including languages of the East coast and south Taiwan.

These proposals have been made on the ground of the convergence of three independent lines of evidence: (1) the implicational hierarchy with the numerals 5-10, shown in Table 1; (2) the systematic resemblances between the consensus numerals for 7-8-9 and the corresponding numerals in Pazeh, described in section 3; and (3), the geographically coherent and processually realistic spatial pattern of settlement shown in Figure 2. (2) is obviously independent from (1) and (3); and (1) and (3) are independent from each other because one could have an implicational hierarchy which did not result in a coherent spatial pattern. None of these independent sets of facts actually 'proves' the phylogeny in Table 4: rather, that phylogeny makes sense of them all: while the same facts have to be regarded as coincidences under other phylogenies. In effect, (1), (2) and (3) are three independent verified predictions of the phylogenetic hypothesis in Table 4. Because it makes more verified predictions than earlier hypotheses, it should be preferred.

References

Adelaar, K. A. 1997. Grammar notes on Siraya, an extinct Formosan language. *Oceanic Linguistics* 36, 2: 362-397.

Bartoli, M. 1925. *Introduzione alla neolinguistica. Princìpi, scopi, metodi*. Geneva: Olschki.

Bellwood, P. 1997. *Prehistory of the Indo-Malaysian archipelago*. Honolulu: University of Hawai'i Press.
-------- 2004. Taiwan, Batanes, Cagayan, Talaud, Maluku, Marianas, Lapita Gardens, and all stations to Polynesia. All aboard the Austronesian Express. *Paper presented at the workshop on Human migrations in continental East Asia and Taiwan*, Geneva, June 10-13, 2004.
Benedict, P.K. 1942. Thai, Kadai and Indonesian: a new alignment in Southeastern Asia. *American Anthropologist, n.s.,* 44: 576-601.
-------- 1972. *Sino-Tibetan: a Conspectus*. Cambridge: University Printing House.
-------- 1975. *Austro-Thai: language and culture*. New Haven: HRAF Press.
-------- 1995. Extra-Austronesian evidence for Formosan etyma. In *Austronesian studies relating to Taiwan*, ed. by Paul J.-K. Li, C.H Tsang, Y.-K. Huang, D.-A. Ho and C.Y. Tseng, pp. 399-454. Symposium series of the Institute of History and Philology, Academia Sinica 3. Taipei: Academia Sinica.
Bloomfield, L. 1933. *Language*. New York: Holt.
Blust, Robert A. 1977. The Proto-Austronesian pronouns and Austronesian subgrouping: a preliminary report. *University of Hawai'i working papers in Linguistics* 9, 2:1-15.
Blust, R. 1995. Sibilant assimilation in Formosan languages and the Proto-Austronesian word for 'nine'. *Oceanic Linguistics* 34:443-453.
-------- 1998. Ca- Reduplication and Proto-Austronesian grammar. *Oceanic Linguistics* 37:29-64.
-------- 1999. Notes on Pazeh phonology and morphology. *Oceanic Linguistics* 38:321-365.
-------- 2001. Malayo-Polynesian: new stones in the wall. *Oceanic Linguistics* 40:151-155.
-------- 2003. *Thao Dictionary*. Nankang: Institute of Linguistics (preparatory office), Academia Sinica.
Bril, I. 2002. *Le nêlêmwa (Nouvelle-Calédonie): Analyse syntaxique et sémantique*. Collection "Langues et Cultures du Pacifique", n° 16. Paris: Peeters.
Buck, C. D. 1949. *A dictionary of selected synonyms in the principal European languages*. Chicago: Chicago University Press.
Carstairs, D. 1899 *edn. Chinese-English dictionary of the vernacular or spoken language of Amoy, with the principal variations of the Chang-chew and Chin-chew dialects.* London: Presbyterian Church of England.
Cavalli-Sforza, L.-L., and A. Ammerman. 1984. *The Neolithic Transition and the Genetics of Populations in Europe*. Princeton: Princeton University Press.
Chambers, J.K. and P. Trudgill. 1998. *Dialectology*. Cambridge: Cambridge University Press.
Chang, Yung-li. 2000. *Gemalan yu cankao yufa*. Taipei: Yuan-liou.
Dahl, O. Ch. 1981. *Early phonetic and phonemic changes in Austronesian*. Instituttet for sammenlingnende kulturforskning Serie B: Skrifter LXIII. Oslo: Universitetsfor-laget.
Dauzat, A. 1922. *La géographie linguistique*. Paris: Flammarion.
Edmondson, J. and G. Thurgood. 1992. Gelao reconstruction and its place in Kadai. Paper presented at the 25th International Conference on Sino-Tibetan Languages and Linguistics, 14-18 Oct. 1992, Berkeley, California.
Egerod, Søren. 1980. *Atayal-English Dictionary*. Scandinavian Institute of Asian Studies monograph 35. London and Malmø: Curzon Press.
Ferrell, R. 1969. *Taiwan Aboriginal groups: problems in cultural and linguistic classification*. Monograph No. 17, Institute of Ethnology, Academia Sinica. Nankang: Academia Sinica.
-------- 1982. *Paiwan dictionary*. Pacific Linguistics C-73. Canberra: A.N.U.

François, Alexandre. 2001. *Contraintes de structures et liberté dans la construction du discours. Une description du mwotlap, langue océanienne du Vanuatu.* Doctoral dissertation. Paris: Université Paris-IV Sorbonne.
Fujian Provincial Museum. 1991. Brief report on the excavation of the Keqiutou site in Pingtan, Fujian [in Chinese]. *Kaogu* 1991, 7:587-599.
Han, K.X. and T. Nakahashi. 1996. A Comparative Study of Ritual Tooth Ablation in Ancient China and Japan. *Anthropological Science* 104,1:43-64.
Harvey, Mark. 1979. Subgroups in Austronesian. BA Honours thesis. Canberra: A.N.U.
-------- 1982. Subgroups in Austronesian. In *Papers from the Third International Conference on Austronesian Linguistics*, vol. 2: *Tracking the travellers.* (Pacific Linguistics C-75), ed. by A Halim, L. Carrington and S.A. Wurm, pp. 47-99. Canberra: ANU.
Haudricourt, A.-G. 1956. De la restitution des initiales dans les langues monosyllabiques : le problème du Thai commun. *Bulletin de la société de Linguistique de Paris* 52:307-322.
Higham, C. 1996. *The Bronze age of southeast Asia*. Cambridge: Cambridge University Press.
Ho, Dah-an. 1998. Taiwan Nandaoyu de Yuyan Guanxi. *Chinese Studies* 16, 2:141-171.
Ho, Dah-an and Yang Hsiu-fang. 1999. Nandaoyu yu Taiwan Nandaoyu. *Introduction to a collection of monographs on Taiwan languages*. Taipei: Yuanliou.
Huang, Lillian. 2000. *Beinan yu cankao yufa*. Taipei: Yuanliou.
Imbault-Huart, C. 1893. *L'île Formose*. Paris: Ernest Leroux.
Ino, Y. 1998. Xun Tai Ri Cheng. In *Ino Yoshinori Fanyu Diaocha Shouce*, ed. by T. Moriguchi, pp. 13-201. Taipei: Southern Materials Center.
Li, Jinfang. 1999. *Buyang Yu Yanjiu*. Beijing: Zhongyang Minzu Daxue.
Li, P. J.-K. 1975. *Rukai texts*. Nankang: Institute of History and Philology special publications No. 64-2.
-------- 1995. Formosan vs. non-Formosan features in some Austronesian Languages in Taiwan. In *Austronesian Studies Relating to Taiwan*, ed. by Paul Jen-kuei Li, Dah-an Ho, Ying-kuei Huang, Cheng-hwa Tsang, and Chiu-yu Tseng, pp. 651-681. Symposium series of the Institute of History and Philology, Academia Sinica, no. 3. Taipei: Academia Sinica.
-------- 2001. Basai yu de diwei [in Chinese]. *Language and Linguistics* 2,2:155-171.
Li, P. J-K, and S. Tsuchida. 2001. *Pazih dictionary*. Nankang: Institute of Linguistics (preparatory office), Academia Sinica.
-------- 2002. *Pazih texts and songs*. Nankang: Institute of Linguistics (preparatory office), Academia Sinica.
Lin, Yingjin. 2000. *Bazehai yu*. Taipei: Yuanliou.
Mabuchi, T. 1954. Migration and distribution of the Formosan aborigines [in Japanese]. *Minzogaku Kenkyu* 18,1-2:123-154; 18,4:23-72.
Matisoff, J.A. 1997. *Sino-Tibetan numeral systems: prefixes, protoforms and problems*. Canberra: Pacific Linguistics B-114.
Mazaudon, M. 1985. Dzongkha number systems. In *Southeast Asian linguistic studies presented to André G. Haudricourt*, ed. by S. Ratanakul, D. Thomas and S. Premsrirat, pp.124-157. Bangkok: Mahidol University.
-------- 2002. Les principes de construction du nombre dans les langues tibéto-birmanes. In *la pluralité*, ed. by J. François, pp.91-119. Paris: mémoires de la Société de Linguistique de Paris, nouvelle série, tome XII.
Meacham, C., and G. Estabrook. 1985. Compatibility methods in systematics. *Annual Review of Ecology and Systematics* 16:431-446.
Ogawa, Naoyoshi, and Erin Asai. 1935. *The myths and traditions of the Formosan native tribes*. Taipei.
Ostapirat, W. 2000. Proto-Kra. *Linguistics of the Tibeto-Burman Area* 23.1.

-------- in press. Kra-dai and austronesian: notes on phonological correspondences and vocabulary distribution. In *The peopling of East Asia: Putting together Archaeology, Linguistics and Genetics* ed. by L. Sagart, R. Blench and A. Sanchez-Mazas. London: RoutledgeCurzon.
Pecoraro, F. 1977. Essai de dictionnaire taroko-français. *Cahier d'Archipel 7*. Paris: S.E.C.M.I.
Peiros, I. 1998. *Comparative Linguistics in Southeast Asia*. Canberra: Pacific Linguistics C-142.
Reid, Lawrence A. 1971. *Philippine minor languages: Word lists and phonologies.* Oceanic Linguistics Special Publication, 8. Honolulu: University of Hawaii Press. xii
-------- 1982. The demise of Proto-Philippines. In *Papers from the Third International Conference on Austronesian Linguistics*, vol. 2: *Tracking the travellers*, ed. by A Halim, L. Carrington and S.A. Wurm, pp. 210-216. Canberra: Pacific Linguistics C-75.
Ross, Malcolm D. 1995. Some current issues in Austronesian Linguistics. In: Darrell T. Tryon. ed.) *Comparative Austronesian Dictionary, part 1, fascicle 1*: 45-120. Berlin, New York: Mouton de Gruyter.
Sagart, L. 1999. *The Roots of Old Chinese.* Current Issues in Linguistic Theory, 184. Amsterdam: John Benjamins.
-------- 2001. Comment: Malayo-Polynesian features in the An-related vocabulary in Kadai. Paper presented at the workshop "Perspectives on the Phylogeny of East Asian Languages", August 28-31, Périgueux.
-------- 2002a. Sino-Tibeto-Austronesian: an updated and improved argument. Paper presented at the 9th International Conference on Austronesian Linguistics, Canberra, January 8-11, 2002.
-------- 2002b. Gan, Hakka and the Formation of Chinese Dialects, in *Dialect Variations in Chinese*, ed. by Dah-an Ho, pp. 129-154. Taipei: Academia Sinica.
-------- 2003. The vocabulary of cereal cultivation and the phylogeny of East Asian languages. *Bulletin of the Indo-Pacific Prehistory Association* 23. Taipei papers, Volume 1:127-136.
-------- in press, a. Tai-Kadai as a subgroup of Austronesian. In *The peopling of East Asia: Putting together Archaeology, Linguistics and Genetics*, ed. by L. Sagart, R. Blench and A. Sanchez-Mazas. London: RoutledgeCurzon.
-------- in press, b. Sino-Tibetan-Austronesian: an updated and improved argument. In *The peopling of East Asia: Putting together Archaeology, Linguistics and Genetics*, ed. by L. Sagart, R. Blench and A. Sanchez-Mazas. London: RoutledgeCurzon.
Solheim, Wilhelm G. 1971. New Light on a Forgotten Past. *National Geographic Magazine*, 139,3:330-39.
Starosta, Stanley. 1985. Verbal inflection versus deverbal nominalization in PAn: the evidence from Tsou. In *Austronesian linguistics at the 15th Pacific Science Congress*, ed. by Andrew Pawley and Lois Carrington. Pacific Linguistics.
-------- 1994. Rukai-Tsouic: subgroup or treetop ? Paper presented at the 7th International Conference on Austronesian Linguistics. Leiden: Leiden University.
-------- 1995. A grammatical subgrouping of Formosan languages. In *Austronesian Studies Relating to Taiwan*, ed. by : Paul Jen-kuei Li, Dah-an Ho, Ying-kuei Huang, Cheng-hwa Tsang, and Chiu-yu Tseng, pp. 683-726. Symposium series of the Institute of History and Philology, Academia Sinica, no. 3. Taipei: Academia Sinica.

-------- 1996. The position of Saaroa in the grammatical subgrouping of Formosan languages. In *Pan-Asiatic Linguistics: Proceedings of the Fourth International Symposium on Languages and Linguistics, Volume III*, ed. by Suwilai Premsrirat, pp. 944-966. Salaya, Thailand: Institute of Language and Culture for Rural development, Mahidol University at Salaya,
-------- 2001. Reduplication and the subgrouping of Formosan languages. Paper presented at the International Symposium on Austronesian Cultures: Issues relating to Taiwan: Taipei, 8-12 December 2001.
Thompson, J. 1873. Notes of a journey in southern Formosa. *Journal of the Royal Geographical Society* XLIII:97-107.
Thurgood, G. 1994. Tai-Kadai and Austronesian: the nature of the historical relationship. *Oceanic Linguistics* 33:345-368.
Topping, D. 1973. *Chamorro reference grammar*. Honolulu: University of Hawaii Press.
Trudgill, P. 1974. *Sociolinguistics: an introduction*. Harmondsworth: Penguin.
Tsang, Cheng-hwa. 1995. New archaeological data from both sides of the Taiwan straits and their implications for the controversy about Austronesian origins and expansion. In *Austronesian Studies Relating to Taiwan*, ed. by Paul Jen-kuei Li, Dah-an Ho, Ying-kuei Huang, Cheng-hwa Tsang, and Chiu-yu Tseng, pp. 185-225. Symposium series of the Institute of History and Philology, Academia Sinica, no. 3. Taipei: Academia Sinica,
-------- in press. Recent discoveries at a Tapenkeng culture site in Taiwan: implications for the problem of Austronesian origins. In *The peopling of East Asia: Putting together Archaeology, Linguistics and Genetics* ed. by L. Sagart, R. Blench and A. Sanchez-Mazas. London: RoutledgeCurzon.
Tsuchida, S. 1976. *Reconstruction of Proto-Tsouic phonology*. Study of languages and cultures of Asia and Africa monograph series #5. Tokyo: Tokyo Daikokugo Daigaku.
Tsuchida, S. Yamada, Y. and T. Moriguchi. 1991. *Linguistic materials of the Formosan sinicized populations I: Siraya and Basai*. Tokyo: The University of Tokyo, Linguistics Department.
Tung, T.H. 1964. *A descriptive study of the Tsou language*. Nankang: Institute of History and Philology, Academia Sinica.
Wright, H. 2004. Early Austronesians and their neighbours on the East Asian mainland: the archaeological context. *Paper presented at the workshop on Human migrations in continental East Asia and Taiwan*, Geneva, June 10-13, 2004.
Zeitoun, E. 2000a. *Lukai yu cankao yufa*. Taipei: Yuan-liou.
-------- 2000b. *Bunong yu cankao yufa*. Taipei: Yuan-liou.
Zhang, Yongli. 2000. *Gemalan yu cankao yufa.* Taipei: Yuan-liou.

	pitu '7'	lima '5'	enem '6'	walu '8'	Siwa '9'	puluq '10'
Luilang	innai	(na)lup	(na)tsulup	patulunai	satulunai	isit
Saisiat	saivuseaha	rrasu	saivusa	makaspat	ra:ha	ranpon
Pazeh	xasebidusa	xasep	xasebuza	xasebaturu, xasebituru	xasebisupat	isit
Favorlang	naito	achab	nataap	maaspat	tannacho	zchiett
Taokas	yweto	hasap	tahap	mahalpat	tanasu	(ta)isid
Atayal	pituʔ	imagal	cziuʔ	spat	qeruʔ	lpuu
Sediq	pito	lima	mataro	maspat	maŋali	maxal
Thao	pitu	rima	ka-turu, makalh-turu-turu	kahspat, maka(lh)-shpa-shpat	tanacu	maqcin
Siraya	pǐttu	rima	nəm	kuixpa	matuda	saat kǐttian
Hoanya	pito	Lima	(mi)nun	(mi)alu	(a)sia	(miata)isi
Papora	pitu	nema	(ne)nom	mahal	(me)siya	(me)tsi
Tsou	pítu	eímo	nómə	vóeu	sío	máskə
Saaroa	(k)upito	(k)ulima	(k)ənəmə	(k)ualo	(k)usia	(ku)ma:ɬə
Kanabu	pitu	rima	nəm	(h)a:ru	si:ya	ma:nə
Bunun	pitu'	hima'	nuum	vau'	siva'	mas'an
Rukai	pitu	Lima	eneme	vaLu	baŋatə	maŋeale
Paiwan	pitju	lima	enem, unem	alu	siva	puluq
Puyuma	pitu	Lima	nem	waLu	iwa	puLu
Amis	pitu	lima	'enem	falu	siwa	polo
Kavalan	pitu	rima	'nem	waru	siwa	betin
Ketagalan	pitu	tsjima	anəm	wasu	siwa	labatan
PMP	*pitu	*lima	*enem	*walu	*siwa	*puluq

Table 1: implicational hierarchy of the numerals 5-10 in Formosan languages and in PMP. Gray cells: presence of the etymon. Sources: Amis: Wu (2000); Atayal: Egerod (1980); Bunun: Zeitoun (2000a); Favorlang=Babuza: Ferrell (1969); Hoanya: Ferrell (1969); Kanakanabu: Ogawa and Asai (1935), as cited in Ferrell (1969); Kavalan: Zhang (2000); Ketagalan-Basai: Yamada, Tsuchida and Moriguchi (1991); Luilang: Ferrell (1969); Paiwan: Ferrell (1982); Papora: Ferrell (1969); Pazeh: Li and Tsuchida (2001); Puyuma: Huang (2000); Ruka (Budai): Zeitoun (2000b); Saaroa: Tsuchida, as cited in Ferrell 1969; Saisiat: Ino (1998: Saitaoyak); Sediq: Pecoraro (1977), Siraya: Adelaar (1997); Taokas: Ferrell (1969); Thao: Blust (2003); Tsou: Tung (1964).

	0. start = PAn	1. schwa > i after -iC	2. pa > wa	3. delete remaining schwas	4. prune left of pretonic syllable	5. prune right of stressed vowel	6. tl > t
7 (5+2)	RaCep-i-t**u**Sa	RaCepit**u**Sa	RaCepit**u**Sa	RaC_pit**u**Sa	_pit**u**Sa	pit**u**_	pit**u**
8 (5+3)	RaCep-a-tel**u**	RaCepatel**u**	RaCe<u>w</u>atel**u**	RaC_wat_l**u**	_watl**u**	watl**u**	wa<u>l</u>**u**
9 (5+4)	RaCep-i-Sep**a**t	RaCepiS<u>i</u>p**a**t	RaCepiSi<u>w</u>**a**t	RaC_piSiw**a**t	_Siw**a**t	Siw**a**_	Siw**a**

Table 2: derivation of *pitu, *walu and *Siwa out of PAn analytic forms

	5+ additive	6+ additive	multiplicative	subtractive
six	5+1	___	2x3	no exx
seven	5+2	6+1	___	no exx
eight	5+3	no exx	2X4	no exx
nine	5+4	no exx	___	10−1

Table 3. The PAn numerals from 6 to 9: analytic forms in Formosan languages

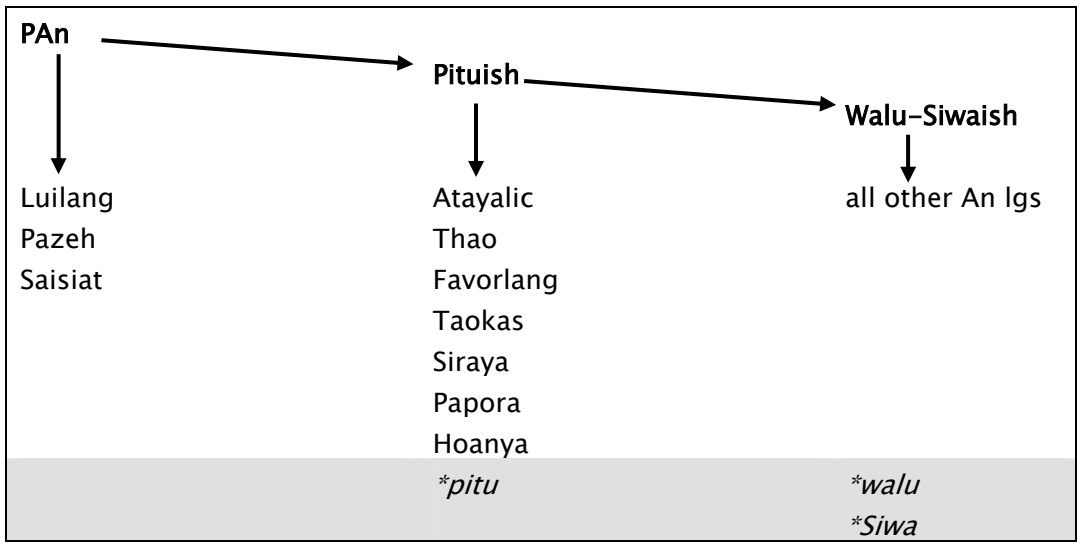

Figure 1. Higher An phylogeny based on three characters, with innovations (gray area) at each node

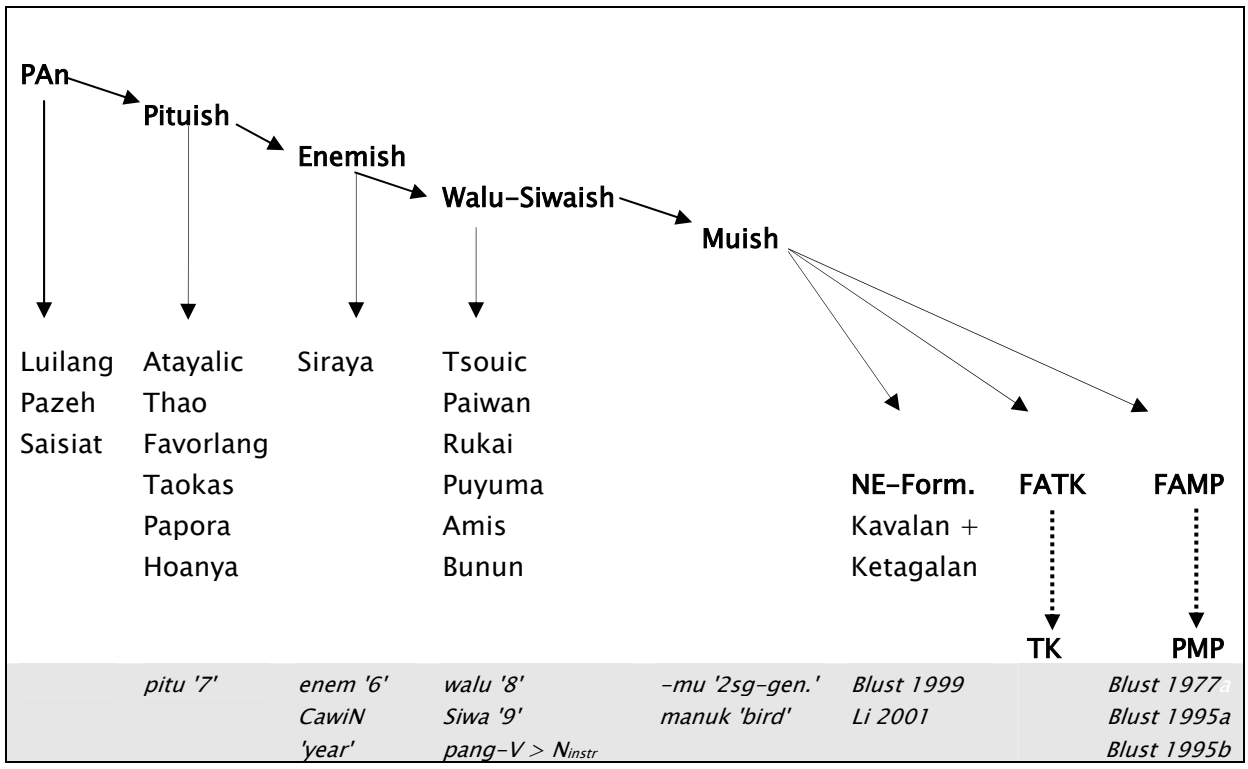

Table 4: Enriched higher An phylogeny based on basic-lexical and morphological innovations (gray area at bottom). Taxon names are in bold type. To the exception of the NE-Formosan languages Kavalan and Ketagalan, no claim is made that the languages whose names are in the same column (for instance Luilang, Pazeh and Saisiat) form a taxon. Dotted arrows indicate an oversea migration. 'FATK' = Formosan Ancestor of Tai-Kadai; 'FAMP' = 'Formosan Ancestor of Malayo-Polynesian'.

	PAn	PMP	Tai	Lakkia
die	maCay	matay	ta:i$_{1}$	plei$_{1}$
eye	maCa	mata	ta$_{1}$	pla$_{1}$
bird		manuk	nok$_{8}$	mlok$_{7}$

Table 5: regularity of sound correspondences in some An and TK basic vocabulary items

	Buyang	PAn	MP
die	ma^{0} tɛ^{54}	maCay	matay
eye	ma^{0} ta^{54}	maCa	mata
bird	ma^{0} nuk^{11}	qayam	manuk
8	ma^{0} ðu^{312}	_____	walu
head	qa^{0} ðu^{11}	quluh	quluh
louse	qa^{0} tu^{54}	kuCu	kutu
fart	qa^{0} tut^{54}	qetut	
raw	qa^{0} ʔdip^{54}	qudip	
bear (n.)	ta^{0} mɛ^{312}	Cumay	
cover (v.)	ta^{0} qup^{11}		WMP ta(ŋ)kup

Table 6. Disyllabic An words in Buyang

	Buyang	PAn	PMP
I	ku^{54}	-ku	-ku
thou	ma^{312}	-Su	-mu

Table 7. Personal pronouns of An and Buyang

	Buyang	PAn	PMP
2	ɕa^{54}	duSa	
3	tu^{54}	telu	telu
4	pa^{54}	Sepat	e(m)pat
5	ma^{312}	RaCep	lima
6	nam^{54}	_____	enem
7	tu^{312}	_____	pitu
8	ma^{0} ðu^{312}	_____	walu
9	va^{11}	_____	siwa
10	put^{54}	sa-iCit	puluq

Table 8. Numerals of An and Buyang

	PMP	P-Kra (Ostapirat 2000)
five	lima	r-ma A
six	enem	X-nəm A
seven	pitu	t-ru A / C-tu A[26]
eight	walu	m-ru A
nine	Siwa	s-ɣwa B
ten	puluq	pwlot D
2sg	-mu	mə A/B
bird	manuk	ɲok D

Table 9: Post-PAn lexical innovations and proto-Kra

[26] see fn. 25.

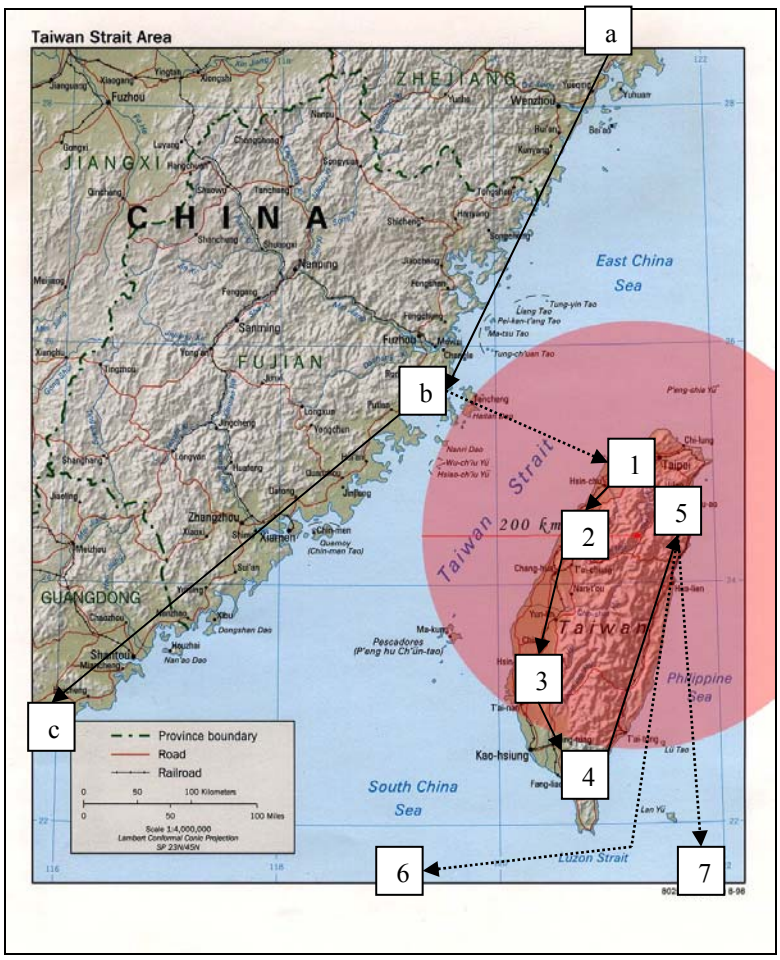

Figure 2. The An settlement of Taiwan with the MP and TK migrations

[a]: The pre-Austronesians, from NE China, expand southward along the SE China seaboard in the 5th and early 4th millennia BCE: they cultivate rice, foxtail millet, exploit marine resources, practice tooth evulsion. [b]: the Nanri and Pingtan islands, from which the top of Mt Xueshan (3884 m., at center of 200-km radius visibility circle) can be seen,[27] are reached. From there one group crosses to Taiwan c. 3500 BCE, while [c] the rest continues expanding in a SW direction towards the Pearl River Delta. [1]: location of earliest An (PAn-speaking) settlements on Taiwan. [2]: location of Pituish, [3] location of Enemish, [4] location of Walu-Siwaish, [5] location of Muish, [6] Tai-Kadai migration, [7] Malayo-Polynesian migration, c. 2000 BCE.

[27] I thank Christophe Coupé who in May 2004 calculated the visibility distance of the Xueshan using a formula used by the French Navy. That the top of the Xueshan is visible from these islands has since been confirmed to me by Prof. Tsang Cheng-hwa, from personal experience (p. c., June 2004).