The 14th International Symposium on Chinese Languages and Linguistics (IsCLL-14)

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

Robert Blust
University of Hawai'i

ABSTRACT. Sagart (2004) introduced a radical new view of the higher-level branching of Austronesian languages that has been challenged by others, but that he has continued to maintain and develop over the succeeding decade. The purpose of this paper is to show that his numeral-based phylogeny is built on straightforward errors of fact, and that once corrections are made in this flawed empirical foundation his attempt to derive the consensus PAN numerals *pitu '7', *walu '8', and *Siwa '9' from earlier compound forms based on *RaCep 'five' requires not the 'mere' six changes that he claims, but rather 12 changes, 10 of which are found in a single form, and thus justify the label *ad hoc* which has been directed at his argument by others. The implicational hierarchy defining numeral systems across Formosan languages presents a greater challenge, but can be seen as a consequence of descending text frequency with ascending magnitude, resulting in lower stability as one moves from 1-10. Even if no viable alternative were found for the implicational hierarchy in Formosan numeral systems, however, the argument fails unless Tai-Kadai is accepted as a member of the AN family, a claim that raises far more serious questions than it answers.

1. A new Austronesian family tree. Sagart (2004) has proposed a novel and strikingly different higher-level subgrouping of the AN languages based on a reinterpretation of the traditional reconstruction of the numerals 1-10. While a PAN full decimal counting system *esa/isa '1', *duSa '2', *telu '3', *Sepat '4', *lima '5', *enem '6', *pitu '7', *walu '8', *Siwa '9', *puluq '10' has been posited in the past and accepted by virtually all AN comparativists, Sagart assumes that PAN had an imperfect decimal system of the form 1, 2, 3, 4, 5, 5+1, 5+2, 5+3, 5+4, 10, in which '5' was *RaCep, '6' is unknown, '7' through '9' were *RaCep-i-tuSa, *RaCep-a-telu, *RaCep-i-Sepat, and '10 was *sa-iCit, a system that has been preserved almost intact by Pazeh, as shown in Table 1, where PAN-1 = consensus reconstruction, and PAN-2 = Sagart's proposed revision:

TABLE 1: PAZEH AS THE BASIS FOR THE NEW AUSTRONESIAN FAMILY TREE

PAN-1	PAN-2	Pazeh

[1] I am indebted to Yen-ling Chen and Paul Jen-kuei Li for assistance with references, to Amy Schafer and Kamil Deen for directing me to sources on the text frequency of English numerals, and to Yuko Otsuka for help with Japanese publications. Victoria Yen-hsin Chen wanted me to write this paper so badly that I finally did (thanks, Victoria). None of them is responsible for any errors in my interpretation or citation of data, nor do they necessarily accept my conclusions.

*esa/isa	*esa/isa	ida	one
*duSa	*duSa/tuSa	dusa	two
*telu	*telu	turu	three
*Sepat	*Sepat	supat	four
*lima	*RaCep	xasəp	five
*enem	?	xasəb uza	six
*pitu	*RaCep-i-tuSa	xasəb-i-dusa	seven
*walu	*RaCep-a-telu	xasəb-a-turu ~ xasəb-i-turu	eight
*Siwa	*RaCep-i-Sepat	xasəb-i-supat	nine
*puluq	*sa-iCit	isit	ten

To derive the PAN-1 numerals *pitu, *walu, *Siwa from their hypothesized compound progenitors Sagart appeals to a general ‘drive to disyllabism’ of the type described in Blust (2007), and six putative sound changes, as shown in table 2:

TABLE 2: THE DERIVATION OF *pitu, *walu, *Siwa FROM EARLIER COMPOUND NUMERALS BUILT ON PAN *RaCep ‘five’ (after Sagart 2004)

	7 (5+2)	8 (5+3)	9 (5+4)
PAN	*RaCep-i-tuSa	*RaCep-a-telu	*RaCep-i-Sepat
(1)	RaCepit**u**Sa	RaCepatelu	RaCepiSipat
(2)	RaCepit**u**Sa	RaCewatelu	RaCepiSiwat
(3)	RaC_pIt**u**Sa	RaC_wat_lu	RaC_piSiwat
(4)	_pituSa	_watlu	_Siwat
(5)	pitu_	watlu	Siwa_
(6)	pitu	walu	Siwa

The parenthesized numbers in the column under ‘PAN’ represent what Sagart (to appear) conceives as stages in the historical derivation of the traditional PAN numerals 7-9. The starting point is PAN, prior to any change giving rise to innovative numerals. The succeeding stages are described as follows:

FIGURE 1: The six ‘arbitrary’ but ‘natural’ sound changes assumed by Sagart (2004) to derive *pitu, *walu, and *Siwa from compound base-5 progenitors

Stage (1) : schwa (*e) > i after –iC
Stage (2) : pa > wa
Stage (3) : delete remaining schwas
Stage (4) : prune left of pretonic syllable
Stage (5) : prune right of stressed vowel

Stage (6) : tl > t

In support of changes 1-6, Sagart (2004:418) holds that "The changes in table 2 (i.e. Fig. 4) 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" that was at work throughout early AN history. Such reductions 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."

While the structure of the Pazeh numeral system may have inspired Sagart's novel ideas about AN subgrouping, he is too sophisticated to have used it without supporting evidence of other kinds. The most critical of these is an intriguing observation that he summarizes as follows (Sagart to appear):

FIGURE 2: IMPLICATIONAL RELATIONSHIPS IN THE APPEARANCE OF WORDS FOR 5-10 IN THE FORMOSAN LANGUAGES (after Sagart 2004, to appear)

*puluq '10' << *Siwa '9' << *walu '8' << *enem '6' << *lima '5' << *pitu '7'
(where A << B means 'a reflex of A implies the presence of a reflex of B')

In other words, AN languages that have a reflex of *puluq '10' also have a reflex of all numerals to the right, and so on as one moves through the implicational string. Sagart interprets this as evidence for sequential innovation, and he attempts to anchor this interpretation in the established knowledge of other academic disciplines by reminding the reader that the 'nesting' of innovative features is found in biological species as well as languages: "Similarly among biological species the presence of hair implies amniotic eggs, which imply four limbs, which imply a bony skeleton, which implies vertebrae."

Just as these implicational relationships in biology mirror the evolutionary history of the animal kingdom, then, with the appearance of endoskeletons preceding the appearance of four limbs from ancestral fins, and a life on land preceding the appearance of eggs that are hatched within the mother's body rather than external to it, implicational relationships in the appearance of numeral forms in the Formosan languages mirror the branching order of the AN family tree: languages with PAN *esa/isa 'one', *duSa 'two', *telu 'three', *Sepat 'four', *RaCep 'five' first added *pitu '7', then replaced *RaCeb with a reflex of *lima 'hand', and then successively added *enem 'six', *walu 'eight' and *Siwa 'nine' (no ordering implied for these two) and finally *puluq 'ten'. In keeping with the type of evidence to which he appeals for this subgrouping hypothesis, Sagart (2004:431) labeled all but one of the nodes in his family tree after the proposed numeral innovations

that define them, as shown in Fig. 3, where the tree structure is preserved, but converted to linear form (Northeast Formosan = Kavalan and Ketagalan/Basai, FATK = Formosan ancestor of Tai-Kadai, and FAMP = Formosan ancestor of Malayo-Polynesian):

FIGURE 3: THE AUSTRONESIAN FAMILY TREE
ACCORDING TO SAGART (2004)

PAN > 1. Luilang, 2. Pazeh, 3. Saisiyat, 4. Pituish
Pituish > 1. Atayalic, 2. Thao, 3. Favorlang, 4. Taokas, 5. Papora, 6. Hoanya, 7. Enemish
Enemish > 1. Siraya, 2. Walu-Siwaish
Walu-Siwaish > 1. Tsouic, 2. Paiwan, 3. Rukai, 4. Puyuma, 5. Amis, 6. Bunun, 7. Muish
Muish > 1. NE-Form., 2. FATK, 3. FAMP

More recently Sagart (2008, to appear) has modified this tree by adding a 'Limaish' node between Pituish and Enemish, by changing the name of the Muish node to 'Puluqish', and by changing the assignment of some languages. In particular, Atayalic and Thao have been reassigned to Limaish, Papora and Hoanya to Walu-Siwaish, Paiwan, Puyuma and Amis to Puluqish, and Northeast Formosan (Kavalan and Ketagalan/Basai) to Walu-Siwaish, as shown in Fig. 4:

FIGURE 4: THE AUSTRONESIAN FAMILY TREE
ACCORDING TO SAGART (2008, to appear)

PAN > 1. Luilang, 2. Pazeh, 3. Saisiyat, 4. Pituish
Pituish > 1. Favorlang/Babuza, 2. Taokas, 3. Limaish
Limaish > 1. Atayalic, 2. Thao, 3. Enemish
Enemish > 1. Siraya, 2. Walu-Siwaish
Walu-Siwaish > 1. Hoanya, 2. Papora, 3. Tsouic (Tsou, Kanakanabu, Saaroa), 4. Rukai, 5. Bunun, 6. Kavalan, 7. Ketagalan, 8. Puluqish
Puluqish > 1. Paiwan, 2. Puyuma, 3. Amis, 4. Tai-Kadai, 5. PMP

There is a certain *a priori* plausibility in the conclusion that Sagart reaches about the AN settlement of Taiwan, namely that first landfall was in the northwest, nearest the Chinese mainland from which the island was clearly reached. One cannot help but notice that the three primary subgroups that he recognizes as containing one language each (Luilang, Pazeh, Saisiyat) are located in the north or northwest portion of the island, and that his Puluqish group occupies the east coast and its southern extension, where the poorest agricultural lands are found, and which consequently was the last part of the island to be settled during the later immigration of the Han Chinese. At first glance, then, there is nothing inherently implausible about these claims.

However, in studying this tree serious problems quickly become apparent. The first of these is that Sagart's tree conflicts with other proposals that are more in keeping with mainstream application of the comparative method. Before considering conflicting subgrouping proposals, however, it should be noted that HPAN has so far been the subject of criticism from two sources, Winter (2010), and Ross (2012).

1.1. Winter (2010) and Ross (2012). Winter raises the following objections: 1) the putative sound changes that Sagart uses to derive *pitu, *walu and *Siwa from earlier compound numerals are *ad hoc* (283), 2) the 'drive to disyllabism' that Sagart invokes to explain the reduction of five-syllable forms to disyllables is well-attested with base morphemes, but not with morphologically complex forms (283), 3) a reanalysis of the numeral system starting with an innovative monomorphemic numeral for '7', while complex forms are still maintained for '6', '8', and '9', is not motivated, and would be typologically bizarre (283), 4) the general similarities between the consensus PAN numerals *pitu, *walu and *Siwa, and the underlined sequences in Pazeh xasəp-i-dusa, xasəp-a-turu, and xasəp-i-supat is most likely due to chance (283-84), 5) the claim that six is a "small number" of special sound changes needed to derive three disyllabic numerals from hypothetical five-syllable compound proto-forms glosses over the fact that the drive to disyllabism is usually achieved by single changes (284), 6) since numerals are borrowed in many languages, borrowing cannot be ruled out as an explanation for the distribution of forms attributed to PAN *RaCep (284-85), 7) Sagart proposes a counterclockwise migration pattern following an initial settlement on the northwest coast, and one would expect the nesting of languages in his tree to match the geographical distribution, but there are cases where it doesn't, as with Taokas, which was traditionally spoken to the north of Pazeh.

In fairness to Sagart, some of Winter's criticisms seem misplaced. With respect to point 4), for example, Winter notes (283-84) that the 'pre-Pituish languages' Luilang, Saisiyat and Pazeh have 15 forms for the numerals 5-10 with 55 syllables, which gives a wide range of choices for the six syllables of *pitu, *walu and *Siwa, making the similarity Sagart has found between these forms and the underlined portions of Pazeh *xasəb-i-dusa*, *xasəb-a-turu*, and *xasəb-i-supat* less remarkable than he claims. But Sagart has not selected *pi, *tu, *wa, *lu, and *Si randomly from the total collection of syllables in the Luilang, Saisiyat and Pazeh words for 5-10. Rather, he has shown that, given certain assumptions about irregular sound change, the consensus PAN numerals *pitu, *walu, and *Siwa are derivable from compound numerals inferred to be ancestral to the Pazeh words for 7-9. Similarly, geographical distributions can change over time, so the fact that

Taokas was traditionally spoken north of Pazeh during the narrow window of time that recordings have been made is insignificant.[2]

However, other criticisms are more telling, and cannot be dismissed so easily. The condemnatory use of 'ad hoc' is found repeatedly in Winter (2010). Moreover, as both Winter (2010) and Ross (2012) point out, Proto-Pituish would be typologically bizarre in having an innovative monomorphemic numeral for '7', while complex forms are still maintained for '6', '8', and '9'. Perhaps more to the point, there is no obvious reason why a PAN system of numerals for 6-9 that closely resembles the known Pazeh system (with 5+1, 5+2, 5+3 and 5+4) would reanalyze the hypothetical compound numerals 7-9 as monomorphemic disyllables, while leaving '6' unaffected, and *enem with no known etymology. Finally, Sagart's apparent satisfaction that he is able to achieve his proposed derivations with "only" six otherwise unfamiliar 'sound changes' is unlikely to convince most historical linguists: this is a large number of changes for deriving three forms and, as will be shown, it is not sufficient to do the job without significant supplementation.[3]

Ross (2012) raises many of the same points as Winter, but also lays particular emphasis on the questionable nature of the Tsouic subgroup first proposed by Ferrell (1969), and subsequently defended by Tsuchida (1976), as this conflicts with his own proposal of a Nuclear Austronesian subgroup which includes all AN languages except Tsou, Rukai and Puyuma. In my view Sagart (2013b, to appear) has successfully defended the integrity of the Tsouic subgroup, which is problematic for the Nuclear Austronesian hypothesis (Ross 2009, 2012), but is essentially neutral with respect to his own arguments.

1.2. 'Arbitrary' vs. 'ad hoc'. Sagart (2004:418) is quick to acknowledge that most of the sound changes he invokes to make his derivations work are 'arbitrary', meaning that they are not motivated by generally applicable phonological processes. However, he objects to the use of the term *ad hoc*, which is commonly used by philosophers of science to refer to theoretical constructs that are posited to dispose of a single observational anomaly without convergent confirmation from other sources (Leplin 1974-75). In his view, although they are arbitrary, changes 1-6 (Fig. 4)

---

[2] Despite it's value in presenting an alternative point of view, Winter's paper, unfortunately, also contains several errors, as where Luilang *patulunai* is given for both '8' and '9', whereas the latter should read *satulunai*, and where the Saisiyat word for '5' is given as *rrasu*, when Ferrell (1969) lists '*asəb*, and Li (1978:196) gives Tungho dialect *asəb*, Taai dialect *Lasəb*. The source of *rrasu* remains unclear.

[3] Sagart (2013:250) has responded to this criticism by claiming, *inter alia*, that informal Rumanian *paişpe* '14', *cinşpe* '15', and *saişpe* '16' derive from *patrusprezece*, *cincisprezece*, *şasesprezece* through seven changes. However, six of these are deletions which may have affected long strings at once, and in any case do not involve the transformation of one segment into another through processes that are otherwise unknown in the language. Moreover, it can plausibly be argued that the replacement of *-prezece* by *–şpe* was a single change --- a type of interpretation that cannot be applied to the AN data Sagart considers.

- 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;[4]
- do not affect the points of articulation of the consonants"

And again (2004:419) he says "I am well aware that the sound changes I am proposing lack the support of recurring sound correspondences. That is unavoidable, as the drive to disyllabism could only have affected a small number of expressions simultaneously. What table 2 [i.e. changes 1-6, RAB] 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."

It is true that even though it is sporadic, change 1 is not likely to raise many eyebrows, since it is the sort of irregular assimilation that happens frequently in natural language data. Similarly, no one is likely to quarrel with Sagart's interstage *tl > *l* as a plausible reduction, since heterorganic consonant clusters are not found in PAN, and if they arose through syncope there would be strong structural pressure to reduce them. However, *pa > *wa*, reportedly in two forms but nowhere else, certainly raises questions about the manipulation of data with a definite end in sight, and like many of the proposed 'sound changes' in Benedict (1975), *tl > *l* is unique, and so cannot possibly be confronted with counter-evidence. As we will see with gathering intensity in examing the details of Sagart's phylogeny, the issue of *ad hoc* 'sound changes' will not go away. Rather, it grows increasingly more serious as the argument becomes empirically better grounded and typologically more plausible.

1.3. Conflict with Western Plains. A Western Plains (WP) branch of Formosan languages was proposed by Tsuchida (1982:9ff), to include Taokas, Babuza, Papora, and Hoanya, all of which are known exclusively from colonial Japanese records that were made shortly before they were engulfed by the rising tide of sinicization among the lowland tribes.[5] Blust (1996) added Thao to this group, and the general outlines of the extended group were accepted by Li (2001).

---

[4] This statement reflects an inadequate appreciation of sound change in AN languages. Among many AN languages that show extensive vocalic change we can mention Tboli and other Bilic languages of the southern Philippines, Sa'ban of northern Sarawak, the Melanau languages of coastal Sarawak, the Chamic languages of mainland Southeast Asia, Rejang of southern Sumatra, Javanese, Gorontalo of northern Sulawesi, Palauan, and nearly all of the Oceanic languages of Micronesia, and much of Vanuatu.

[5] It is generally agreed that Babuza was a later continuation of the language that 17th century Dutch missionaries called 'Favorlang', for which fuller records are available (summarized in Ogawa 2003).

All three writers agree that WP is defined by at least two shared mergers: 1. merger of PAN *s and *t, found only in Hoanya, Babuza, Papora, Taokas, and Thao, and 2. merger of PAN *n and *ŋ, found in all of these languages except Hoanya, where dialect forms are given with both *n* and *ŋ* (Tsuchida 1982:45, 134), and in Kulon (Tsuchida 1985), which Sagart does not classify. Sagart (2004:429) discusses only the second of these mergers, and concludes that "there is nothing in the present phylogeny to contradict Blust's Western Plains group." Reference to Fig. 6 shows that this was true in the original phylogeny, where all five languages were assigned to Pituish, but in improving this proposal Sagart (2008, to appear) reassigned these languages to three different branches: Babuza and Taokas to Pituish, Thao to Limaish, and Hoanya and Papora to Walu-Siwaish (Fig. 7). The claim that WP is not in conflict with the numeral-based higher phylogeny of AN languages is thus no longer true.

Even more strikingly, the words for '8' (PWP *maka-Sepat, found in Taokas *maka-ipat*, Babuza *maa-spat*, and Thao *makalh-shpa-shpat*), and '9' (PWP *tanaCu, found in Taokas *tanaso* and Thao *tanacu*) are numeral innovations cutting across Pituish and Limaish. Sagart (2004:415) discusses these in passing without noting the conflict they present to his phylogeny because at that time the conflict did not yet exist. Now that it does, however, he is conspicuously silent about the matter (Sagart to appear). Add to this the innovations *iCu replacing PAN *ia '3sg' in both Pituish and Limaish languages (Taokas *iso*, Babuza *icho*, Thao *c-icu*), *taun replacing PAN *Rumaq 'house' in Pituish, Limaish, and probably Walu-Siwaish languages (Taokas *yadaun*, *don*, Babuza *tun*, *ton*, Thao *taun*, Hoanya *tuhun*, *tohun*), and *NaSuq replacing PAN *beRas 'husked rice in Babuza and Thao, and a serious problem with the numeral-based classification becomes apparent even before we consider the data upon which it is based.

1.4. Conflict with East Formosan. The next problem in Sagart'classification is the separation of Siraya under Enemish, from Kavalan/Ketagalan (= Kavalan/Basai in current terminology) under Walu-Siwaish, from Amis under *Puluqish. All of these languages are members of the East Formosan subgroup proposed by Blust (1999a), and accepted in broad outline by Li (2004). The most significant piece of evidence for this group, and the one that connects all of its members is the merger of PAN *j and *n. In many cases a single merger would be considered inadequate to justify a subgrouping claim, but this merger has far greater significance, since it is found in no other AN language. Given the numeral-based phylogeny this unique innovation becomes a product of three parallel changes, one affecting Siraya, another affecting Kavalan/Basai, and a third affecting Amis, leaving us to ask why a phonetically peculiar innovation would happen three times in a geographically constricted area, but nowhere else in a language family that contains over 1,200 member languages.

Sagart (2004:429) acknowledges that the merger of *j with *n took place in languages that "cannot be a monophyletic taxon in my phylogeny," and he therefore considers the

possibility that it could have been contact-induced. Because he finds this unconvincing, but needs to dismiss the evidence for East Formosan to salvage his subgrouping theory, he proposes an imaginative escape route: "Another 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 conservative zones: Amis and Kavalan-Ketagalan on the east coast, and Siraya on the west, 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 that gives the appearance of a palatal stop merging with *n."

Like a number of Sagart's proposals this is certainly a novel claim, and in fact given the evidence for *j, it is an extraordinary claim. By implication, the stop and nasal inventory he proposes for PAN contains five voiceless stops *p, *t, *C, *k, *q, four voiced stops *b, *d, *z, *g, and at least five nasals: *m, *n, *N, *ñ and *ŋ. However, there are several problems with this proposal.

The first problem is one of plausible phoneme inventory: Sagart's suggestion that *j was a palatal nasal ignores the fact that a palatal nasal *ñ has been reconstructed based on a different set of sound correspondences, and although it has merged with *N in most Formosan languages, Tsuchida (1976:307) has claimed that *N and *ñ are distinguished by both Tsou and Saaroa. If this is accurate then Sagart's claim that *j was a palatal nasal implies that PAN had six nasals *m, *n, *N, *$ñ_1$, *$ñ_2$, and *ŋ. Needless to say, not only is a distinction between two palatal nasals phonetically unlikely (and unattested anywhere in the AN family), but an implied phoneme inventory with six nasals introduces a typological anomaly, namely a language in which the number of place features for nasals exceeds that for oral consonants (Ferguson 1966:57).

The second problem is how to reconcile Formosan and Malayo-Polynesian reflexes of *j and *ñ. PMP *j and *ñ are patently distinct, with radically different reflex profiles (Blust 2009:572, 576), and if one takes Sagart at his word it becomes necessary to assume that PMP *j derives from a PAN palatal nasal, while PMP *ñ reflects something else. The only PAN source that PMP *ñ could possibly have is *N, but it is well-known that PAN *N and *n merged as PMP *n, as shown in Figure 8 and scores of PAN/PMP pairs such as *Natad/natad 'cleared ground, yard', *Niwaŋ/niwaŋ 'slender, skinny', *Nuka/nuka 'wound', *aNak/anak 'child', *baNaR/banaR 'a vine, *Smilax* sp.', *daNum/danum 'fresh water', *paNaw/panaw 'go, walk away, depart', *paNij/panij 'wing', *puduN/pudun 'ball of thread', *qaNiC/qanit 'animal skin, hide', *SiNuq/hinuq 'beads', *taNek/tanek 'to

cook', *tuNa/tuna 'freshwater eel', *bulaN/bulan 'moon, month', *quzaN/quzan 'rain', or *tiaN/tian 'abdomen' (Blust and Trussel, ongoing). In short, Sagart's proposal that PAN *j was a palatal nasal violates the comparative method in failing to account for fundamental sound correspondences.

FIGURE 5: PAN *j IN RELATION TO THE NASALS *n, *N and *ñ IN PAN AND PMP
(PAN-1 = standard reconstruction, PAN-2 = Sagart's alternative)

PAN-1	*j	*n	*N	*ñ
PAN-2	*ñ	*n	*N	??
PMP	*j	*n	*n	*ñ

Entirely apart from the two problems already seen (one of duplicate proto-nasals for radically different sound correspondences, and a second for leaving the origin of PMP *ñ in limbo), Sagart's rather desperate attempt to account for the merger of PAN *j and *n by regarding *j as a palatal nasal creates a number of phonetic problems. Following the nodes in his own most recent tree, it does not take a great deal of research to see that PAN *j is reflected as Pazeh, Saisiyat *z* under the PAN node, as Babuza *d*, Taokas *t* under the Pituish node, as *g* or zero in Atayalic, and *ð* in Thao under the Limaish node, as *n* in Siraya under the Enemish node, as Hoanya *dz*, Papora *d*, Tsou zero, Kanakanabu *l* or zero, as Saaroa *ɬ* (voiceless lateral), Rukai *g* or zero, Bunun zero and Kavalan *n* under the Walu-Siwaish node, and finally as Paiwan *d*, Puyuma *d*, Amis *n* under the Puluqish node before it finally became [g^y] in PMP. All of this is presented in Blust (1999a:43), which Sagart has surely seen. For the convenience of the reader this pattern of reflexes is summarized in Fig. 6:

FIGURE 6: REFLEXES OF PAN *j IN FORMOSAN LANGUAGES
(PAN-1 = standard phonetic interpretation, PAN-2 = Sagart's phonetic interpretation)

PAN-1	*j [g^y]	
PAN-2	*ñ [ɲ]	
	z	(Pazeh, Saisiyat)
Pituish	d	(Babuza)
	t	(Taokas)
Limaish	g, Ø	(Atayalic)
	ð	(Thao)
Enemish	n	(Siraya)
Walu-Siwaish	dz	(Hoanya)
	d	(Papora)
	Ø	(Tsou)
	l,Ø	(Kanakanabu)
	ɬ	(Saaroa)
	g, Ø	(Rukai)

	Ø	(Bunun)
	n	(Kavalan)
Puluqish	d	(Paiwan, Puyuma)
	n	(Amis)

By any reckoning it would be unexpected for a palatal nasal to produce the reflexes Sagart wishes to derive from it in Formosan languages, namely *z, d, t, g*, zero, *ð, n, dz, l*, and *ɬ*, with six of the ten reflexes being obstruents, including a voiced velar stop which, given the structure of Sagart's tree, implies *ñ > *g* independently in both Atayalic and Rukai. Not only is this phonetically unconvincing, but for purposes of comparison, consider the relatively uniform reflexes of PMP *ñ in a far larger sample of languages: *ñ, n*, and *y*, the latter only in historically secondary word-final position. And to round out the picture, reflexes of PMP *j, which Sagart is willing to acknowledge as a palatalized velar stop, are very similar to those in Formosan languages, with widely separated languages reflecting *j as *g* (Atta, Ilokano and other languages of northern Luzon, the Batak languages and Rejang of Sumatra), and others reflecting it as *d* (Ivatan of the Batanes Islands, Kelabit of northern Sarawak, Malay, the Sangiric languages of northern Sulawesi), as *s* (Buginese, nearly all Oceanic languages), or various other less common developments (Blust 2009:572).

Sagart's efforts to trivialize the evidence for East Formosan, a subgroup that conflicts with his numeral-based phylogeny, clearly violate basic principles of the comparative method, namely 1. that the same proto-phoneme (*ñ) cannot represent more than one well-attested set of sound correspondences in the same environment, 2. that all languages in the same family must be interrelatable by a single set of sound correspondences (not one set for Formosan and another for Malayo-Polynesian languages), and 3. that the phonetic value of a proto-phoneme must be able to account plausibly for the phonetic values of its reflexes. The merger of PAN *j and *n has great value for subgrouping precisely because in a language family with over 1,200 members it is unique to East Formosan languages, and efforts to make this appear otherwise through arbitrary assumptions about the phonetics of *j, or *ad hoc* sound changes from PAN [ɲ] to PMP [g^{y}] will strike most linguists as a desperate attempt to avoid facing counterevidence to one's claims rather than as a serious attempt to explain data.

1.5. The position of Tai-Kadai, I. We come now to the queen of Sagart's numeral-based phylogeny: the position of Tai-Kadai. The first claim that Tai-Kadai and AN are genetically related was made by Benedict (1942) in a seminal paper that has spawned decades of controversy. This is largely because after leaving the matter untouched for over 30 years Benedict (1975) returned to the problem, but with a very different mindset. While the first paper was restrained in its claims and tantalizing in the evidence it presented, the second paper abandoned the comparative method in the apparent hope of achieving victory by hyperbole. Large numbers of extremely speculative etymologies replaced the solid core with which Benedict began, and for many scholars what may well

be a valid hypothesis became obscured and devalued by what was essentially an exercise in academic hoodwinking (for a valuable critique, see Matisoff 1990).

Despite the excesses in Benedict's exuberant, but unconvincing approach to establishing genetic relationship, some scholars continued to see the value of the core etymologies, and the 'Austro-Tai hypothesis' as the claim had come to be called, attracted the attention of a wider circle of scholars. Thurgood (1994) agreed that Tai-Kadai and AN appear to have a historical connection, but he argued on the basis of alleged irregularities in tonal correspondneces that the observed similarities are due to borrowing. While this implied that AN languages had once been present in southern China, it did not establish a larger language phylum that includes both language families.

More recently a small but convincing body of evidence building on the original proposal of Benedict (1942) has been progressively strengthened, primarily through the work of Ostapirat (2005, 2013), who favors the name 'Kra-Dai' for this language family, and who has restored the Austro-Tai hypothesis to scientific credibility by addressing the data through a far stricter application of the comparative method. Moreover, Ostapirat (2005) has argued that the tonal correspondences Thurgood (1994) considered irregular are in fact regular, thus implying that the accepted historical relationship was one of common descent rather than contact. One of the key developments in making this possible has been the work of Chinese linguists in documenting more of the Kadai or Kra languages of southern China, of which Buyang has proven especially valuable. Sagart (2004) cites Buyang data to establish a genetic relationship of Tai-Kadai or Kra-Dai languages with AN, and this position was accepted as likely in Blust (2009:702-704).

Although more conservative historical linguists may still balk at accepting Austro-Tai , up to this point linguists such as Ostapirat or myself have no quarrel with Sagart. Where the two camps part company is with regard to the place of Tai-Kadai within the larger phylogenetic schema. To Ostapirat (2005, 2013) or most others who have commented on the matter, this relationship is a distant one, so distant as to barely remain capable of demonstration. However, as shown in Figures 3 and 4, Sagart sees Tai-Kadai is a primary branch of Puluqish. By the logic of his phylogeny, then, AN languages such as Paiwan, Puyuma, Amis, Tagalog or Malay are equidistant from one another and Tai-Kadai, and the same languages (including all descendants of PMP) are actually *more closely* related toTai-Kadai languages than they are to Formosan languages such as Kavalan, Bunun, Rukai, Siraya or Thao.

It takes little effort to show that languages such as Paiwan, Puyuma or Amis share several hundred cognates with MP languages such as Tagalog or Malay, or with other Formosan languages such as Kavalan, Bunun, Rukai or Thao (www.trussel2.com/ACD), whereas the most generous counts of probable cognates linking Tai-Kadai and AN are currently on the order of 25-30 solid comparisons (Ostapirat 2005, 2013). For Sagart's views on this matter to attract a serious following there must be a convincing explanation why Tai-

Kadai shares so few cognates with other members of the Austro-Tai superfamily, a matter first raised pointedly by Peiros (1998:103). Sagart's answer (2004:434) is worth quoting in full, as it illustrates a style he has shown repeatedly in dealing with counterevidence to his claims: "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 that has not left any other descendant (although macrophylic connections to Austroasiatic or Miao-Yao are a distinct possibility). In the course of that period of contact, a large part of the original vocabulary of TK was lost, with only the most basic part resisting relexification."

If ever there was a *deus ex machina* for annihilating counterevidence this is it: a theory creates expectations that conflict with observation, and the inconvenient observation is then eliminated with a magic wand. It goes without saying that the scientific credibility of any claim which appeals to evidence that is not accessible to observation is zero.[6]

2. The linguistic evidence. The conceptual basis for Sagart's tree begins with the observation that the first three languages provide no evidence for the traditional PAN numerals above 'four'. Rather than treating these languages as innovative Sagart assumes that *all other* languages have post-PAN innovations for 5-10, as seen in table 3, where again PAN-1 = standard reconstruction and PAN-2 = Sagart's alternative. For the sake of consistency the schwa is written with its IPA symbol in all attested languages; for Saisiyat the phonologically more conservative Taai dialect is chosen, and next to the standard form of Pazeh the Kahabu dialect (Ferrell 1968) is added:

TABLE 3: THE NUMERALS 1-10 IN THE NON-PITUISH AUSTRONESIAN LANGUAGES

PAN-1 '4'	*esa '1'	*duSa '2'	*telu '3'	*Sepat
PAN-2	*esa	*duSa/tuSa	*telu	*Sepat
Luilang	sa(ka) suva	tsusa	tuḷu	
Pazeh	ida	dusa	turu	supat
Pazeh K.	ʔadaŋ	dusaʔ	turuʔ	səpat
Saisiyat	ʔæhæʔ ʃəpat	roʃaʔ	toLoʔ	

[6]A more recent email (3/6/14) asking his current position elicited the following response: "Yes, I continue to think that the reason for the low cognate count between Kra-Dai and the other Puluqish languages is partial relexification from a local language, perhaps Austroasiatic or para-Austroasiatic, and later by Chinese. At the same time I think the cognate count between Kra-Dai and the rest of Puluqish is going to rise as the sound correspondences become clearer. Why do you ask ? did you hear someone say I had changed my mind on that ? that is not the case."

PAN-1	lima ‘5’	*enem ‘6’	*pitu ‘7’	*walu ‘8’
PAN-2	*RaCep	?	*RaCep-i-tuSa	*RaCep-a-telu
Luilang	(na)lup	(na)tsulup	innai	patulunai
Pazeh	xasəp	xasəb uza	xasəb-i-dusa	xasəb-i-turu
Pazeh K.	xasəb	xasəb uzaʔ	xasəb-i-dusaʔ	xasəb-i-turuʔ
Saisiyat	Lasəb	ʃayboʃiL	ʃayboʃiL-o-ʔæhæʔ	ka-ʃpat

PAN-1	*Siwa ‘9’	*puluq ‘10’
PAN-2	*RaCep-i-Sepat	*sa-iCid (?)
Luilang	satulunai	isit
Pazeh	xasəb-i-supat	isit
Pazeh K.	xasəb-i-səpat	ʔisid
Saisiyat	Lææʔhæʔ	laŋpəz

Except for the doublet for ‘two’, then, Sagart’s reconstruction of the PAN numerals 1-4 is identical to the traditional view. Where the two begin to diverge is with the number ‘5’, and that is the first point we need to consider. Sagart’s PAN *RaCep ‘five’ is based on cognate forms in Pazeh, Saisiyat, Favorlang and Taokas, and on the assumption that these languages belong to different primary subgroups of AN.

The first thing to note about this comparison is that standard Pazeh has undergone final devoicing, and shows intervocalic voicing of stops, making the voicing value for *xasəp* historically indeterminate (Blust 1999b:326-27). By contrast, both Pazeh Kahabu and Saisiyat maintain the voicing distinction in final stops, as seen in table 4:

TABLE 4: EVIDENCE FOR VOICING DISTINCTION OF FINAL STOPS IN PAZEH KAHABU AND SAISIYAT

PAN	Pazeh	Pazeh Kahabu	Saisiyat
*-b/d			
*debdeb	zəbəzəp	zəbəzəb ‘chest’	
*qaCeb			ʔæsəb ‘deadfall trap’
*qeNeb	aləp	ʔaləb ‘door’	ʔiləb ‘to close’ (root *-Neb ‘door’)
*Nihib			ka-lhib ‘rock shelter’
*CeRab			s<om>Lab ‘to belch’
	talup	tałub ‘juice’	
	bunat	bunad ‘sand’	
	lapət	malapəd ‘lightning’	
	laŋat	łaŋad ‘name’	
*lahud	rahut	rahud ‘downstream’	læhœr ‘downhill’

	rumut	rumud ‘flesh, meat’	
*qelud	urut	urud ‘pillar, post’	kæ-ʔlʸor ‘housepost’
*-p/t			
*layap			L<om>ayap ‘to fly’
*qaNup	alup	m-aLəp ‘to hunt’	ʔ<œm>alop ‘to hunt’
*quSeNap			ʔœʃalap ‘fish scale’
*qetut			ʔətut ‘to fart’
	barət	pa-barut ‘to answer’	
	kakamut	kakamut ‘finger’	
	makət	makət ‘muntjac deer’	
	maŋit	maŋit ‘weep’	
*Sepat	supat	səpat	ʃəpat ‘four’
	sidukut	si-dukut ‘small knife’	

For reasons that now seem transparent, Sagart (2004:416) uses the voicing ambiguity for Pazeh final stops to argue that *xasəp* reflects a PAN word that ended with *-p*. However, since both Pazeh Kahabu and Saisiyat preserve the PAN voicing distinction in final stops, and point to *RaCeb, simple application of the standard comparative method shows that Sagart’s *RaCep should be *RaCeb. Although it has no bearing on the argument, a parallel analysis holds for his *sa-iCit, which should instead be *sa-iCid, as shown by Pazeh Kahabu *ʔisid* and Taokas *ta-isid* ‘10’.[7]

Further data that raise questions about the history of Pazeh *xasəp*, Saisiyat *Lasəb* ‘five’ are *achab* ‘five’ in 17th century Favorlang, *hasip* in some dialects of late 19th century Hoanya (but *lima* in others; Tsuchida 1982), and *hasap ~ kasap* ‘five’ in late 19th century Taokas (Ferrell 1969, Li 2003). As noted by Li (1995b:139), none of these forms appears to be regular, since *R normally became Favorlang *g* (*kaRaC > *agach* ‘a bite’, *daRaq *tagga* ‘blood’, *kaRaŋ > *aggan* ‘crab’, *baRah > *bagga o chau* ‘ember’, *baRaq > *bagga* ‘lungs’, *huRaC > *oggach* ‘vein’), Taokas *l*, *x*, or possibly *h* (*kaRaŋ > *ka-kalaŋ* ‘crab’, *daRaq > *taxa ~ itahha* ‘blood’), and so far as the limited data permit us to see, *e did not normally become Hoanya *i* (*lipen > *ripun* ‘tooth’, *beNbeN > *bulbul* ‘banana’. Since the sound correspondences do not match, these forms are consequently best treated as unrelated, or as products of borrowing, possibly from Pazeh.

2.1. The derivations. We have now established that Sagart’s PAN *RaCep ‘five’ should be *RaCeb, and that his PAN *sa-iCit should be *sa-iCid, but it remains unclear whether

[7] Sagart (2004:416) tries to dodge the evidence for a final voiced stop in *RaCeb: “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.” But rules of final obstruent voicing are virtually unknown in the world’s languages (Blevins 2004), and as just shown here, the voicing distinction for final stops is *preserved* in both Pazeh Kahabu and Saisiyat (data for Favorlang are more limited, but *qudip > *orix* ‘alive’ suggests the same).

these forms have a PAN lineage, or are later innovations that spread by contact. The next logical step is to investigate the evidence for the numerals 6-9.

2.1.1. Six. By any reckoning the word for 'six' is an embarrassment for the numeral-based phylogeny. If PAN was like Pazeh, the numerals 6-9 should have had the structure 5+1, 5+2, 5+3 and 5+4. Sagart is happy to propose such compound numerals for 7-9, but is at a loss to do this for the number '6', which is simply left as a question mark (tables 16, 18). Yet, if *RaCep-i-tuSa evolved into *pitu we have every reason to expect that *RaCep-i-esa would have evolved into the number '6' which, given the 'stages' of Fig. 4, should be *pisa. Since *enem and *pisa share no similarity it is clear that the consensus PAN word for 'six' had some other, unknown source, and that the process of contracting five-syllable compound numerals to produce their better-known disyllabic descendants was itself unpredictable. Even while being forced to this conclusion, however, we are confronted by the observation that 'six' is *xasəb usa* (5+1) in both Pazeh and Pazeh Kahabu, a form that must have replaced *enem, raising further unanswered questions.

2.1.2. Seven. To strengthen his derivation of *pitu from a compound numeral based on PAN *RaCeb 'five', Sagart argues for the existence of a doublet *tuSa next to the well-established *duSa 'two', but there are at least two problems with this suggestion.

First, there is no good evidence for a variant *tuSa 'two'. Virtually all AN languages, including Luilang, Pazeh and Saisiyat, have regular reflexes of PAN *duSa, but Thao *tusha* and Amis *tosa* 'two' break this pattern. As noted elsewhere (Blust 1995c:450, 2003:80), the initial consonant in the Thao word for 'two' is almost certainly a product of what Bloomfield (1933:422-23) called 'contamination', and what Matisoff (1995) has called a 'run' in the onsets of successive numerals. Unlike most lexical items, numerals occur in sets governed by a strict sequential order that is followed when counting in the abstract (one, two, three, four, five.....). Many languages across a variety of families show a tendency to assimilate the onset consonants of adjacent numerals, either in anticipation or perserveration. Regular sound change in English, for example, should have produced *whour*, *five*, but the phonetic similarity of a voiceless bilabial glide or fricative and a voiceless labiodental fricative in adjacent numerals produced *four, five* to accommodate the rhythmic demands of rapid speech. In an earlier publication (Blust 1995) I have appealed to the same mechanism to explain why the initial consonant of PMP *siwa '9' invariably corresponds to a reflex of PAN *S in Formosan languages that have a cognate form.

In the case of Thao, the expected outcome of PAN *esa, *duSa, *telu was [ta], [ʃúʃa], [toró] (with '2' < [súʃa] by sibilant assimilation; cf. Blust 1995c). Rapid recitation of these numerals immediately shows the rhythmic disharmony of monosyllable-disyllable-disyllable, *t-ʃ-t*, and stress that shifts from paroxytone to oxytone in the number '3'. To remedy this problem Thao did three things: 1) it reduplicated [ta] to [táta], 2) it changed [ʃúʃa] to [túʃa], and 3) it changed [toró] to [tóro], producing a more harmonic result in

serial counting, from [ta, ʃúʃa, toró] to [táta, túʃa, tóro]. In short, Thao *tusha* does not provide reliable comparative evidence for PAN *tuSa.

The initial consonant of Amis *tosa* 'two' appears to owe its divergence from expectation to the same cause, namely that the word for 'two' was sandwiched between words for 'one' and 'three' that have voiceless alveolar obstruent onsets. Given the innovation *cacai ~ cecai* 'one' ([tsətsáiʔ] in my fieldnotes on Central Amis) expected **rosa (where /r/ is an alveolar flap) would have created an onset sequence *ts-r-t* with shared place features for all three segments, but shared manner features for only the peripheral ones. Given what we know about contamination in the history of other numeral systems, the articulatory disharmony of this sequence probably would have exerted an assimilatory pressure on the onset of '2'. Other Formosan languages which have a word for 'one' that begins with a voiceless alveolar obstruent are Tsou, Kanakanabu, Saaroa, Taokas, and Papora. However, Papora has an innovative word for 'two' (*nia*) without an obstruent onset, the onset of the Tsou word for 'two' (*ruso*) is described by Tsuchida (1976:86) as 'a voiced retroflex frictionless continuant as in English', and the regular reflex of PAN *d in both Kanakanabu and Saaroa is /c/, a 'voiceless palato-alveolar affricate' (Tsuchida 1976:27, 60). The Papora and Tsou words for 'two' thus lacks the phonetic properties that would favor contamination by appearing between *coni* 'one' and *turu* 'three', and the other two Tsouic languages would reflect PAN *duSa with *c-* in any case. Finally, Taokas *taanu, rua, turu* might appear to resist the type of contamination seen in Thao *tusha*, Amis *tosa*, but this depends on the phonetic properties of the Taokas rhotic, and since this word was recorded as *rua, dua, sua,* and *gua* by different researchers (Tsuchida 1982:34) it appears likely that its phonetic properties were quite variable, possibly including a uvular articulation. To summarize, the two Formosan languages that have a word for 'two' beginning with *t both have precisely the kind of precondition that one finds in many languages for contamination in the onsets of successive numerals, and for this reason it is very difficult to rule out this explanation for the irregularity in the initial consonant of this numeral.

Given these preferred alternatives, then, which involve processes of cross-morphemic assimilation known to affect sequential numerals in many of the world's languages, Sagart's derivation of *pitu should more properly be from *RaCeb-i-duSa. This raises the second problem for Sagart's hypothetical history of *pitu, since if we apply the six derivational 'rules' that he claims to be responsible for the reduction of earlier compound numerals to *pitu, *walu and *Siwa we get *bidu, not *pitu, and to remedy this situation we need additional derivational 'rules' devoicing *b and *d only in this word. Finally, even if Amis *tosa* is accepted as reflecting an earlier *tuSa that somehow figured in the history of *pitu we would still have no explanation for the initial *p, and a pre-Amis *tuSa, which is hardly sufficient by itself to infer a PAN doublet, would not be available in time for Proto-Pituish to use it in the proposed derivation.

2.1.3. Eight. Sagart derives *walu, the consensus PAN word for ‘eight’, from *RaCep-a-telu. We have already seen that *RaCep should be *RaCeb, and this changes Sagart’s compound numeral to *RaCeb-a-telu. Although the substitution of *RaCeb for *RaCep creates additional problems for Sagart’s derivation of *pitu ‘seven’, it has little effect on the derivation of *walu ‘eight’, which is problematic for other reasons.

The first problem with the derivation of *walu from *RaCeb-a-telu is in the proposed change *pa > *wa*, which Sagart (2004:418) finds a particularly compelling piece of evidence for his claims, since it is capable of explaining the origin of *w* in *both* *walu and *Siwa. However, given *RaCeb rather than *RaCep, the picture changes in ways that become far more damaging to him that one might imagine, since in deriving *walu from interstage *b-a-tlu, and *Siwa from interstage *Sipat, *pa > *wa* actually represents *two* irregular changes, *ba > *wa* and *pa > *wa* (or, better, *b > *w* and *p > *w*, since, as will be shown, the *b of *RaCeb-a-telu never preceded *a in its derivational history). Sagart’s claim (2004:418) that “The relatively marked lenition –pa- > -wa- affects two forms” is thus not only false, but doubly ironic, since to assert such an unusual context-dependent innovation for *both* *b and *p, each in a single morpheme, is an unmistakeable sign of an artificial analysis rather than a “natural change”.

The second problem with the derivation of *walu from *RaCeb-a-telu is the claim that the ligature in this word was *–a-*. Sagart bases this on the appearance of Pazeh variants *xasəb-i-dusa ~ xasəb-a-dusa* in Li and Tsuchida (2001, 2002), maintaining on the basis of its reported dominance in texts that the latter type is “the primary spoken form” and the former variant “its analogically leveled variant” (Sagart 2004:416). He adds that a variant with the numeral ligature *–a-* does not occur with any other Pazeh numeral built on ‘five’. However, when one inspects the texts in Li and Tsuchida (2002) this is not what one finds at all.

Sagart (2004:416) states that *xaseb-a-turu* occurs twice in Li and Tsuchida (2002), and that “a variant with the numeral ligature *–a-* does not occur with any other Pazeh numeral built on ‘five’.” However, an inspection of the entire book shows that these two examples are the only additive numerals based on ‘five’ in the texts and songs, and neither citation means ‘8’. Rather, both citations occur in a story that includes an event separated by a passage of 18 years. The first of these (page 49, sentence 14) reads:

adaŋ	a	dali,	adaŋ	a	dali,	maidəh	kutab-	a	daurik
one	Lig	day	one	Lig	day	soon	twinkling	Lig	eye

a	isiz-	a	xasəb	a	turu	a	kawas
Lig	ten	Lig	five	Lig	three	Lig	year

‘Day after day, it was soon eighteen (years) at the twinkling of eyes’ (which can perhaps be rendered more felicitously as ‘The days passed, and in the blinking of an eye it had been 18 years’).

The second example (page 53, sentence 35) repeats the terminal portion of the above, beginning with *isiz-*.[8] The first thing that matters here is that the number which Sagart reported as ‘8’ is in fact ‘18’. The second thing that matters is that the expression ‘18 years’ in these two citations is the only example of a compound numeral anywhere in Li and Tsuchida (2002), thus making it meaningless to say that a variant with the numeral ligature *–a-* does not occur with any other Pazeh numeral built on ‘five’. There simply are no examples of other compound numerals in this collection of texts and songs for comparison. The third thing that matters is that Ferrell (1968) gives Pazeh Kahabu ‘11’ (and, by implication, 12-19), as well as ‘20’, ‘30’ (and, by implication, 40-90), and these appear as: *ʔisid a ʔadaŋ* ‘11’ (and, by implication, *ʔisid a dusaʔ* ‘12’, *ʔisid a turuʔ* ‘13’, etc.), *dusaʔ isid* ‘20’, *turu a isid* ‘30’ (and, by implication, *supat a isid* ‘40’, *xasəb a isid* ‘50’, etc.). Excusing Ferrell’s inconsistent use of glottal stops, what emerges from this data is that *–a-* appears to be the *normal* ligature for numbers in the teens, and multiples of ten. The choice of *isiz a xasəb a turu* ‘18’, and the description of it as ‘8’ in Sagart’s argument is thus not only biased toward a predetermined conclusion, it is also an unfair misrepresentation of critically important data for readers who do not take the trouble to consult the primary source and do the necessary spadework themselves. In short, the facts are as follows: the ligature for ‘7’, ‘8’ and ‘9’ is *–i-*, while that for teens and multiples of ten is *–a-*, and Sagart’s *RaCep-a-telu must be revised to *RaCeb-i-telu.[9]

2.1.4. Nine. The last of the consensus PAN numerals that Sagart proposes to derive from an earlier compound based on ‘five’ is *RaCeb-i-Sepat > *Siwa ‘9’, and this is also problematic in multiple ways.

Sagart’s derivation of *Siwa essentially claims that the only part of *RaCeb-i-Sepat to survive is the original numeral ‘four’, but with the first vowel replaced by /i/ through sporadic assimilation. While a sporadic assimilation of this kind is certainly possible, it must be emphasized that it has no independent support. The change was not recurrent, as seen in PAN *bineSiq ‘seed rice’, PAN (Sagart’s Proto-Limaish) *likeS ‘mosquito’, or

---

[8] Note that this combining form for ‘10’ provides further evidence that Sagart’s *iCit ‘10’ should instead be *iCid.

[9] Li and Tsuchida (2002) do not help the matter, since their corpus contains no examples of numerals 11-19 other than ‘18’, yet in their dictionary (Li and Tsuchida 2001) they list ’11-19’ in their English finder list with the invariant form *isit* as the Pazeh equivalent (eleven : *isit*, twelve : *isit*, thirteen : *isit*, etc.), an inappropriate gloss which may indicate that they simply failed to elicit these numerals. To make matters even more confusing (and misleading), they give the Pazeh dictionary entry **xaseb a turu** (= **xasebituru** < **xasep** + **turu**) ‘eight’, and under ‘eight’ in the English finder list they give *xaseb a turu* (= *xasebituru*), *xasep*, which, based on their own textual examples, is clearly false.

PAN (Sagart's Proto-Walu-Siwaish) *qinep 'lie down to sleep', and the sole motivation for assuming it evidently is to remove some of the obstacles to deriving *Siwa from *RaCeb-i-Sepat. Allowing a sporadic sound change clearly increases the role of chance in relating two forms, but even if this is accepted there are other serious obstacles to the derivation of *Siwa from an earlier compound based on 'five'.

The second problem with this derivation is that it is made possible only by accepting another sporadic sound change, namely *pa- > *wa*-. As noted in the derivation of *walu '8', Sagart's original claim that *pa > *wa* happened in the derivations of both *walu and *Siwa no longer works, since *walu must now be derived from *RaCeb-i-telu, with an assumed *b > *w*, while *Siwa is derived from interstage *Sipat, where the hypothesized change must be *p > *w*.

2.2. The 'six stages' revisited. As noted by Winter (2010:284), Sagart (2004) appears pleased that he is able to derive *pitu, *walu, and *Siwa through changes that can be represented as happening at just six 'stages'. It has already been shown that one of these changes, *pa > *wa* in derivations for '8' and '9' must now be represented as two changes, *b > *w*, and *p > *w*, thus increasing this number to 7. However, it has also been shown that *RaCep must be reconstructed as *RaCeb, that a PAN variant *tuSa 'two' probably never existed, and that the ligature for all compound primary numerals was /i/. It is clear that each of these corrections will have implications for the complexity of the derivations, and these will be examined shortly. But first we need to consider the three 'sweeping' changes in Fig. 4: 'delete remaining schwas', prune left of pretonic syllable' and 'prune right of stressed vowel'. Each of these is problematic for various reasons.

3.2.1.5.1. Delete remaining schwas? Sagart's claim that "expressions of four or more syllables were reduced to disyllables" due to a preferred disyllabism for lexical bases overlooks the *qali/kali- forms in Formosan and Malayo-Polynesian languages which contain at least four syllables, and show no loss of schwa in stem forms, as in PAN *qaNi-meCaq, PMP *qali-metaq 'paddy leech', or PAN, PMP *qati-mela 'flea', and possibly *quNi-medaw 'dizzy' (Blust 2001:22-23, 28, 46). One might still suppose that given a recurrent disyllabism in the numerals 1-5 (table 18) the sudden departure from this pattern in the compound numerals would have exerted sufficient pressure to reduce these forms by foreshortening, even if this was not through the specific steps he proposes, although what he actually claims is more general (and false).

3.2.1.5.2. First pruning: prune left of pretonic syllable? The second 'sweeping' change that Sagart proposes (change 4) is one of two 'prunings', one to the left and the other to the right. Both of these proposed changes are stress-dependent, which requires an explicitly defended theory of stress placement in PAN. With regard to this matter Sagart (2004:418, fn. 9) says "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* '2 children', *dusa saw* '2 persons', *dusa ilas* '2 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',[10] *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)."

This is a surprising set of statements. First, there is no evidence of any kind to support the claim that PAN *duSa reflects earlier *duS plus the ligature *a. Not only is this a speculation lacking empirical support, but it is contrary to the almost complete absence of monosyllabic content morphemes in PAN, PMP or other AN proto-languages. Second, it is not true that "Pazeh /a/ is optional (indeed, mostly absent) between *dusa* '2' and a following noun." Unlike Sagart, who never had the opportunity to hear Pazeh spoken, I was fortunate to work with the last speaker, and have published based on that fieldwork (Blust 1999b). Among other expressions relevant to this claim in my fieldnotes are *dusa a isit* '20', *dusa a haten* '200', and *yaku dusa a rakihan* 'I have two children'.

In effect, then, Sagart assumes final stress as a general rule, but penultimate stress in the derivation of *pitu, since otherwise he would miss his target by an even wider margin than he already does. In view of this claim it is worth quoting Sagart in another context, where he defends his derivations from the criticisms of Winter (2010) that these 'work' only because he appeals to multiple *ad hoc* sound changes: "Under the comparative method a sound change can apply to a very limited number of words, if the context triggering it is uncommon. Part of the context conditioning the changes I proposed is a particularly fast speech rate due to pronouncing 6-10 on the same tempo as 1-5, all disyllables, in rhythmic counting." Yet what Sagart is claiming in his derivation of *pitu is that the PAN form of this numeral differed in stress from all others surrounding it, a clear violation of the prosodic conditions he appeals to in proposing his first pruning. Moreover, pruning left of the pretonic syllable of interstage *b-i-duSá yields *duSa, not *pitu, producing a situation that is clearly impossible in any language, namely identical numeral forms for '2' and '7'.

3.2.1.5.3. Second pruning: prune right of stressed vowel? We still have not gotten to the bottom of the problems that Sagart's proposals present. Putting aside the failed derivation of *pitu from an earlier compound numeral, we can assume final stress in the derivation of *walu and *Siwa from *RaCeb-a-telu and *RaCeb-i-Sepat respectively. However, Sagart then asks his readers to assume a second 'pruning', this one to the right of the stressed vowel.

---

[10] Apparently a typo in Li and Tsuchida (2002); cp. *supad a isit* in my own fieldnotes from the same speaker (Mrs. Pan Jin-yu).

Phonetically, one would expect 'prunings' not to affect full vowels, but rather to reflect a two-stage process in which full vowels are first reduced to schwa, which is extra-short, and then deleted. Sagart proposes his change 5 in order to drop the final syllable of interstage *b-i-dúSa, and he claims (2004:418) that his 'right-pruning' is not *ad hoc* since it serves a purpose in two different derivations, namely in *b-i-dúSa > *bidu (his *pitu), and in *Siwát > *Siwa. However, this is the same change in name only. I am unaware of a phonological process in any natural language that deletes a syllable in some forms and a coda in others. These are simply not the same type of change.

The loss of final segments, whether these are vowels or consonants, typically occurs in languages with non-final stress, and a common historical cycle is for final consonants to disappear, followed by the loss of final vowels, as in many of the languages of the Admiralty Islands, Micronesia and Vanuatu. The 'rule' that Sagart proposes makes no phonetic sense, and amounts to little more than a solution 'on paper'. Furthermore, it is obvious that this was not a general process, since the final consonant of *Sepat '4' was retained, even though the two words would have been virtually identical up to the last stage of Sagart's highly artificial derivation (*Sepat '4', *Siwat '9').

3.2.1.6. Living with the new reality. However he decides to proceed with his ideas about the numeral-based phylogeny, it is clear that Sagart has little choice but to live with the new reality. This means that he must recognize the errors in his proposal with regard to at least the following points:

1. The form assigned to PAN as *RaCep '5' was actually *RaCeb.
2. There is no reliable evidence for a PAN variant *tuSa '2'.
3. There is no empirical basis for assigning penultimate stress to *RaCep-i-túSa when all other primary numerals were oxytone; if such an aberrant prosodic feature ever existed it would soon be assimilated to the rhythmic requirements of rapid recitation in serial counting.
4. All compound primary numerals in Pazeh take the ligature *–i-*; the ligature *–a-* is reserved for additive and multiplicative derivatives of '10'.
5. 'Prune right of stressed vowel', which deletes a syllable in some words but a coda in others, is not a possible phonological rule, and therefore not a possible sound change.

It will be enlightening to see what we can do to salvage the numeral-based phylogeny in the face of these new obstacles. To do this the derivations of *pitu, *walu and *Siwa defended in Sagart (2004) and in all of his subsequent publications, are given in table 5, but with the conditions altered to take account of points 1-5 above. This will show how far the results of Sagart's derivations now are from their desired targets. Table 6 contains a new set of derivations that achieves the desired output, but at a far higher cost in terms of *ad hoc* assumptions.

Note that the ‘new reality’ reflected in table 5 admits four of the original six ‘arbitrary’ changes proposed by Sagart, disallowing only *p > *w*, which is now unique to the derivation of ‘9’, and ‘prune right of stressed vowel’, which is a graphic contrivance, not a possible sound change. The unique changes *e > *i*/iC_ in ‘9’ and *tl > *l* in ‘8’ are allowed on the grounds that sporadic assimilations are common, and if a secondary *tl cluster arose it would most likely be reduced to the liquid member, as in PAN *qiCeluR > *itlúg > Ibanag *illug*, Atta *illuk*, Gaddang, Yogad *i:lug* ‘egg’. Bolding indicates stressed vowels, and names of changes are abbreviated in some cases: *e > Ø = delete remaining schwas, SS in ‘7’ = stress shift to penult in ‘7’ (and presumably ‘2’), PL = prune left of pretonic syllable:

TABLE 5: SAGART’S DERIVATIONS OF PAN *pitu, *walu and *Siwa
IN LIGHT OF THE NEW REALITY, TAKE 1

PAN		RaCeb-i-es**a**	RaCeb-i-duS**a**	RaCeb-i-tel**u**	RaCeb-i-Sep**a**t
(1)	*e > i/iC_				RaCebiS̲ip**a**t
(2)	*e > Ø	RaC_bis**a**	RaC_biduS**a**	RaC_bitl**u**	RaC_biSip**a**t
(3)	PL	_bis**a**	_duS**a**	bitl**u**	Sip**a**t
(4)	*tl > l	bis**a**	duS**a**	bi̲l**u**	Sip**a**t
Proto-Puluqish		bis**a**	duS**a**	bi̲l**u**	Sip**a**t
PMP		bis**a**	duS**a**	bil**u**	sip**a**t
		‘6’	‘7’	‘8’	‘9’

As can be seen, with such seemingly minor changes as requiring a derivation for ‘6’ comparable to those for ‘7-9’, and accepting points 1-5 because of the need to correct earlier mistakes, the results fail miserably: desired *enem is **bisa; desired *pitu is **duSa, and hence identical to ‘2’, a clearly impossible situation; desired *walu is **bilu, with only the second syllable matching prediction; and desired *Siwa is **Sipat, which is dangerously close to *Sepat ‘4’.

How can we salvage Sagart’s original intention while remaining faithful to the ‘new reality’? Table 6 succeeds in doing just that --- not with his original six steps, which Winter (2010) already considered excessive in deriving just three numerals, but with 12. Moreover, 10 of these (all but nos. 3 and 6) are changes invented solely to account for a single form, while leaving others untouched, as where *b > *p* affects the derivation of ‘7’ but not ‘8’, or *i > *a*, which affects the derivation of ‘8’, but not ‘7’ or ‘9’. Finally, the derivation of *enem is left in limbo if the structure of the Pazeh numeral system is taken seriously as the model for PAN:

TABLE 6: SAGART’S DERIVATIONS OF PAN *pitu, *walu and *Siwa
IN LIGHT OF THE NEW REALITY, TAKE 2

PAN	RaCeb-i-es**a**	RaCeb-i-duS**a**	RaCeb-i-tel**u**	RaCeb-i-Sep**a**t
(1) *e > i/iC_				RaCebiSi̲p**a**t
(2) *p > w			RaCebitel**u**	RaCebiSi̲w**a**t
(3) *e > Ø	RaC_bis**a**	RaC_biduS**a**	RaC_bit_l**u**	RaC_biSiw**a**t
(4) SS in '7'		RaC_bid**u**Sa		
(6) PL	_bis**a**	_biduS**a**	_bitl**u**	_Siw**a**t
(6) –Sa > Ø		bid**u**_	bitl**u**	Siw**a**t
(7) –t > Ø				Siw**a**_
(8) *tl > l		bid**u**	bi̲l**u**	Siw**a**
(9) *b > p		pid**u**		
(10) *d > t		pit**u**		
(11) *i > a			bal**u**	
(12) *b > w			wal**u**	
Proto-Puluqish	(enem)	pit**u**	wal**u**	Siw**a**
PMP	(en**e**m)	pit**u**	wal**u**	siw**a**
	'6'	'7'	'8'	'9'

To summarize, Sagart's derivations of *pitu, *walu and *Siwa are products of an undeniably ingenious and resourceful argument, but one that is made possible only by misrepresenting primary data (cf. points 1-5), and by introducing multiple *ad hoc* hypotheses. In the light of a basic 'reality check' on the data, then, it is apparent that to derive the consensus PAN numerals 7-9 from earlier compound forms requires not six, but twelve 'sound changes', at least ten of which are posited to account for single forms.

3.2.2. The implications. In some ways one wonders why Sagart took such pains to provide morphological histories for *pitu, *walu, and *Siwa rather than simply allowing them to appear *ex nihilo*, as with *enem '6'. The reason seems to be that he dislikes the idea of lexical innovation as an independent process in language history. In any case, whether he allows *pitu, *walu and *Siwa to arise from pure invention (as must have happened at some point in the evolution of language) or derives them in the highly contentious way he has proposed, it seems clear that the strongest part of his argument is the embedding of forms for 5-10 in sets of languages that fit one within the other like Russian dolls. It is this feature that gives Sagart the greatest confidence of directionality, since in his view this type of embedding could only be a product of adding new forms, never of losing old ones.

However, there is another interpretation of the embedding issue that should be examined.

Consider the data in table 7, where ‘X’ marks a reflex of the traditional PAN numerals 1-10, and R marks reflexes of *RaCeb.[11]

TABLE 7: PATTERNING OF THE TRADITIONAL PAN NUMERALS 1-10 IN FORMOSAN LANGUAGES

	1	2	3	4	5	6	7	8	9	10
Proto-Atayalic		X	X	X	X		X			
Saisiyat	X	X	X	X	R					
Pazeh		X	X	X	R					
Thao	X	X	X	X	X		X			
Kavalan	X	X	X	X	X	X	X	X	X	
Amis		X	X	X	X	X	X	X	X	(X)
Bunun	X	X	X	X	X	X	X	X	X	
Tsou		X	X	X	X	X	X	X	X	
Kanakanabu		X	X	X	X	X	X	X	X	
Saaroa		X	X	X	X	X	X	X	X	
Taokas		X	X	X	R		X			
Babuza		X	X	X	R		X			
Papora			X	X	X	X	X	X?		
Hoanya		X	X	X	X		X	?	X	
Siraya		X	X	X	X	X	X			
Rukai	X	X	X	X	X	X	X	X		(X)
Puyuma	X	X	X	X	X	X	X	X	X	X
Paiwan	X	X	X	X	X	X	X	X	X	X
TOTAL	7	17	18	18	14	11	16	10	8	3/5

As the numerical values at the bottom of table 7 show, there is a clear distributional pattern for reflexes of *esa/isa, *duSa, *telu, *Sepat, *lima, *enem, *pitu, *walu, *Siwa and *puluq. Apart from ‘1’, which is unstable in many of the world’s languages due to its tendency to evolve into or from an indefinite article, and ‘7’, which is also out of place in the numeral-based phylogeny, there is steady decline from 2-10. Sagart sees this in terms of a theory of addition: first *pitu arose by contraction and other changes from a fictitious *RaCep-i-tuSa, then *lima ‘hand’ gave rise to a new word for ‘5’, then *enem appeared from somewhere, then *walu and *Siwa arose from processes parallel to those that

[11] For Amis (X) marks the fact that although ‘10’ is represented by an innovation *moʔtep*, higher multiples of ten use a reflex of *puluq, as with *tosa poloq* ‘20’, *tolo poloq* ‘30’, etc. (data from fieldnotes on Central Amis, collected January-May, 2002). For Rukai (X) indicates an irregular reflex of PAN *q in this form, although internal correspondences permit the reconstruction of Proto-Rukai *poḷoko ‘10’ (Li 1977). The latter may be a loan, but if so its source is unclear and its regularity across Rukai dialects/languages implies borrowing before the attested Rukai communities diversified from an immediate common ancestor.

produced *pitu (even though *enem had meanwhile arisen in an unrelated way), then *puluq replaced *iCid ‘10’. However, this whole argument can be turned on its head.

It is intuitively likely that higher numerals have lower text frequency, which in turn correlates with lower stability. If both of these assumptions are essentially correct much of the pattern seen in table 7 follows automatically from differences in text frequency: ‘10’ would be replaced first because it is used less often than ‘8’ or ‘9’, these numerals would be replaced next because they are used less often than ‘7’, and so on. As already noted, ‘7’ is out of place in this frequency-based sequence, but this is true in Sagart’s account as well, and frequency studies for numerals need to be conducted to determine whether this disconformity is a frequency effect, or a consequence of the greater likelihood of ‘6’ and ‘8’ being replaced by multiplicative numerals (2x3, 2x4).

The validity of this assumption can be tested empirically, at least for English, by simply typing the numbers ‘one’ through ‘ten’ into Google, which immediately supplies the number of hits in titles. On March 10, 2014 this yielded the following results:

FIGURE 7: RELATIVE FREQUENCY OF ENGLISH NUMERALS 1-10 ON GOOGLE

Numeral	number of results
one	3,490,000,000
two	1,470,000,000
three	914,000,000
four	542,000,000
five	519,000,000
six	395,000,000
seven	235,000,000
eight	175,000,000
nine	145,000,000
ten	471,000,000

As Figure 7 shows, there is a steady decrease in frequency in moving from 1-9, but a spike again at ‘10’, which drops sharply when entering the teens. The Google Books Ngram Viewer (https://books.google.com/ngrams) provides graphic displays based on corpus data from digitized books, and this shows somewhat different frequency counts of English numerals written out as full words, with the relative frequency of the first ten numerals in English given as 1-6, 8, 10, 9, 7.

In a paper that is exceptional for citing frequency data for AN numerals, Yamada (1991) considers both the unaffixed basic numerals and numerals formed by Ca- reduplication, which refer exclusively to human referents, or more broadly to animate referents in different Formosan and Philippine languages. Among languages for which this

information is given are Amis, Itbayaten and Ivatan, as shown in Figure 8. All data apparently is extracted from short texts; numerals are listed sequentially from 1-10, with number of tokens under language names:

FIGURE 8: RELATIVE FREQUENCY OF NUMERALS 1-10
IN THREE AUSTRONESIAN LANGUAGES

	Amis	Itbayaten	Ivatan
1-plain	1	Ø	Ø
-Ca-	12	172	221
2-plain	6	Ø	Ø
-Ca	21	11	64
3-plain	7	19	Ø
-Ca	2	3	56
4-plain	1	5	Ø
-Ca	1	Ø	19
5-plain	5	4	Ø
-Ca	Ø	Ø	13
6-plain	Ø	Ø	Ø
-Ca	Ø	Ø	8
7-plain	2	5	Ø
-Ca	Ø	3	10
8-plain	1	Ø	Ø
-Ca	Ø	Ø	Ø
9-plain	Ø	Ø	Ø
-Ca	Ø	Ø	2
10-plain	2	1	Ø
-Ca	Ø	Ø	Ø

Although it is based on limited data compared to English, what Yamada's paper shows clearly is that lower numerals tend to have higher text frequencies than higher numerals. Rearranging the data of Figure 8 by descending frequency and combining the plain and

Ca- reduplicated forms since they have the same numerical value, the results are: Amis 2, 1, 3, 5, 4/7/10, 8, 6/9; Itbayaten 1, 3, 2, 7, 4, 5, 10, 6/8/9; Ivatan 1, 2, 3, 4, 5, 7, 6, 9, 8/10.

It has been known for well over a century that many of the languages of Borneo, as well as the Malayo-Chamic languages that originated on the same island, have innovations in the numerals 7-9 or 7-10, as in the Uma Juman dialect of Kayan (*ji* '1', *duaʔ* '2', *təluʔ* '3', *pat* '4', *limaʔ* '5', *nəm* '6', *tusu* '7', *sayaʔ* '8', *pitan* '9', *pulu* '10'; Blust 1977a), the Sebop dialect of Kenyah (*jah* '1', *duah* '2', *təluʔ* '3', *pat* '4', *ləmah* '5', *nəm* '6', *tujək* '7', *ayah* '8', *piʔah* '9', *jəjap '10'*; Blust n.d.), most other languages of Sarawak and some of Sabah (Ray 1913:58-64), or Malay/Indonesian (*satu* '1', *dua* '2', *tiga* '3', *əmpat* '4', *lima* '5', *ənam* '6', *tujuh* '7', *dəlapan* '8', *sambilan* '9', *sə-puluh* '10', where '8' < *dua alap-an 'two taken away', and '9' < *esa ambil-an 'one taken away').

The patterns shown here are not conclusive. Indeed, the numeral '10' appears to be more stable than some lower numerals. However, it is clear that higher numerals tend to have lower text frequencies, which correlates with lower stability, and that this accounts for much of the structure seen in table 7. The major remaining puzzle for those who accept the consensus PAN numerals is why *puluq has been so unstable in Formosan languages as compared with MP languages, apart from those that have replaced the original decimal system, presumably under contact influence. Whatever explanation is ultimately found for this observation, it hardly seems adequate in itself, or even with a scattering of other observations, to justify the radically unorthodox position that Sagart has advocated.

3.3. The position of Tai-Kadai, II. The classification of Tai-Kadai (TK) as part of AN is based largely on three assertions: 1. that PTK, like PMP, reflects PAN *-mu '2pl gen.' as '2sg. gen.' through a politeness shift first proposed in Blust (1977), 2. the claim that TK languages have replaced PAN *qayam with *manuk 'bird', and 3. the observation that TK languages seem to have cognates of the AN numerals *lima '5', *enem '6', *pitu '7', *walu '8', *Siwa '9', and perhaps *puluq '10', all of which Sagart treats as post-PAN developments . Each of these putative shared innovations is problematic in more than one way.

3.3.1. The first Austronesian politeness shift: *-mu '2pl.' to *-mu '2sg.'. The first piece of evidence for including TK within the AN language family is the observation that both Buyang –ma[312] '2sg.' and PMP *-mu '2sg. gen.' appear to reflect the first AN politeness shift, an innovation in which PAN *-mu '2pl. gen.' replaced *-Su '2sg gen.'

while the corresponding nominative pronoun *kamu remained 2pl. (Blust 1977b).[12] However, given the brevity of this form and the fact that *a* : *u* is not a recurrent sound correspondence, this comparison obviously should be treated with caution. Furthermore, since Sagart (2004, 2008, to appear) sees PTK and PMP as coordinate branches of Puluqish, the proposed innovation that he imagines to be evidence for subgrouping TK must have been independent in Proto-Tai-Kadai and PMP if it is accepted at all, as other Puluqish languages do not reflect it (Ross 2006:534).

3.3.2. From *qayam to *manuk 'bird'. The second observation that Sagart uses to place TK within the AN family concerns the distribution of two words for 'bird'. Most Formosan languages, including those at the top of his tree, reflect *qayam: Pazeh *ayam*, Tsou *zomə*, Saaroa *ałamə*, Bunun *qaðam*, Rukai (Mantauran) *a-alamə*, Kavalan *alam*, Puyuma (Nanwang) *ʔayam* 'bird', Puyuma (Tamalakaw) *Hayam* 'bird', *Haya-Hayam* 'animal (other than human beings)', Amis *ʔayam* 'bird; chicken', Paiwan *qayam* 'any omen bird', *qaya-qayam* 'bird (generic)'. This form is rare outside Taiwan, but appears in at least Bontok *ay-áyam* and as *aya* in a few South Halmahera-West New Guinea languages (Ansus, Serui-Laut, Woi, Wandamen), where it means 'bird'. However, reflexes of *manuk/maNuk 'bird' are found in Tai-Kadai languages such as Buyang *ma^{o}nuk*[11] 'bird', in a two-language clade that Sagart calls 'Northeast Formosan': Basai *manuk(ə)*, Trobiawan *manukka* 'bird', and in PMP *manuk 'chicken', *manu-manuk 'bird'. What do we do with such recalcitrant data?

Sagart (2004:425) tries to reach an accommodation with the complexity of the evidence in the following way: "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 ... 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."

This was a key piece of evidence for Muish (Fig. 5), which was later abandoned through the reassignment of Northeast Formosan to Walu-Siwaish, and of Paiwan, Puyuma and Amis to Puluqish alongside PTK and PMP (Fig. 6). However, it remains as evidence for

[12] Basing himself on a single citation from a text in undated and unpublished fieldnotes collected by the Japanese linguist Erin Asai prior to the Second World War Sagart (2004:425), citing Li (1995a:667), claims that Trobiawan of northeast Taiwan also shows the second AN politeness shift, but this disagrees with the closely related Kavalan, which reflects *-Su '2sg. gen.' (Sagart 2004:425), and is out of keeping with all other evidence (Ross 2006:533-534). With reference to Blust (1977), Sagart (2004:425) says "In a more recent paper (1995) he acknowledges that **-Su did not disappear as 'your' (SG) in MP languages, but maintains that *-mu is an MP innovation. The coexistence in MP languages of reflexes of *-Su and *-mu as 2SG-genitive pronouns probably means that both existed side-by-side in PMP." However, Sagart's reference fails to distinguish two different 1995 publications, and I find no statement in either of these that agrees with the claim he attributes to me.

including TK *within* AN rather than as a sister group, since in Sagart's interpretation it was a post-PAN innovation. But this interpretation is flawed in multiple ways.

First, as noted by Blust (2002), a reference which is not cited in this, or subsequent treatments of the problem, PMP *manuk meant 'chicken', not 'bird' (cf. ,PMP *manu-manuk 'bird'). Sagart's proposed reconstruction of PMP *manuk as 'wild bird' and *qayam as 'domesticated bird' thus essentially reverses the glosses justified by the available comparative evidence. Second, and partly for the reason just stated, PAN *qayam 'bird' is best treated as distinct from PWMP *qayam 'domesticated animal', a term with reflexes that refer not only to chickens (Malay *ayam*), but also to dogs (Bikol *áyam* 'dog, canine'), domesticated pigs (Murik *ayam* 'domesticated pig'), or pets in general (Sarangani Manobo, Iban *ayam* 'pet', Palauan *ou-ʔárm* 'keep (animal) as a pet; raise animal'). Third, although PMP *manuk surely meant 'chicken', and the PAN reconstruction in Blust and Trussel (ongoing) has carried this gloss, it is unlikely that PAN *manuk had this meaning, since the Southeast Asian red jungle fowl was not domesticated until at least 5,400 BP in southern China or mainland Southeast Asia (Storey, Athens, Bryant, Carson, Emery, et al. 2012). In other words, chickens probably did not become available to AN speakers until after the MP migration to the northern Philippines (virtually all words for 'chicken' in Formosan languages are onomatopoetic, and presumably came in sometime after the aboriginal settlement of the island).

A hypothesis which is better able to explain the comparative data for both TK and AN languages is therefore that *manuk (or *maNuk) was the Proto-Austro-Tai word for 'bird'. This word survived into PAN, where *qayam was innovated, possibly as a cover term for 'omen bird', which is its meaning in Paiwan. Omen birds were important in traditional AN-speaking societies in Taiwan, Borneo, and probably elsewhere, and the antiquity of this cultural trait can be inferred from the reconstruction PAN *SiSiN 'an omen bird: *Alcippe* spp.' (Li 1995a:657, Blust and Trussel, ongoing). Both *manuk and *qayam survived in PMP, with *manuk applied to the newfound and economically valuable chicken, and *qaya-qayam (from *qayam 'any omen bird') co-existing as a dispreferred alternative to *manu-manuk (derived from *manuk 'chicken') in the meaning 'bird'. Reflexes of *manuk are thus poor evidence for situating TK within the AN family, because the claim that they replaced PAN *qayam in the meaning 'bird' is open to other, more plausible interpretations.

3.3.3. Tai-Kadai cognates of PMP 5-10. The third observation that Sagart uses to place TK within the AN family rather than outside it goes back to Benedict (1942), namely the existence of Kra-Dai numerals that appear to have the same origin as PMP *lima, *enem, *pitu, *walu, *siwa and *puluq, as with Buyang *ma*^{312}, *nam*^{54}, *tu*^{312}, *ma^{o}du*^{312}, *va*^{11}, *put*^{54}. Because he accepts these as cognates the entire higher phylogeny of AN based on the implicational hierarchy of nested numerals would unravel if TK were placed *outside* AN, since this would force acceptance of the PAN traditional numerals as retentions from an earlier Proto-Austro-Tai. To avoid this disastrous consequence for the numeral-based

higher phylogeny of AN he places TK within Puluqish, and the problem disappears. As already noted, the fact that *other* problems then appear, as why the cognate density of TK with other Austro-Tai languages is so low, is dismissed with a wave of the hand.[13]

3.4. Can Sagart's phylogeny be saved? In considering whether Sagart's unorthodox position can be saved, it will perhaps be helpful to the reader to summarize the problems it creates. These are:

TABLE 8: PROBLEMS WITH SAGART'S HIGHER PHYLOGENY OF AUSTRONESIAN AND THE POSITION OF TAI-KADAI[14]

I. Conflict with Western Plains
    1. ignores evidence of *s/t merger
    2. treats *n/ŋ merger as having no subgrouping value
    3. ignores WP innovation for '4'
    4. ignores WP innovation for '9'
    5. ignores WP innovation for '3sg'
    6. ignores WP innovation for 'house'
    7. ignores WP innovation for 'husked rice'

II. Conflict with East Formosan
    1. interpretation of *j forces the reconstruction of two PAN palatal nasals
    2. interpretation of *j leads to more place features for PAN nasals than for stops
    3. interpretation of *j leads to phonetically bizarre reflexes in Formosan languages
    4. Sagart's reconstruction provides no source for PMP *ñ

III. The position of Tai-Kadai I
    1. offers no plausible explanation for low cognate density with subgroup mates

IV. The derivations
    1. whatever its level of reconstruction, *RaCep is unambiguously *RaCeb
    2. several reflexes of *RaCeb in the western plains appear to be loans
    3. there is no valid comparative evidence for *tuSa
    4. Sagart's model predicts incorrectly that Proto-Enemish 'six' should be *bisa
    5. *pitu cannot reflect *RaCeb-i-duSa without *ad hoc* devoicing of *b and *d

---

[13] Ostapirat (2005:126) has drawn attention to other observations which he believes point to TK being a sister group of the AN family rather than a member of it. Although his discussion of a possible *t/C retention in TK languages is incompatible with Sagart's derivation of PTK and PMP from Muish, in the revised phylogeny (2008, to appear) both can be derived from Puluqish, since Paiwan and Puyuma, which are also assigned to this group, retain the *t/C distinction. Similarly, reflexes of *Cumay 'bear' are found in both Formosan and TK languages, but never in MP. However, since languages that Sagart assigns to Puluqish have reflexes of this form there is no conflict; rather, what Ostapirat has established in reaction to Sagart's proposals is that TK cannot be a descendant of PMP, but that was never claimed either in Sagart (2004) or in later incarnations of the revised phylogeny.

[14] All problems cited here are with reference to Sagart's *revised* proposals (2008, to appear). Other problems are apparent in considering his original statement of the numeral-based phylogeny.

6. the claim that interstage *b-i-dúSa had penultimate stress is based on a morphological analysis that has no known support
7. the loss of –**Sa* in interstage *b-i-duSa under the second pruning is *ad hoc*
8. *RaCeb-a-telu '8' is based on spurious data
9. *ba > *wa* is *ad hoc*
10. *e > *i*/iC_ is *ad hoc*
11. *pa > wa is *ad hoc*
12. the loss of *-t in interstage *Siwát under the second pruning is *ad hoc*

V. The implications

1. the 'nesting' of numerals 1-10 in Formosan languages reflects text frequency

VI. The position of Tai-Kadai II

1. PAN *-mu '2pl gen' to PTK/PMP *-mu '2sg gen' requires parallel sound changes with no clear motivation; Buyang *–ma* '2sg' may be due to chance
2. *manuk in the meaning 'bird' is not an innovation shared by Muish or Puluqish languages with PMP
3. *lima, *enem, *pitu, *walu, *Siwa and *puluq cannot convincingly be shown to be innovations in TK rather than retentions from a more remote Austro-Tai proto-language

Any one of the broad categories marked by Roman numerals in table 8 would be seriously damaging to Sagart's argument; the fact that there are five/six is crippling.

If Sagart wishes to pursue his argument further he must abandon the misguided derivations of *pitu, *walu, and *Siwa, and allow the numerals 7-9 to appear *ex nihilo*, as is already the case for '6'. This would leave the implicational hierarchy for the cross-linguistic distribution of 5-10 as the sole basis for the structure of his tree. However, for this to work Tai-Kadai must be part of AN, not external to it, since if Austro-Tai is valid, as envisioned by Benedict (1942), the entire implicational hierarchy would necessarily follow as a product of loss rather than addition of the traditional PAN numerals. If Sagart's subgrouping claims are to have a future, then, they depend critically on the position of Tai-Kadai. Needless to say, most linguists probably consider his position on this point the most questionable part of his entire theory. To convince others that TK is part of AN some explanation for the low cognate density between TK and AN languages must be found that is less dismissive than claiming massive relexification from an unknown and unknowable source. That is simply not serious science. And, in the foreseeable future, there appears to be no ready explanation for this observation other than that the Tai-Kadai languages are distant relatives of AN that split off from the parent stock well over 7,000 years ago before the rice-growing cultures of the lower Yangze river expanded southward along the coast to eventually settle Taiwan.

REFERENCES

[1] Benedict, Paul K. 1942. Thai, Kadai and Indonesian: a new alignment in southeastern Asia. *American Anthropologist* 44: 576-601.

[2] _____. 1975. *Austro-Thai: language and culture, with a glossary of roots*. New Haven: Human Relations Area Files.

[3] Blevins, Juliette. 2004. *Evolutionary phonology: the emergence of sound patterns.* Cambridge: Cambridge University Press.

[4] Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rinehart and Winston.

[5] Blust, Robert. 1977a. Sketches of the morphology and phonology of Bornean languages 1: Uma Juman (Kayan). *Papers in Bornean and Western Austronesian Languages, no. 2*: 7-122. Canberra: Pacific Linguistics (PL A33).

[6] _____. 1977b. The Proto-Austronesian pronouns and Austronesian subgrouping: a preliminary report. *Working Papers in Linguistics* 9.2: 1 -15. Honolulu: Department of Linguistics, University of Hawai'i.

[7] _____. 1995. Sibilant assimilation in Formosan languages and the Proto-Austronesian word for "nine": a discourse on method. *Oceanic Linguistics* 34: 443-453.

[8] _____. 1996. Some remarks on the linguistic position of Thao. *Oceanic Linguistics* 35: 272-294.

[9] _____. 1999a. Subgrouping, circularity and extinction: some issues in Austronesian comparative linguistics. In Elizabeth Zeitoun and Paul Jen-kuei Li, eds., Selected Papers from the Eighth International Conference on Austronesian Linguistics: 31-94. Taipei: Symposium Series of the Institute of Linguistics (Preparatory Office), Academia Sinica, no. 1.

[10] _____. 1999b. Notes on Pazeh phonology and morphology. *Oceanic Linguistics* 38: 321-365.

[11] _____. 2001. Historical morphology and the spirit world: the *qali/kali- prefixes in Austronesian languages. In Joel Bradshaw and Kenneth L. Rehg, eds, *Issues in Austronesian morphology: a focusschrift for Byron W. Bender*: 15-73. Canberra: Pacific Linguistics (PL 519).

[12] _____. 2002. The history of faunal terms in Austronesian languages. *Oceanic Linguistics* 41: 89-139.

[13] _____. 2003. *Thao dictionary*. Language and Linguistics Monograph Series no. A5. Taipei: Institute of Linguistics (Preparatory Office), Academia Sinica.

[14] _____. 2007. Disyllabic attractors and anti-antigemination in Austronesian sound change. *Phonology* 24:1-36.

[15] _____. 2009. *The Austronesian languages*. Canberra: Pacific Linguistics (PL 602).

[16] _____. n.d. Fieldnotes on languages of central and northern Sarawak (collected April-November, 1971).

[17] _____, and Stephen Trussel. Ongoing. *Austronesian comparative dictionary* (online open access site at: www.trussel2.com/ACD).

[18] Ferguson, Charles A. 1966 [1963]. Assumptions about nasals: a sample study in phonological universals. In Joseph H. Greenberg, ed., *Universals of Language*, 2nd ed.:53-60. Cambridge, Massachusetts: The M.I.T. Press.

[19] Ferrell, Raleigh. 1968. The Pazeh Kahabu language. *Bulletin of the Department of Archaeology and Anthropology, National Taiwan University* 31/32: 73-97.

[20] _____. 1969. *Taiwan aboriginal groups: problems in cultural and linguistic classification*. Taipei: Institute of Ethnology, Academia Sinica, Monograph 17.

[21] Leplin, Jarrett. 1974-75. The concept of an ad hoc hypothesis. *Studies in the history and philosophy of science* 5: 309-345.

[22] Li, Paul Jen-kuei. 1978. A comparative vocabulary of Saisiyat dialects. *Bulletin of the Institute of History and Philology, Academia Sinica* 49.2:133-199.

[23] _____. 1995a. Formosan vs. non-Formosan features in some Austronesian languages of Taiwan. In Paul Jen-kuei Li, Cheng-hwa Tsang, Ying-kuei Huang, Dah-an Ho, and Chiu-yu Tseng, eds., *Austronesian Studies relating to Taiwan*: 651-681. Symposium Series of the Institute of History and Philology, Academia Sinica, no. 3. Taipei: Academia Sinica.

[24] _____. 1995b. Numerals in Formosan languages. *Oceanic Linguistics* 45: 133-152.

[25] _____. 2001. The linguistic position of Thao --- with some remarks on Blust's (1996) paper on Thao. In Su-chuan Jan adn Ying-hai Pan, eds., *Papers presented at the Symposium on the Plains Aborigines and Taiwan history*:165-184 [in Chinese]. Reprinted in Paul Jen-kuei Li. 2004. *Selected Papers on Formosan Languages*, vol. 2: 891-906. Language and Linguistics Monograph Series, no. C3. Taipei: Institute of Linguistics, Academia Sinica.

[26] _____. 2004. Origins of the East Formosan peoples: Basay, Kavalan, Amis and Siraya. *Language and Linguistics* 5.2:363-376.

[27] _____, and Shigeru Tsuchida. 2001. *Pazih dictionary*. Language and Linguistics Monograph series A2. Taipei: Institute of Linguistics (Preparatory Office), Academia Sinica.

[28] _____. 2002. Pazeh texts and songs. Language and Linguistics Monograph series A2-2. Taipei: Institute of Linguistics (Preparatory Office), Academia Sinica.

[29] Matisoff, James A. 1990. On megalocomparison. *Language* 66: 106-120.

[30] _____. 1995. Sino-Tibetan numerals and the play of prefixes. *Bulletin of the National Museum of Ethnology* (Osaka) 20.1: 105-252.

[31] Ogawa, Naoyoshi. 2003. *English-Favorlang vocabulary by Naoyoshi Ogawa, with an introduction by Paul Li*. Asia and African Lexicon Series, no. 43. Tokyo: Research Institute for Languages and Cultures of Asia and Africa.

[32] Ostapirat, Weera. 2005. Kra-Dai and Austronesian: Notes on phonological correspondences and vocabulary distribution. In Sagart, Blench and Sanchez-Mazas: 107-131.

[33] _____. 2013. *Austro-Tai revisited*. Paper presented at the 23rd Annual Meeting of the Southeast Asian Linguistics Society, May 29-31, 2013, Chulalongkorn University.

[34] Peiros, Ilia. 1998. *Comparative linguistics in Southeast Asia*. Canberra: Pacific Linguistics (C-142).

[35] Ross, Malcolm D. 2006. Reconstructing the case-marking and personal pronoun systems of Proto Austronesian. In Henry Y. Chang, Lillian M. Huang and Dah-an Ho, eds., *Streams convering into an ocean: festschrift in honor of Professor Paul Jen-kuei Li on his 70th birthday*: 521-563. Language and Linguistics Monograph Series no. W-5. Taipei: Institute of Linguistics, Academia Sinica.

[36] _____. 2012. In defense of Nuclear Austronesian (and against Tsouic). *Languages and Linguistics* 13.6: 1253-1330.

[37] Sagart, Laurent. 2004. The higher phylogeny of Austronesian and the position of Tai-Kadai. *Oceanic Linguistics* 43: 411-444.

[38] _____. 2008. The expansion of setaria farmers in East Asia: a linguistic and archaeological model. In Alicia Sanchez-Mazas, Roger Blench, Malcolm Ross, Ilia Peiros and Marie Lin, eds., *Past human migrations in East Asia: matching archaeology, linguistics and genetics*: 133-157. London: Routledge.

[39] _____. 2013a. The higher phylogeny of Austronesian: a response to Winter. *Oceanic Linguistics* 52: 249-255.

[40] _____. 2013b. Is Puyuma a primary branch of Austronesian: a rejoinder. *Oceanic Linguistics* 52: 481-492.

[41] _____. To appear. In defense of the numeral-based model of Austronesian phylogeny, and of Tsouic. Ms. 25 pp. *Language and Linguistics* 15.

[42] Storey, Alice A., J. Stephen Athens, David Bryant, Mike Carson, Kitty Emery, Susan de France, Charles Higham, Leon Huynen, Michiko Intoh, Sharyn Jones, Patrick V. Kirch, Thegn Ladefoged, Patrick McCoy, Arturo Morales-Muñiz, Daniel Quiroz, Elizabeth Reitz, Judith Robins, Richard Walter, Elizabeth Matisoo-Smith. 2012. Investigating the global dispersal of chickens in prehistory using ancient mitochondrial DNA signatures. *PLOS One*: 1-12 (electronic publication).

[43] Thurgood, Graham. 1994. Tai-Kadai and Austronesian: the nature of the historical relationship. *Oceanic Linguistics* 33:345-368.

[44] Tsuchida, Shigeru. 1976. *Reconstruction of Proto-Tsouic phonology*. Study of Languages & Cultures of Asia and Africa Monograph series, no. 5. Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa.

[45] _____. 1982. *A comparative vocabulary of Austronesian languages of sinicized ethnic groups in Taiwan, Part I: West Taiwan*. Tokyo: Memoirs of the Faculty of Letters, University of Tokyo, no. 7.

[46] _____. 1985. Kulon: yet another Austronesian language in Taiwan? *Bulletin of the Institute of Ethnology, Academia Sinica*, no. 60: 1-59.

[47] Winter, Bodo. 2010. A note on the higher phylogeny of Austronesian. *Oceanic*

Linguistics 49: 282-287.

[48] Yamada, Yukihiro. 1991. The numeral systems of the Formosan and the Philippine languages. *Bulletin of the Himeji Dokkyo University Department of Linguistics*, vol. 4: 119-135.

[49] Yeh, Mei-li. 2000. *Saixia yu cankao yufa* (Saisiyat grammar). Taipei: Yuan-liou.