3 *Norm-referenced lexicostatistics and Chamic*[1]

Anthony P. Grant

1. Norm-referenced lexicostatistics: introduction, history and methodology.

The lexicostatistical techniques that are used for analysis of materials in historical and comparative linguistics, which were first developed in their modern form by the American structuralist Morris Swadesh (and which were first made readily available in Swadesh 1950, see also Swadesh 1955 for a protracted exposition) have enjoyed mixed fortunes in the last half-century of historical linguistic work, although they are currently enjoying a certain degree of revival. (Glottochronology, with which lexicostatistics is often used and sometimes confused although the use of neither technique of necessity entails use of the other, is currently much less popular. Yet glottochronological dates of separation between languages and within proto-languages are still cited with reverence by non-linguistic specialists in other fields such as archaeology and anthropology, who impute to them a degree of methodological accuracy and overall reliability which few linguists would now agree with.)

The 100-item and 200-item lists (and to a lesser extent the older 215-item list) that were drawn up by Swadesh in the 1950s are still those which are used most frequently. This remains the case half a century on, even though it has long been recognised that they are not equally appropriate for all languages. Sometimes this is because of 'cultural gaps' in some languages. Often, however, it is because of differing semantic patterns, in certain fields at least, from those which were promulgated and incorporated onto the lists by Swadesh on the basis of his firsthand experiences of particular languages. Up to the time when Swadesh was assembling this list (a little before 1950[2]) this involved languages of Europe, North America, Mexico and (in part) the Far East, more specifically Mandarin and Burmese, both of which he had worked upon for the US military during WWII.

Consequently, a number of scholars have elaborated somewhat different gloss lists which are better suited to capturing certain of the semantic characteristics of a particular family of languages. This has been done on at least two occasions for the historical investigation of interrelationships within Austronesian languages. The renowned work of Dyen (1962 and especially Dyen 1965), which attempted to present a genetic classification of the Malayo-Polynesian languages by using lexicostatistical materials, used a 196-item list, namely the Swadesh 200-item list minus 'that' (the demonstrative adjective, which is not always distinguished from 'this' in these languages, though often split into different

[1] I would like to thank Bob Blust, Robert K. Headley, Russell Murray, Peter Patrick, Graham Thurgood and David Zorc and the staff of the Special Collections Reading Room at the School of Oriental and African Studies, University of London, for their assistance with aspects of the production of this paper. Any infelicities are of course my own responsibility.

[2] The first mention of Swadesh's use of this technique was in 1948, at a Viking Fund Supper Club presentation which he gave in New York that year.

Grant, A.P. "Norm-referenced lexicostatistics and Chamic". In Grant, A. and Sidwell, P. editors, *Chamic and Beyond: Studies in mainland Austronesian languages*.
PL-569:105-146. Pacific Linguistics, The Australian National University, 2005. DOI:*10.15144/PL-569.105*

forms depending upon the distance from the speaker, the visibility of the object referred to, and so on), and the tropically inappropriate 'ice', 'freeze', and 'snow'. Similarly-structured searches among the overtly-expressed morphological features of Malayo-Polynesian languages were not carried out in extenso. Nevertheless, on the basis of the findings from this lexicostatistical experiment Dyen posited the existence of 40 primary groups of Austronesian, with their area of greatest diversity (according to the findings of this lexicostatistical experiment) being in New Guinea, which he therefore proposed as the Austronesian *Urheimat*. In contrast, one of the 40 groups, the Malayopolynesian Linkage, accounted in Dyen's scheme for more than half of the languages surveyed, including practically all those languages which are now regarded as Western Malayopolynesian.[3] Dyen's vision was a view which has won remarkably little acceptance, despite Dyen's eminence in Austronesian linguistics. The reason for this is that Dyen was wrong in the inferences which he had drawn from the use which he had made of lexicostatistics (a point which was first made clear in Grace 1966, although Grace's valid reasons for his criticisms did not include an analysis of the faultiness of Dyen's lexicostatistical methodology).

In terms of the technique employed, what Dyen had used in his comparisons was *pair-referenced lexicostatistics*. In Dyen's investigation, each gloss in each Malayo-Polynesian language was compared by computer with the same gloss in every other Malayo-Polynesian language, so that each gloss in Itbayaten of the northern Philippines was compared with the appropriate gloss in Chru of Vietnam[4], Atayal of Formosa, Nauruan of Micronesia, and hundreds of other languages. What the glosses in these languages were not compared with, however, was the equivalent forms in any kind of a reconstructed proto-language at any level.

In the methodology underpinning this work Dyen was comparing Language A with Language B, Language B with Language C, Language C with Language D, and so on. This strategy is interesting in itself and can bring forth fascinating intimations of lower-level linguistic relationships (for pair-referenced lexicostatistics is very useful in certain spheres), and Dyen's concept of the 'critical percentage' (the greatest percentage of cognates which one language that is being surveyed has with any other language which is being surveyed) is valuable. But it is the wrong kind of lexicostatistical methodology to be used for what Dyen was trying to achieve, and without firstly using the right sort of methodology, his wider aims for his research and such findings as emerged from them were futile.

What Dyen did not attempt to do in the course of his lexicostatistical studies was make us of any information which would have enabled him to indicate which of the elements in these languages went back to a proto-language and which other elements were borrowings from current or previously surrounding languages (both Austronesian and non-Austronesian), later internally-driven lexical developments, or forms confined to sub-

---

[3] This is paradoxical and counterfactual because Western Malayo-Polynesian is not a proven subgroup, as it is not distinguished by the possession of any shared innovations, and therefore has to be defined negatively as being that subset of Malayo-Polynesian languages which does not possess the shared innovations of Oceanic for instance, or of Central Malayo-Polynesian. (Nevertheless Western Malayo-Polynesian does contain several well-defined subgroups of its own: Malayo-Chamic is one such.) I call such negatively-defined large groups 'antigroups'.

[4] Chru was the only Chamic language, apart from Acehnese, for which Dyen had access to a lexicostatistical list, and Dyen's findings did not pick up on the special historical connection between these two.

branches of Malayo-Polynesian (MP) or whatever. Furthermore he was interested in the number of cognates which were to be found between pairs of languages, but he was concerned with absolute figures and not with forms. The actual cognates, and the degree in each instance to which they were replicated in the vocabularies of one language or another, did not enter the picture and they were not exemplified. The result is an internally-enclosed and self-referential analysis, which has the potential to give observers a misleading picture of the relevant genetic linguistic relationships.

In short, Dyen was using an approach which was too purely quantitative, whereas the nature of the task required recourse to more qualitative methods. These methods took note of the quantitative findings which could be gathered fairly quickly, but did not confine themselves to them, going instead beneath the surface to analyse the kinds and the relative historical statuses (PAn, PMP, Proto-Malayo-Chamic, etc.) of the forms which two languages shared.

If it had been the case, for example, that in a hypothetical family Languages A and B shared 25% of the cognates on the list, and that Languages B and C shared 25%, and that Languages C and D shared 25%, but that none of the actual shared cognates were to be found in more than any two of these languages or in any more than one of the pairs listed above, then this highly significant fact, which might at least superficially cast serious and reasonable doubt upon the ultimate unity in origin of A, B, C, and D, would not have been clear from the tables of percentages presented in Dyen's study.[5] Looking at these tables of percentages of forms which are common to any two particular Austronesian languages in each case, we cannot tell from such figures which items among the commonly-shared forms are inherited from Proto-Austronesian, which other forms reconstruct back only to Proto-Malayo-Polynesian, and which other of those forms are first found in a daughter-language of Proto-Malayo-Polynesian, such as what we now call Proto-Oceanic. And we may assume that on certain occasions those words which are common to two contiguous languages and which are taken by Dyen as being cognates jointly inherited from a parent language may actually have been introduced from one to another, and it is sometimes possible that they may even have come into both languages from a third language.

Dyen was an admirer of the achievements in the Malayo-Polynesian reconstruction work of Otto Dempwolff (as, to a large extent, am I). In Dyen's published work he has given little indication that the doubts the essential correctness of the visible fruits of Dempwolff's remarkable intellectual achievements; though he does revise and improve many of the spellings of Dempwolff's PMP reconstructions, he does not doubt that they are correct and valid. Yet crucially he did not compare the gloss list for any language with those available for each item in the three volumes of Dempwolff (1934-1938). This was a lost opportunity which had considerable consequences for much later work on Malayo-Polynesian subgrouping.

Had Dyen referenced the items on each list to their occurrence or non-occurrence on (and their cognacy with) a list of equivalents which used elements derived from Dempwolff's list, he would have been practising a kind of *norm-referenced*

[5] It is always theoretically possible for two languages which are descended from the same parent language, but which belong to different subgroups and which are both low scorers in regard to lexical retention from the parent language, to have a cognacy rate of 0%, although I do not know of any certain examples of this

lexicostatistics[6], a technique in which the forms in each language are compared with the forms in the same control language, control case, or 'norm language'. This type of lexicostatistics can be seen as a development from the lexicostatistical principle which is also used as an essential part of traditional glottochronology, namely that a wordlist from one historical state of a language, which is taken as the control case or norm, is compared with a wordlist from a later historical state of the same language or with several such states of the same language (which are each compared the forms from the earlier stage of this language). When this has been done, then the number and proportion of forms remaining in the later state (or states) and that have been perpetuated from the former state, that is from the control case language, is calculated. In this particular study, however, glottochronological techniques are not going to be used.

In such a scenario as one using Dempwolff's reconstructed Malayo-Polynesian proto-forms (for want of better reconstructions), the ideal norm (or the language) against which the forms in each language were being compared, one language after another, would be an assumed and reconstructed proto-language which had been arrived at independently of the investigation of any of the daughter-languages under discussion. Using such a method early in his examination would have enabled Dyen to spot numerous recurrences of the same widespread but non-Proto-Malayo-Polynesian (and therefore not directly inherited) morphs in various languages, and this might have led to the earlier reconstruction of such important subgroups as Oceanic. Such a technique, measuring the proportion of forms which a particular language has retained from a list of forms from its proto-language, is something which Robert Blust has done in certain of his papers (for instance Blust 1993). Most importantly, Blust has shown that the number and proportion of retentions varies from one set of Malayo-Polynesian languages to the next (see also Blust 2000b for an illustration of this.).

Epistemologically at least such a comparison would have been something of a risky exercise, since one is dealing with an abstraction (namely Dempwolff's inductive reconstruction of Proto-Malayo-Polynesian), the degree of whose similarity to an assumed but unrecorded entity is uncertain (and was even more uncertain at that time). Furthermore, one is comparing elements of this abstraction with data from attested languages. Nevertheless, as one attempts to do this kind of historical reconstruction of the linguistic manifestations of actual speech-community splits, the use of such a technique demonstrates the similarities (admittedly both retentions and innovations of various sorts, including those borrowings found in more than one language) of different languages to a particular reference point, and is a valid approximation to the facts.

The findings of norm-referenced lexicostatistics are best seen displayed overtly, for instance in the form of a grid. This has been done by Miller (1984), using a modification of the Swadesh 100-word list, in an attempt to subgroup a couple of dozen Uto-Aztecan languages in North and Central America, and a modification of Miller's model (a model which is closer to the technique used in Miller, Carpenter and Foley 1971) is the one which I have pursued here. The primary purpose in such comparisons is to spot similar forms, and

[6] This term was introduced in Bennett (1998), to describe a kind of lexicostatistics that he applied to Semitic languages, in which the number of forms, inherited from a proto-language (which was the norm against which each of the modern languages was referenced), that remained in the lexicon of a modern language, was counted for each language that was surveyed. For instance it might be the case that out of 5 forms reconstructed back to Proto-Semitic, Language A retained 4 but Language B only retained 2 while Language C retained 3.

more specifically, to spot cognates between languages which represent retentions, and thereafter to distinguish them from those which represent innovations. The result is a kind of 'multilateral comparison' (a term which was made famous by Greenberg 1987), but it is one in which there is a norm language used in the comparisons, a language which may have true historical or other non-trivial significance to the project. (A similar technique was used at about the same time by Hooley 1971 in his classification of the Austronesian languages of Morobe Province in present-day Papua New Guinea, although Hooley used numerals rather than letters to separate out words belonging to different cognate sets, and he did not use special indicators in his tables of forms for missing glosses, unique forms, or loan elements as Miller did.)

Although Miller had previously published a long list of 'formulist' reconstructions of Proto-Uto-Aztecan forms (Miller 1967), he did not employ the results of this in his 1984 work. Consequently a PUA (= Proto-Uto-Aztecan) column is not provided as a norm language in his table of cognates and similarities, which is presented in grid form, and the reflexes of the forms in Miller's list are not compared with those which had already been reconstructed for PUA. In fact, Miller does not cite the actual forms used for the expression of each gloss in each language. Instead, what Miller did was to start from the leftmost and most northerly languages in his table, the Numic languages of eastern California and the Great Basin, and to assign the letter 'a' to the word which is used in this language, so that the reflex of each word in this leftmost language is always marked with 'a'. If the next language used a form of a different word to express the same concept, then 'b' is used, and if a further language uses a form of a word which is different from both of these then 'c' is used, and the process continues this way.

When drawing up his table Miller used the symbol 0 for cases in which a form for a particular gloss in a particular language was not available to him, so that a particular slot or cell had to be left empty, while he capitalised the letters in cases representing words in the list which had been borrowed from another language. Instances in which the gloss for a particular item was represented by a form which was exclusive to that language and which was found in no other language in the sample, were represented with 'x'; there could be more than one 'x' in each line of the list (sometimes there were half a dozen or more). If a loanword was only attested in one language, it too could be capitalised as X. We may call the grid which results from these procedures a *cognate grid* or *cognacy grid.* Cognate grids may not necessarily result from the application of principles of norm-referenced lexicostatistics (and we have seen that Miller was not using such norms), but they can be developed for use in data regression after the application of this kind of lexicostatistical discovery procedure.

For Austronesian languages the default lexicostatistical list used nowadays is that drawn up in Blust (1981), a brilliant and still unpublished paper. The list draws upon the work of another Borneanist, Alfred B. Hudson (Hudson 1967), which used a 203-item list, but goes beyond it in terms of its range of applicability, and versions in English and Malay have been widely circulated. In addition, Blust (1993), a paper which uses this list as a basis, provides reconstructed Proto-Malayo-Polynesian translations or equivalents for every item on the 200-gloss semantically-arranged list, and well over half of these forms (at least 116: Robert Blust, personal communication, 1997) are also attested in some or all Formosan languages and can thus be reconstructed back to Proto-Austronesian, with appropriate phonological adjustments. Almost 85% of the items on Blust's list are to be found listed on either the 100-item or 200-item Swadesh lists, while the remaining forms

are (with a couple of exceptions) well-suited to one's expectations of the assumed semantic primes of the lexica of Austronesian languages. Blust's list is much better suited to this particular task and to these particular languages than the one which Dyen used or adapted (though of course it would also have been perfectly feasible to practise norm-referenced lexicostatistics using Dyen's 196-item list), and Blust's list is the one which I have used below.

Blust (2000b) has made a terminologically useful distinction between *horizontal lexicostatistics* and *vertical lexicostatistics*. The former technique is the one which is more widely used nowadays (though this was not always so). This technique compares lexical data from languages which are supposed to have been attested in the same time period and to be roughly contemporaneous. Meanwhile the latter technique compares lexical data from an earlier stage of a particular language with data from other languages which are assumed to be descendants from this language. Comparisons between material from Classical Latin on the one hand (Latin being the control case or norm) and French, Spanish, Italian and so on, on the other, would be an example of the use of vertical lexicostatistics. Comparisons between French, Spanish and Italian would be instances of horizontal lexicostatistics. The study offered in this paper uses horizontal lexicostatistics as a point of departure, since most of the languages compared are contemporaries of one another, but additionally it incorporates the findings which vertical lexicostatistics (and more specifically, which the use of the Blust 200-item list) can give us.

For Blust (2000b: 320) horizontal lexicostatistics is characterised by a known retention rate (which Swadesh had long since set at 0.81 per millennium, or 81/100 items are supposed to be retained from the word list after a thousand years), an unknown period of divergence between the two or more contemporary languages that were being surveyed (indeed we may say that the time when these diverged was the question to which we were seeking an answer), and an ability to calculate these figures horizontally. With vertical lexicostatistics the rate of retention was unknown, but the time of divergence between the control case language and the descendant language(s) was supposed to be known, and the figures could be calculated vertically. The unspoken assumption is that in vertical lexicostatistics all the languages concerned diverge from the ancestral language to approximately the same degree. But this is not the case with horizontal lexicostatistics, and this is supposed to enable us to subgroup languages (and then to construct family trees) according to their depth or recency of split from one another.

Combining the strengths of historical investigation and of the use of a cognate grid in norm-referenced lexicostatistics in which the norm comprises items from a reconstructed language allows one to take advantage of the strengths of the various subfields: the use of a well-selected lexical sample (a choice of material which is especially germane in the case of languages which have minimal inflectional morphology of the sort relied upon for historical linguistic purposes by diachronists), and the ability more clearly to see patterns of lexical distribution within a chosen sample of languages.

There is also the benefit that can be provided by working from a set of reconstructed forms, which (if we have enough historical background information to make assumptions secure) allows one to recognise whether the equivalent form in a modern language which is being surveyed is an inherited form that the proto-anguage contained, or whether it is one or another kind of innovation. Different kinds of such innovations would include borrowing (including the borrowing of a form which is cognate to one which might have been found in the lexicon of the proto-language under discussion, and thus a 'false

cognate'), internally-developed form, or whatever. Indeed Blust (2000b) pointed out that it is the inability of horizontal lexicostatistics to be able to let us distinguish between inherited forms and other forms which are innovations shared between two or more languages (but not between all the languages that are being surveyed), which vitiates this technique. With the use of vertical lexicostatistics this confusion of the historical status of elements does not happen.

2. The Chamic languages in their historical and contact setting.

The Chamic languages, long overlooked or misclassified as Austroasiatic by linguists as recently as Sebeok (1942), have received considerable recent attention in the Austronesian linguistic literature, thanks very largely to the work of Graham Thurgood over the past decade (for instance in Thurgood 1996; the work which he has carried out is encapsulated in Thurgood 1999; see latterly also Thurgood to appear a, b, c, Thurgood and Li 2003). Thurgood's work, rooted as it is in historical phonology and the use of 'top-down' and also 'bottom-up' modes of reconstruction[7], and with its copious references to parallel forms in Malay (which shares a number of non-trivial phonological developments of Proto-Malayo-Polynesian sounds with those which are found in Chamic languages), demonstrates beyond reasonable doubt that the speakers of the ancestor of the Chamic languages left Borneo (where its sister-languages were spoken) a few centuries before Christ. This is also what Proto-Malayic had done, although the movement of the speakers of Proto-Malayic from Borneo took place probably some centuries after the departure of the speakers of Proto-Chamic (or maybe Pre-Proto-Chamic).

The linguistic evidence which can be gleaned from the responses to the Blust list and from other sources also shows the skeptical observer that the speakers of Chamic languages, like those of the Malayic languages which are its closest genetic relatives, have returned to mainland Asia after their ancestors spent millennia in the islands, rather than having remained in Asia in situ for millennia.[8]

The most widely-spoken Chamic language is Acehnese of extreme northern Sumatra, with over 2 million speakers (its relationship with other Chamic languages, which is beyond doubt, is discussed in part in Durie 1990). It is one of two Chamic languages which has left Indochina, the other being Tsat or (as it is called in Putonghua) Huihui, a language spoken by a Muslim minority of a few thousands in two villages on the extreme south coast of Hainan, China, who descend at least in part from Chamic-speakers who

---

[7] 'Top-down' reconstruction, starting with forms which can reasonably be assumed to have occurred in a proto-language and then tracing their phonological histories in the various daughter languages, is preferable in Chamic languages, because it is certain that they are all related to one another, and because many of the customary reconstructional techniques of historical linguistics are difficult to apply to items in Chamic languages as a result of the varying but often dramatic effects of changes in the forms of the syllable, especially in the presyllable segments. For example Cham, just like Malay, has *lima* but Jarai has *rema*, Rade has *ema* and Tsat has *ma33* for proto-Chamic **lima* 'hand, five' (PMP **qalimah*). These changes are perfectly in accordance with the developments of Proto-Chamic historical phonology in each language, even though in other phonological environments PMP *l would become /l/ in all the languages concerned.

[8] Proto-Chamic and Proto-Malay share the same diagnostic reflexes of PMP sounds such as **q*, **Z*, **R*, **D*, **b-* and also the same innovative shapes of PMP words such as **wahiR* 'water', and **qaqaɣ* 'leg, foot', features which allow them to be subgrouped together against languages such as Moken of the Mergui Arcipelago, Burma, and Javanese.

migrated from what is now Vietnam maybe a millennium ago[9]. Of the other languages, the most widely known is Cham, which was formerly the language of Champa, a series of kingdoms of Hindu/Buddhist cultural affiliation, part of the East Asian Indosphere, the southern part of which was finally brought to its knees in 1471 by Khmer invasions (the northern kingdoms had succumbed to the incursions of the Vietnamese in 982, when the Vietnamese were themselves responding to pressure from the Chinese to the north). Cham survives in two differentiated dialects which now have the status of separate languages. These areEastern Cham or Phan Rang Cham of Phan Rang, formerly known as Panduranga, in coastal Vietnam, and the emigrant Western Cham of the area around Tonle Sap in Cambodia, and of Chau Đoc and other Khmer-speaking areas in the Vietnamese part of the Mekong Delta. An earlier form of the language, as it was spoken before the dialectal division and before the strong impact of Vietnamese on Eastern Cham, was (and to a slight extent still is) used as a written language by male Chams, employing a distinctive alphabet of Indic origin.

Other Chamic languages are Jarai and Rade/Rhade/Ede, which are spoken in the Vietnamese highlands, Haroi, which has moved to the highlands from coastal Vietnam, and two other languages or language groups spoken in areas near the Vietnamese coast, namely Chru and Roglai (the latter includes several forms of speech, notably Northern Roglai, which is the best described form, Southern Roglai, and the aberrant Cat Gia Roglai, all of them used in coastal regions of Vietnam). These languages are all clearly related, as even a cursory inspection of wordlists shows, but just as clearly they exhibit an impressive array of variation and diversity, especially in regard to the developments in each language of features of Chamic historical phonology. Nevertheless there are phonological developments from Proto-Malayo-Polynesian (hereafter PMP), such as the change of initial PMP *n-* (itself a very rare sound word-initially) to *l-*, which are common to all Chamic languages including Acehnese and which, regionally at least, mark them out from other Malayo-Polynesian languages in the area (including Malay in this instance). It should be understood, though, that these changes are not exclusive to Chamic throughout the whole Austronesian world.

To the best of my understanding, almost none of the languages listed above are mutually intelligible. Phan Rang Cham and Western Cham may be a partial exception, as these may be interintelligible, although Western Cham has absorbed a large amount of lexicon from Khmer, including epistemic particles and other grammatical morphs, and none of this is found in Eastern Cham. Meanwhile male speakers of Cat Gia Roglai are bilingual in Phan Rang Cham: the situation is discussed in Lee (1998), but this societal bilingualism does not constitute mutual intelligibility of the languages involved.

Phonologically the most aberrant Chamic languages are Tsat (this aberrancy has come about as a result of influence from non-Chamic languages such as Hlai and

[9] But there may have been more than one wave of migrants from Champa to Hainan, and it is further possible that several centuries may have elapsed between migrations to Hainan (Pang 1998). Nor need the different waves of migrants of necessity have come from the same region in Champa. Indeed, as Graham Thurgood pointed out (personal communication, 22 March 2002), there is evidence of some dialect mixture within the Chamic component of Tsat, with some southern forms (for instance the numeral 'hundred') being mixed at a later period into the basically more northern language which gave rise to Tsat as we now know it (the lower numerals show more distinctly Northern Chamic traits, insofar as diagnostic forms are available for inspection).

Hainanese), Rade, and Cat Gia Roglai (or Cac Gia Roglai). In the last two cases there is no obvious external linguistic motivator for the striking and surprising – and, it must be noted, very different – sets of developments in their historical phonology. We cannot state that they have modified their phonologies in order for the resulting system to resemble more closely the phonological system of any particular neighbouring language. The fact that the remarkable presyllabic phonological constraints in Rade resemble those of the Mon-Khmer language Chong of Laos and eastern Thailand, which is a Pearic language, is almost certainly a coincidence. This is because there are several Mon-Khmer and other languages separating the areas populated by speakers of Rade and Chong, and these separating languages do not share these highly marked presyllabic constraints (see Thurgood 1999). Matisoff (2001) has, however, pointed out that there are some apparent shared innovations, in terms of the kinds of massive erosion that the forms undergo, between the construction of presyllabic onsets in Rade and those found in Tsat.

There are published and unpublished descriptive materials available for all of these languages, but only Acehnese and to a lesser extent Tsat and Phan Rang Cham are well-described in regard to lexical coverage. The provision of text collections, and grammatical descriptions are rare for these languages, and only Acehnese and (rarely) Phan Rang Cham are used in writing. I refer to the Chamic languages apart from Acehnese and Tsat as Indochinese Chamic languages; I would point out that I use this term as no more than a geographical expression and I would assert that no genetic considerations, suggesting that Indochinese Chamic languages constitute a single genetic subgroup, should be read into it. It is simply that they are Chamic languages which remained in Indochina throughout.

Typologically and especially phonologically Chamic languages resemble Mon-Khmer languages (including the Bahnaric languages with which Proto-Chamic was in prolonged and intimate contact, as well as Khmer and Vietnamese, with at least one of which most speakers of Chamic languages have been in contact).[10] In fact they look superficially like Mon-Khmer languages much more than they resemble the Western Malayo-Polynesian languages of Borneo, including such languages as Proto-Malayic, from which they have derived. Even more so than what has happened with many Malay dialects, the Chamic languages have been integrated into the Southeast Asian linguistic area more and more over the past couple of thousand years. The result of this is that they now exhibit such Southeast Asian areal characteristics as numeral classifiers, which are also found in

[10] The customary classification of Mon-Khmer languages within Austroasiatic recognised eleven groups organised into four larger branches: Northern Mon-Khmer, Southern Mon-Khmer, Eastern Mon-Khmer and Viet-Muong or Vietic. The first branch includes Khasi, Palaungic and Wa, and these languages have not been involved with Chamic languages. Southern languages are Monic, Nicobarese, and the Aslian languages of Malaya (the latter have influenced Acehnese but have not otherwise been involved with Chamic languages). Eastern Mon-Khmer groups are Pearic, Khmer (the closest relative of Pearic), Bahnaric languages (with two major divisions) and Katuic. (Southern and Eastern Mon-Khmer languages are themselves regarded as being the two branches of South-Eastern Mon-Khmer, a grouping which is parallel to Northern Mon-Khmer and to Vietic.) Vietic languages constitute the fourth branch, although it is possible that they are most closely related to Katuic languages. Eastern Mon-Khmer languages, specifically Bahnaric and Katuic, and in many cases latterly Vietic (specifically Vietnamese) have been the languages which have exclusively exerted the Mon-Khmer influence on all languages including Acehnese, though the latter has, as previously stated, been in later contact with Aslian languages (though not with Vietnamese). Paul Sidwell (personal communication) indicates that Katuic languages exerted strong influence upon Bahnaric languages.

Malay and in some North Sarawak languages but which are not part of the structures of many other Austronesian languages. This absence is true even those languages which contain a considerable stratum of loans from Chinese languages: this is the case for instance of Tagalog, which lacks numeral classifiers (although it does contain numerous loans from Hokkien Chinese).

The Chamic languages have furthermore adopted or acquired many of those salient typological characteristics of Mon-Khmer languages which are not also pan-Southeast Asian typological features which cut across genetic lines. (It is a reasonable asusmption that Mon-Khmer languages are the major source for Southeast Asianisms in the Chamic languages.) These features include the prevalence in the vocabulary of monosyllabic and sesquisyllabic contentive stems, the presence and widespread use of glottalic consonants and of many vowel nuclei alien to most Austronesian languages, the use of derivational infixation (rather than the more primarily inflectional infixation found in many Austronesian languages), and most significantly, distinctive registral patterns – patterns which have sometimes (as also with Mon-Khmer languages such as Vietnamese) led to the development of partial or full tone systems. This development has happened independently in Tsat, Phan Rang Cham and to a slight extent in Jarai.

The vocabulary of most of the Chamic languages contains a greater number of lexical items of Mon-Khmer origin than there are those of Austronesian, Malayo-Polynesian or even Malayic origin (these latter numbering a few hundred at most). Even the proportion of *undoubted* Mon-Khmer elements in the reconstructed vocabulary of Proto-Chamic is well over 15% (I counted 205 assured Mon-Khmer-derived items out of 755 Proto-Chamic and post-Proto-Chamic forms that are listed in Thurgood 1999, and there are over 200 further forms which may be of Mon-Khmer derivation). It is certain that all Chamic languages have been recipients of this 'partial relexification', as many core items that are of (say) Bahnaric origin are also found in Acehnese, as are a number of basic pan-Chamic forms which are of uncertain origin. And most of what few productive (or even unproductive) elements of bound morphology there are either derive from Mon-Khmer languages or else are very close in both form and meaning in Mon-Khmer and Austronesian languages. By contrast, most of the small battery of inherited Western Malayo-Polynesian affixation has either been lost completely, or at best is preserved in a few frozen stems and is no longer productive. There may be Austronesian languages that have retained fewer elements from their Proto-Austronesian lineage than the Chamic languages have, but there cannot be many of them (such languages are found in Papua New Guinea and the Solomons).

A fairly strong case could be made for claiming that the Chamic languages are mixed languages (and that they are to some extent even intertwined languages in the sense in which the term is used in Bakker and Mous eds. 1994, since some of what little bound morphology they have is taken from Mon-Khmer languages). It is possible that such 'linguistic mixture' has taken place here because the earliest Cham communities were built up mostly by exogenous men (the Chams were notorious pirates), who were in a position socially, politically and technologically to dominate the members of the communities upon which they had intruded, and who intermarried with indigenous Mon-Khmer-speaking women, upon whom they imposed their Austronesian language once they had established coastal communities.

For its part, Tsat, a language which started out being very similar to Northern Roglai, has become typologically more and more like Hainanese Chinese and latterly more

like Mandarin Chinese as time has progressed, and this direction of change is manifested both in lexicon and morphosyntax (Thurgood and Li 2003). The salient features of Tsat segmental and canonical phonology look like a subset of those of a modern Southern Chinese language, and this has extended to the development of a full five-tone system whose origins Thurgood (1999) reconstructs on the basis of changes in Tsat historical phonology since the language's separation from other Chamic languages. The relics of PMP prefixes and infixes which can be found in other Chamic languages have been completely lost from sight in Tsat, since the words which contained such forms have undergone far-reaching sound changes, to the extent that the syllabic canons and prevocalic consonantal forms which are now permitted in Tsat are a subset of those permitted in Hainanese. Only the use of internal reconstruction and subsequent comparison with corresponding forms in other Chamic languages can shed light on the underlying phonological forms of Tsat words, so that only reconstruction from the top down could show the clear Austronesian origin of more than a small number of Tsat forms.

By contrast, what makes it possible for us to classify the Chamic languages genetically as Austronesian or even Malayo-Chamic is their possession of a few hundred morphs, very few of them bound (such inherited bound morphology as Chamic languages have is no longer productive and much of it has been lost completely) and the bulk of them lexical items which centre in the most frequently-used and generally the most culturally-neutral items of the vocabulary of Chamic languages. Yet even this most basic lexical element is not exclusively a Malayo-Chamic domain, as the table below makes clear. Much of the Chamic lexicon of all kinds, including very many high-frequency verbs, derives from Mon-Khmer languages, and this includes numerous forms which are found in most or all Chamic languages, and with the impact of (especially) Vietnamese on modern Chamic languages, this proportion is growing even more. There is an ineluctable Mon-Khmer element (over 10% of the total at a conservative estimate) in the portion of vocabulary which is common Chamic, which is reconstructible back to Proto-Chamic and which appears on the Blust list. This percentage is surprisingly large for such a loan stratum which can be found in a securely-reconstructed proto-language.

Furthermore a considerable proportion of the lexicon of any Chamic language (and this is a stratum which is less well-represented in the most basic lexicon, but certainly far from absent even here) is made up of forms which have not been properly etymologised, but which have no cognates in any Austronesian languages (nor yet have clear etyma for these any been found in Mon-Khmer languages). But at the same time these very words often possess certain surface phonological characteristics, such as implosive stops or particular vocalic nuclei, which are typical of Mon-Khmer elements in Chamic languages but which are rarely if ever found in items belonging to the slender yet genetic Austronesian stratum in Chamic. The presence of such phonological features in these items suggests that these words entered Chamic languages either at or some time after the period of intense Chamic contact with Mon-Khmer languages, and after the rise to prominence of the monosyllabic contentive. There is a small but nonetheless significant stratum, smaller than that deriving from Mon-Khmer languages, of forms which are reconstructible to proto-Chamic and which are also found on the Blust list.

A considerable proportion of common free grammatical morphs in Chamic languages are as yet of uncertain origin (and a number of these are common to Acehnese and other Chamic languages, so that they must reconstruct back to Proto-Chamic), and some others derive from Mon-Khmer languages. This latter group of forms comprises both

those forms which are common to all or most Chamic languages, and a later but sizeable number of free elements, for instance certain negators and some modal verbs, which have been borrowed into individual Chamic languages from Khmer (in the case of Western Cham) or Vietnamese (in the case of all the Indochinese Chamic languages spoken in Vietnam) since the split up of the Chamic languages about two millennia ago. The incursion of Vietnamese and Khmer elements into Chamic languages is apparently a matter of only a few centuries' age. Yet other such free morphs, including the numerals, are inherited from Proto-Malayo-Chamic (and some higher numerals have been diffused from Chamic into neighbouring Mon-Khmer languages).

Many of the same Bahnaric elements are common to all Chamic languages and therefore reconstruct to Proto-Chamic, into which they are loans, and this much could be demonstrated many times over by the employment of Venn diagrams or by using other demonstrations of the principles of set theory. A second set of Mon-Khmer forms is found in most or all Indochinese Chamic languages (that is, all save Acehnese and Tsat). Having split from the other languages more than a millennium ago, having lost all contact with other Chamic languages and with its speakers having migrated to Sumatra by way of eastern Malaya, Acehnese contains elements from Aslian Mon-Khmer languages, which were probably dominant in that part of Malaya at the time, but it contains an especially large number of loans from Malay (these including some forms which have replaced the more characteristic and inherited Proto-Chamic forms and which therefore count as instances of relexification), Sanskrit and Arabic. There are also (fide Thurgood 1999) a number of loans in Acehnese which derive from Katuic languages, and which are not found elsewhere in Chamic languages.[11]

The other migrant Chamic language, that is Tsat, contains a few stray elements of Hlai (a pre-Chinese language group of Hainan which is distantly related to the Tai languages) and many more from Chinese languages. Thurgood and Li (to appear) note the presence of four layers of loans into Tsat. Chronologically the first layer derived from Hlai. The second layer was taken from Hainanese Min Chinese, the third layer was taken in the course of the 20th century from the Mandarin spoken by the military personnel who were settled near the Tsat villages, and which was acquired by Tsat from contact of its speakers with these personnel, and the fourth and most pervasive layer derives from standard Mandarin (Putonghua) as taught in all Chinese schools. This final layer has wrought strong typological changes upon Tsat (some of which are exemplified in Thurgood and Li 2003), though Tsat may already have developed a tone system even under influence from the multitonal Hlai. This influence has been actuated by the spread of universal Putonghua-medium state-education among the Tsat rather than by intermarriage with native speakers of Putonghua, since Tsat speakers are endogamous Muslims whereas Han Chinese are, officially at least, atheist and therefore Chinese men at least are not permitted to marry Muslims. (The incursion of Vietnamese and Khmer elements into the lexica of the Indochinese Chamic languages rather unsurprisingly postdates the separation of Tsat and Acehnese from other Chamic languages, since the lexica of Acehnese and Tsat contain no unambiguously Khmer or Vietic forms.)

Haroi has borrowed heavily from Bahnar and Hrê, both of them being Bahnaric languages, Haroi and Bahnar have both developed restructured register, and Haroi-

[11] Paul Sidwell (personal communication, April 2003) assures me that earlier claims that there is a pan-Chamic component that is of exclusively Katuic origin are largely incorrect.

speakers have most in common culturally with speakers of Bahnar – indeed Haroi culture *is* Bahnar culture, and the Haroi have sometimes been known as the 'Bahnar Cham'. Western Cham contains added elements from Khmer, which are not found in Eastern Cham, while other Chamic languages have borrowed heavily from Vietnamese, and through this have recently acquired elements originally from French and English. Cham in both its modern forms (and additionally in the traditional written form) also contains a number of elements from Malay, since all Western Chams and many Eastern Chams are Muslims who used Malay as a liturgical language after their conversion to Islam. These are not usually to be mistaken for inherited Malayo-Chamic elements, however, because of the semantic fields in which these Malay borrowings enter (namely religious and similar cultural considerations). Jarai, Rade and Northern Roglai do not appear to have been especially adlexified or even relexified by the absorption of innumerable words from neighbouring Mon-Khmer languages; the main source of new words in these languages is Vietnamese. Rade had had some role as a lingua franca in part of the Vietnamese highlands (Tharp 1980) and may have been a donor language to some (Mon-Khmer) languages rather than being a recipient language.

All the forms which are of Malayo-Polynesian origin and which occur in Chamic languages have either been inherited from Proto-Malayo-Chamic, which is the common ancestor of Malay languages and Chamic languages (and have sometimes subsequently been lost in Malay), or else they are secondary possessions. More specifically, they are borrowings into these languages from Malay, and therefore are loans but not true cognates. In addition some forms which derive from Proto-Malayo-Chamic are found in Malay and in Chamic languages but cannot be reconstructed further back, which suggests that they are innovations within Proto-Malayo-Chamic. (A couple of dozen Malayo-Chamic lexical innovations are given in Blust 1992.) The vast majority of forms on the Blust list which have been retained in Malay lects are also found in Chamic languages, and vice versa. Together Malayic and Chamic have retained some 60% of the 200-item Blust list PMP reconstructions, and the bulk of these retentions are found both in Malayic (which retains 116 of the 200 forms) and Chamic, as indeed are most of the small set of phonologically modified retentions, such as *kaki* 'leg, foot' from PMP **qaqaɣ*.

Despite some superficial coincidental similarities, there is absolutely no lexical or other linguistic evidence in the inherited component of Chamic languages to suggest that Chamic languages subgroup especially closely with Formosan languages, or with one or another subset of Philippine languages, much less with Oceanic or other Central or Eastern Malayo-Polynesian languages. Such similarities in phonological developments as we sometimes find occurring between Chamic and (say) Oceanic are coincidental and of independent development, and do not indicate a special non-trivial historical relationship or period of shared development.

As far as we are aware, Malay has not borrowed any items from Acehnese or Chamic languages, though Vietnamese (though to a very slight extent) and some other Mon-Khmer languages have done so; for instance the Vietnamese word for 'island' *cù lao* is probably a loan from Cham *pulaw*. (Malay *pulau* is also a possibility as a source, though.) The phonological form of the Vietnamese word shows that it was probably borrowed at a time before /p-/ was a permissible or legitimate syllable-initial consonant in Vietnamese (where original /p/ had apparently changed into /f-/), as it was to become in the 19th century with the incursion of borrowings, especially nouns, from French.

3. The use of the Blust list for historical explorations in Proto-Malayo-Polynesian, Proto-Malayic and Chamic languages: aims and operations.

Given the primary consideration - or the primary obstacle - that bound inflectional morphology, which is the kind of evidence which is most prized by diachronists who are attempting to prove the genetic relationship of two or more languages, is almost absent in the Chamic languages, and that much of what little bound morphology there is appears to be borrowed, the best that we can do is to explore some of the possibilities inherent in comparing basic lexicons. (If there were more inflectional morphology available for us to work with, then we would give that part of the languages preferential treatment in a study like this.)

The 200-item Blust list is well-suited to the purpose of comparing basic portions of these languages, although longer lists could also be used and these would tell us even more about the history (and especially the external history) of Chamic languages. It should be noted that evidence from the Chamic languages and Acehnese played little or no part in the original elaboration of the Proto-Malayo-Polynesian forms which are displayed in Blust (1993), so that these can be analysed objectively using this method. There is no risk of circular reasoning in this regard.

What we are trying to do is to see what can be gained from employing a combination of several techniques which are being employed sequentially, in order to put into practice a kind of multilateral comparison. We are employing lexicostatistics (though not glottochronology), and we are referencing each entry to its occurrence or non-occurrence on the equivalent wordlist for the norm which we are using (in this case the norm being used is Blust's reconstruction of Proto-Malayo-Polynesian, against which the reflexes of the glosses in the various modern Chamic languages are mapped). What is more, we are indicating the cognacy of each item to the norm or to other equivalents by the alphabetic code which was detailed above in the discussion of Wick Miller's work.

In this case, though, I am not using 'x' as a marker of lexical singularity; instead I am giving a separate letter to every discrete form, whether it is unique to one language or used among two, more or among all the sampled languages. Whichever gaps remain after my strenuous and studious attempt to fill them will be marked with 0, and loans, which in the case of the Chamic languages are mostly from Mon-Khmer languages (while in Malay they are mostly from Arabic or Sanskrit), will be indicated in a special column at the right of the table. I am comparing the cognacy of these Chamic forms (including the Acehnese forms), wherever possible, with the equivalent forms in Proto-Malayo-Polynesian as reconstructed by Blust, and with those in standard Malay, in an attempt to derive a more nuanced picture of the interrelationships within Chamic. Where possible, plausible loans into Acehnese from Malay (for instance, those which do not follow the sound correspondences that have been drawn up as obtaining between Proto-Chamic and Acehnese in Thurgood 1999 but which are nonetheless similar in shape to elements to be found in Malay) are also indicated. This is because these forms do not count as proper cognates but need to be recognised, somewhat paradoxically, as 'non-cognate' because they have entered Acehnese from Malay as loans.

I have also, for the sake of interest, sampled and surveyed a few further forms across Chamic languages, over and above the Blust list gloss forms. All of these are taken from the traditional Swadesh lists. These spare forms are 'to sing', 'five' (which in Malayo-Chamic languages is generally distinct from the form for 'hand'), and 'to play' (a form which I selected specifically as it is one of the most lexically diverse forms or

'characters' in Indo-European: Ringe at al. 2002. It certainly does not share that distinction in Chamic, since most Chamic languages use a reflex of PMP *maqin). I have also collected both the inclusive and exclusive forms of the pronoun 'we'; this last is a distinction that is of Proto-Austronesian vintage, and one which is perpetuated in very many of the modern languages, including the Chamic ones apart from Tsat (Mon-Khmer languages often make this distinction too whereas Chinese does not, and this typological parallel may account for its preservation in most Chamic languages). The inclusive 1pl form is represented as item 204. (The items numbered above 200 have not been further included in my calculations, though the distribution of forms within them and the variety of forms to express them within Chamic are facts duly noted.)

There are certain considerations and certain desiderata to be taken note of when using the Blust list in this operation. The desiderata constitute the aims and objectives of this project. I wished to see whether there was a valid Malayo-Chamic grouping within Malayo-Polynesian. I also wanted to see whether Chamic constituted a subgroup on its own, whether subgroups within Chamic could be identified and defined on the basis of lexical evidence, and where Acehnese fitted into all of this (and an important if secondary consideration was the extent to which Acehnese basic lexicon might have been influenced by later contact with Malay). I was also interested in seeing whether there were any PMP forms that were still preserved in Chamic which were not to be found in Malay, and if there was such a set of forms, I wanted to attempt to see why they were missing from Malay – had they been replaced in Malay by internal creations, or by external diffusions (lexical borrowings)?

In presenting my findings in the table I have started off with providing a code for the Proto-Malayo-Polynesian forms, which are uniformly logged here as 'a', because they come first on the chart. Next to the right come the letters indicating the cognacy or otherwise of these forms with those for Standard Malay (with loans into all languages indicated where known), and after this follows the column for the cognacy firstly with PMP, and secondarily with Malay, of the equivalent forms in Acehnese. I have continued to do this for the equivalents in several other Chamic languages: Western and Phan Rang Cham, Jarai, Rade, Northern Roglai, Tsat, Haroi and Chru. The sample of languages which I have surveyed is purposely limited, not least because of space constraints, and I have not provided comparable lexicostatistical information on other potentially interesting and relevant Western Malayo-Polynesian languages such as Madurese, Javanese or Tagalog, most of which, incidentally, appear to have preserved fewer of the 200 PMP forms in the Blust list than Phan Rang Cham has.

An important consideration in this study is the relative availability of the relevant kinds of lexical data. My sources were fullest for Phan Rang Cham (Moussay 1971), Acehnese (Daud and Durie 1997), Rade (Tharp et al 1980, also Egerod 1978 and Shintani 1981) and Northern Roglai (for which I used the list in Collins 1969 and some data from Bochet and Dournes 1953), and I have all the forms available for the lists for Jarai (Lafont 1968) and for Western Cham and Tsat as well, the latter thanks to the kindness of Robert K. Headley and Graham Thurgood respectively. Lexical data on Haroi were taken from Thurgood (1999) and Tegenhardt-Mundhenk and Goschnick (1977), and those for Chru come from Thurgood's book, from Fuller (1977) and also from Tín (1955), which also provides a Jarai glossary, together with lists in French and (the language of alphabetisation, and the source of the orthography for entries in Jarai and Chru) also Vietnamese. Thurgood's book was the main source for my data on Western Cham, together with papers

in Thomas (ed., 1977, 1997) and Headley (1991), plus a few forms which Robert K. Headley gave me in personal communication, while for Tsat I used Zheng (1997, a source which was unavailable to Thurgood at the time of writing his book) plus one datum from Benedict (1941) for a single Tsat form which I was unable to find in Zheng's book.

The table could have been fuller. But I have reluctantly omitted a column of forms from the earlier stage of Written Cham (which shares a very high degree of lexical similarity with the two modern Cham languages) because I have too many gaps in my data, and I have available even fewer forms which are attested for Inscriptional Cham or Old Cham. I have also desisted from including a column of Proto-Chamic forms, whether they be those reconstructed by Lee or Thurgood, because I feel that an analysis of the distribution of particular forms across individual Chamic languages is the best first step towards reconstructing this portion of Proto-Chamic lexicon. It should be noted that the data which I use in this study have almost all been gathered by investigators working within the last 50 years or less, so that this exercise is a comparison of materials of roughly contemporary vintage. Many further forms that were not otherwise available to me were graciously provided by Graham Thurgood in personal communications.

The lexical material in Thurgood's book was the starting point for this work, and the basis and source for most of the entries on the grid. Since Thurgood's concerns there are primarily comparative rather than purely descriptive, it means that certain lexical forms which occur only in one Chamic language or which are not otherwise historically interesting are not going to be listed in his lexical lists, no matter how high the forms' text frequency may be. Such forms would include for instance the so-called 'lexical orphans' which may have been present in the proto-language but which are attested only in one modern daughter language. Others would be forms which have developed independently, which are unique to a particular language and are recorded for no other, or alternatively Mon-Khmer or other borrowings which no other Chamic language has taken up. On the other hand, it is unlikely that very many comparative Chamic cognate forms which are essential to this study, especially those of Austronesian origin, cannot be found in the lists in Thurgood's book, just as long as the forms for the relevant English glosses have been included in his lexical appendices in the first place.

We should also remember that, Phan Rang Cham apart (see the dictionary by Bui Khanh The 1995), we do not have voluminous lexica for any Chamic language of the kind which is available for Malay, and that indeed it may be the case some forms whose presence is alluded to in the table may have widely-known cognates in other Chamic languages, but that these cognates have simply not come to our attention because they are not noted in the available literature. All the columns in the table are at least 85% complete; the one with the most gaps is the Haroi. By contrast, the columns for Phan Rang Cham, Western Cham, Rade, Jarai, Northern Roglai, Chru and Tsat, and of course those for Malay and Acehnese, are complete and most of the rest are nearly so. Gaps in the Chamic lexical data which are available to me at the moment are infuriating, as they always are, but here they are not serious enough to distort or impugn the validity of the use of the particular methodology employed and the overall findings of this study.

4. Identifying some problems in Chamic lexicostatistical investigation.

Another consideration in this study was the suitability of the Blust list as manifested in the problems inherent in getting good forms for glosses. The semantic spaces of Chamic languages and that which is assumed for reconstructed PMP did not always coincide, although I did not substitute any of Blust's forms. Decisions sometimes had to be made as to what kinds of 'cutting' (chopping, hewing, splitting, slicing, etc.), 'lying down' (full length, prone or supine), throwing' (hurling, releasing an arrow discarding, throwing underarm as distinct to overarm throwing, or whatever) or 'turning' (spinning, revolving, flipping over, all of these either intransitive or transitive) were involved. There is also the issue of whether 'to smell' is intransitive (in which case the form is most likely Proto-Malayo-Polynesian) or transitive, in which case one chooses between a form meaning 'to sniff, snuffle' from PMP, or 'to sniff at, to kiss' from Mon-Khmer (but reflected also in Malay[12]). Furthermore, the semantic distinction between 'long in distance' and 'far', which is retained in Malay, seemed to be encapsulated in the same word in some (though not all) Chamic languages, while the distinction between 'wide/broad' and 'thick' does not seem to be made lexically in all Chamic varieties.

One or two forms are apparently compounds involving one or more forms which are attested elsewhere on the list. This is the case with the form for 'lake' in some language ('big water'), while in some languages 'to kill' is expressed by a form analysable as 'CAUSATIVE-to.die', thus involving a form which had already been found on the list. Furthermore, one or two forms which reconstruct to PMP are still recorded both for Chamic and Malay, but have developed new semantics in both languages. For instance the widespread Proto-Austronesian and PMP stem **qulu* 'head', which is realised as *hulu* in Malay, has been replaced by a Mon-Khmer loan, namely *'akó'*, in the whole of Chamic and by a Sanskrit loan in Malay (and for that matter in Khmer), at least as far as the name of the anatomical part is concerned. Yet it still occurs in certain kinds of compounds in both languages (and it is used as 'head' in most metaphorical senses in Malay). For instance there is Chamic *dihlau*, Malay *d(ih)ulu,* both of these being forms with the meaning 'formerly', literally 'at+head' in Proto-Malayo-Polynesian (PMP **di + qulu*). Malay *dulu* has subsequently been grammaticalised as an indicator of completive aspect.

Other distinctions which are less easy to capture using the Blust list are those which involve pronouns, especially personal and interrogative ones. In Chamic languages the interrogative pronouns are often bimorphemic words involving a general interrogative form and a specifier which indicates such a sense as 'place', 'time', 'manner' or whatever. Consequently the same morpheme occurs in several glosses on this list, and this replication of the same interrogative morph happens in several Chamic languages. The Blust list glosses provide for only two demonstrative positions, namely proximal and distal, yet many of the languages here have at least three such forms in both pronouns and adverbs. As to personal pronouns, the Blust list assumes a system which involves a two-way distinction of number and a three-way distinction of person, without special reference being made to a distinction between inclusive and exclusive senses of 'we'. The system in Chamic languages is rather different. Except in the first person plural (where an inclusive/exclusive distinction is regularly made), number in pronouns is of secondary importance, although three persons are regularly distinguished. The primary division in

[12] A catalogue and analysis of the Mon-Khmer component in most or all varieties of Malay, which is not massive but not negligible either, is long overdue.

most Chamic languages is between polite or formal versus ordinary pronouns – and this is a distinction which is by no means unknown in Southeast Asia. In addition, the ordinary word for 'I' in Chamic languages is the normal PMP one, whereas the formal word for 'I' is derived from the PMP form meaning 'slave', thereby perpetuating a trope which is also found (inter alia) in Vietnamese *tôi* and Malay *sa(ha)ya* (this last being a borrowing from Sanskrit).

An exception to this patterning is provided by Tsat, which has developed new 2pl and 3pl pronouns by combining the relevant singular pronouns with a suffixed *za:ng,* a Malayo-Chamic form meaning 'person' (cf. Malay *orang,* Phan Rang Cham *uraang* 'person'). This is exactly what many forms of Min Chinese (including Hokkien and Hainanese) have done. Coincidentally it is also what has happened in those varieties of Malay which have also been in touch with Hokkien, or which have developed at a later date from such varieties, such as Betawi of Jakarta in the first instance, and Baba Malay, Sri Lanka Malay and Cocos Malay in the second instance (each of which are developments from Betawi; the observations are based on personal communication from Graham Thurgood in the first case, and Adelaar 1991, 1996 for Sri Lanka Malay and Cocos Malay) in the second. Such dialects have, for instance, *dia orang* '3sg-person' for 'they'. Although some speakers of Tsat have contact with formal Malay through Islamic teaching, we cannot assume that this particular structural parallelism has developed or been percolated through the effects of Tsat contact with Malay, because this construction is not typical of the particular formal Malay lect to which Tsat speakers have been exposed through religious work, which would use the Standard Malay 3pl personal pronoun *mereka.* Instead, what we have here is the development of the same structure within a pronominal system as the result of influence from the same kind of Chinese language upon two related languages, but we see that it developed separately in two areas and on two occasions where the same kind of language (in other words, a Western Malayo-Polynesian one) happened to have been influenced by varieties of Min Chinese.

The results of this investigation are presented in the table below. What then are the outcomes of this experiment? We can imagine a set of outcomes each being displayed on a number of occasions in the result in the table. The first outcome shows the Chamic languages retaining forms inherited from PMP. The second shows them retaining forms inherited from Proto-Malayo-Chamic, in which these forms had developed. The third outcome shows Acehnese having the same form for a particular gloss as other Chamic languages do, but Malay differing from these (and maybe also from PMP). The fourth scenario would have the Indochinese Chamic languages (with or without Tsat) sharing forms which are not also found in Acehnese, and which we assume to have developed at a time when Acehnese had split away from the other Chamic languages, which were all still in contact with one another and which were in a position to diffuse items to one another. Some of these innovations may be loans, as may some forms which bind Acehnese and other Chamic languages together against Malay and PMP. Another series of outcomes would indicate the development of subgroupings within Chamic, say a Jarai-Rade subgroup, which are marked out by the development of shared lexical innovations (including loans), which have replaced forms which have otherwise been conserved in other varieties. Another set of outcomes would show Tsat as being either conservative or, more probably, especially lexically innovative (as the result of borrowings) against the consensus of the evidence of the Indochinese Chamic languages. If it had conserved forms whereas Indochinese Chamic languages had all shared in the introduction of an innovated

form, this might have some historical significance. And there is the possibility that for a certain period of time all the Chamic languages had gone their separate ways and were still doing so (though latterly many were following some of the same paths of conformity as a result of sharing cultural borrowings from Vietnamese, which was the language of power in most or all the communities under discussion). We can find instances of all of these scenarios in the table below, although I should point out that the direct impact of Vietnamese on the contents of the basic Blust list lexica of any of these languages is negligible.

5. Some observations on the results.

How does the use of this bundle of techniques work out in practice? What can we learn from its application? For a start, the rows of straight 'a's which run through Malay, Acehnese and the other Chamic languages (with occasional interruptions due to lexical replacement in one or more languages) show that the Austronesian (and more certainly the Malayo-Polynesian) affinities of Chamic languages are manifested very clearly in the lexicon (and in most of what remained of the inflectional morphology of these languages). Indeed there are even a few cases in which the lexicon of modern Malay has replaced or shed a preexisting PMP form, which has nonetheless been retained in Chamic languages (and in these instances sometimes Acehnese has borrowed the Malay form, while on other occasions it has retained the same inherited form as the Chamic languages). This is the case, for instance, with the word meaning 'to bathe'.

These instances of lexical replacements of inherited forms will have occurred at some time after the split of Chamic and Malayic, a split which occurred a few centuries before the birth of Christ. In this respect it is significant that some of the forms which have been lost in Malay have been replaced there by words which derive from languages with which Proto-Malayo-Chamic could not have been in contact, namely Sanskrit and Arabic.[13]

In an analysis of the items entered on the grids I counted 108 forms (out of 200 glosses), occurring in one or all the Chamic languages (Acehnese apart) which trace back to Proto-Malayo-Polynesian, and three of these forms consistently show phonological irregularities which accord with those for the same cognate forms in Malay[14], giving some credence to the establishment of a special Malayo-Chamic group. As far as I can tell, all but one of these forms (the exception is the form for 'flesh, meat' deriving from PMP **hesi*) also occur in Phan Rang Cham, while one further inherited PMP form (the reflex of PMP **naɲuy* 'to bathe', above) also occurs only in Acehnese but has been replaced by other forms in the remaining Chamic languages.[15] In addition, there are a number of forms

[13] The lexica of the Chamic languages have provided minimal evidence for the reconstruction of Proto-Austronesian and Proto-Malayo-Polynesian, so that they have not been explored much, and indeed there are rather few inherited forms which are preserved in Chamic languages which cannot also be found in Malayic lects.

[14] For instance they may be stems which in both sets of languages incorporate the form of an infix (present at PMP level, but obsolete as a productive morphological device by the time of the first attestations of Malay in the late 7th century AD) into a newly metanalysed stem. The form for 'to dream' is an example of this.

[15] Compare this tightness of bunching with the situation in Oceanic, in which the vast majority of forms which have been reconstructed for the Blust list for PMP are attested as inherited forms in at least one Oceanic language (and the Samoan list has almost half of these, involving 84 of the 200 forms reconstructed in the list for PMP and an even greater proportion of those

(I counted 30 such) which do not occur in PMP but which also occur in Malay as well as in Chamic languages, and the existence of this cluster of lexical innovations bolsters the claims for a Malayo-Chamic group too.[16] 22 forms on the list are shared between Acehnese and some or all of the other Chamic languages, but do not occur in Malay or in PMP, although several of these are loans from Mon-Khmer languages rather than being innovated forms that were first generated at the Common Chamic level. Still, they strengthen the evidence for a historical genetic link between Acehnese and the Indochinese (and Tsat) Chamic languages (while in one further case, Acehnese and Tsat have preserved a PMP form which has been lost in Indochinese Chamic).

By contrast, at least 44 Chamic forms, many of them pan-Chamic, certainly or probably derive from a Mon-Khmer language. Another pan-Chamic form ('dust') derives from Sanskrit by way of its having previously been borrowed into Mon-Khmer languages such as Khmer, and at least 10 further glosses have equivalents which are pan-Chamic in spread, but for which an origin has yet to be found in any known language. Meanwhile 2 further Blust list glosses are variously expressed in Chamic languages, sometimes being expressed by Mon-Khmer forms and sometimes by widespread forms, which are found in several Chamic languages, and which are of unknown origin.

---

reconstructible back to Proto-Oceanic), but where most Oceanic languages lack most of these forms, while some of the forms which can be reconstructed back are found only in a few Oceanic languages.

[16] These 108 forms inherited from Proto-Malayo-Polynesian constitute an unknown but certainly high proportion (at least one third of such forms and probably much higher) of the total of such forms which any or all Chamic languages have inherited from their ultimate genetic ancestor (Proto-Austronesian) or which have been acquired from this ancestor's descendants (Proto-Malayo-Polynesian, Proto-Malayo-Chamic) which are nonetheless antecedents of Proto-Chamic. By comparison, on the Blust list for Standard Malay 112 items out of 200 derive from PMP – and this is the highest proportion of such forms which has so far been recorded for any Malayo-Polynesian language (Blust 1990). This compares with 89/200 retained PMP forms for the Blust list for Chamorro, for example, and with a miserable 10 retained PMP forms out of 194 attested equivalents in the Blust list for Kaulong of New Britain, a language whose Austronesian affinities (within the Pasismanua languages of the Oceanic branch) have never been in doubt. This last figure is less than ¼ the number of *attested Mon-Khmer loans* which are to be found among the Chamic-language equivalents of the Blust list! (The estimate of Headley 1976, to the effect that Mon-Khmer loans accounted for about a tenth of the reconstructed Proto-Chamic lexicon, is set too low.) 3 further forms deriving from PMP, which are replaced in Standard Malay by loans from other languages, occur on Blust lists for some non-standard Malay varieties. According to my calculations the comparable score for Acehnese is 110/200, though some of these 'inherited forms' may actually be unrecognised Malay back-borrowings into Acehnese. The bulk of the recognised Mon-Khmer elements in the Acehnese list are also found in other Chamic languages and reconstruct back to Proto-Chamic, and this is also true of some of the forms which are as yet of 'unknown' origin. As speakers of Acehnese never returned to Champa, the presence of such forms in Acehnese can only be explained by reference to a previous period of common development between Chamic languages and Acehnese, during which contact with Mon-Khmer languages occurred, leading to lexical transfer. On the Blust list some 7 forms which are of PMP origin but which are not recorded in Malay are attested in at least one Chamic language; in Malay these have either been replaced by loans (*nama* 'name' from Sanskrit, expressed in Jakartanese by the Javanese loan *ngaran*, a form which is cognate to the lost Proto-Malay form) or by innovated forms. Blust (1981a) provides scores for two Chamic languages; according to his calculations he assesses Acehnese as retaining 81 items out of 200 and Jarai as retaining 73 out of 200.

What is most striking is that the Chamic languages show a very high degree of lexical similarity and internal lexical homogeneity, no matter what the origin of the individual lexical items may be, and this is especially true when Acehnese and Tsat are removed from the picture. I found 162 instances out of 200 in which either all the Chamic languages in the table from Phan Rang Cham onwards shared the same form (of whatever etymological origin), or else all these languages bar one for which I had an attestation of a gloss for the particular form used forms of the same origin. Proto-Chamic equivalents, which are reconstructible at least to the period after which Acehnese had split off from the other languages and often much further back, could be reconstructed for at least 85% of the items on the Blust list simply by using traditional methods and by drawing upon traditional kinds of evidence for proving the existence and shape of lexical reconstructions.

In addition to the forms on the Blust list which go back to PMP and for which reflexes can be found in at least one Chamic language, there are 13 further forms which are post-PMP but which are found in Malay and in Indochinese Chamic languages as well as in Acehnese, so that they reconstruct back to Proto-Malayo-Chamic. There are 19 further forms which are common to Acehnese and other Chamic languages but which do not occur outside this subgroup so that they are not found in Malay, and there are at least 45 further forms which are common to two or more Chamic languages outside of Acehnese, and many of these have Mon-Khmer etymologies, as have some of the 19 forms which are common to Acehnese and other Chamic varieties. Indeed the stratum of forms of Mon-Khmer origin which are found in all Chamic languages is bigger than the stratum of common Malayo-Chamic innovations, and the stratum of forms of common Chamic heritage but of unknown origin is also broader than that. A few further forms probably reconstruct back to Proto-Chamic on the basis of their widespread distribution in modern Chamic languages, but they are not found in Cham proper or in Acehnese. And it is possible that these numbers are themselves underestimated, and that the lexical uniformity within Chamic (and especially within those varieties still spoken in Indochina) may be greater than these suggest. By contrast, there are few forms on the Blust list which it would be almost impossible to reconstruct using the judicious application of standard historical linguistic methods. But there are also some glosses (the verb 'to throw' being an especially vivid example, in Chamic as in many other language families) which exhibit a very large number of different forms among the dataset for this form for the modern Chamic languages. Indeed, if we had data for Haroi forms meaning 'to throw', it is likely that there would be more than the six separate forms listed which have so far been attested for the Chamic languages surveyed (let alone the other forms that have been noted for PMP, Malay and Acehnese, which all differ from one another). But we need to be mindful of the fact that 'to throw' is one of those forms for which many languages have more than one equivalent, depending upon the nature of the item thrown, the trajectory of the throwing action, the question of whether the item projected hits its target or not and so on. We cannot blandly assume that the semantic range covered by any, most or all of the forms meaning 'to throw' in the various Chamic languages is identical in any or all the languages.

To some extent this widespread core lexical similarity within Chamic languages is a continuation of the manifestation of other clear cognacies. It is beyond doubt that the various Malay lects and Chamic lects subgroup with one another against other Malayo-Polynesian languages, and that they share some common and irregular developments of inherited forms which other MP languages do not. It is beyond doubt that Acehnese and

Chamic fit together in a subgroup against Malay and with one another (we may state this securely despite the presence of a few high-frequency Malay loans in Acehnese, though there are none to be found in corresponding lexical strata in Chamic. This is unless the word for 'green' in Phan Rang Cham and Western Cham is an unrecognised borrowing from Malay rather than a retention from Proto-Malayo-Chamic, in which group it would be an innovation against the inherited PMP form). It is clear that Acehnese has gone its own way in matters of lexical change, loss and replacement for a time, at least before being 'swamped with Malay loans' (Blust 2000a)[17], and it is clear that Tsat fits in lexically with Chamic despite the wide typological and the lesser lexical differences between Tsat and even Northern Roglai, its probable closest genetic relative – in which Tsat is the innovating partner. The presence in the Tsat lexicon of a subset of the same Mon-Khmer-derived loans which one finds in Northern Roglai (and which also generally occur in other Chamic languages) is highly significant here as an indicator of Tsat's Chamic affinities and origins.

What makes this considerable lexical uniformity within Chamic (a uniformity which is somewhat underplayed by the lexicostatistical results presented in Tables 2a and 2b) so remarkable is the fact that it is accompanied by an impressive degree of contact-induced phonological diversity from one Chamic language to the next. (There is less internal diversity in regard to Chamic morphological systems, apart from the conservative features of Acehnese morphology.) It is highly unlikely that Rade and Jarai, or Haroi and Phan Rang Cham, or whatever, are mutually intelligible, despite the similarities of their basic lexica, and much of this is due to the different outcomes of each language's reflexes from Proto-Chamic.

It is fitting that Thurgood had to reconstruct Proto-Chamic phonology from the top down (and also from the bottom up), working from hypothesised Proto-Chamic forms which more often than not bore a strong resemblance to those which are still found in more conservative dialects of Malay. This is because any attempt at reconstructing Proto-Chamic simply by working from the bottom upwards, using only the evidence of the modern Chamic languages (even if Tsat were excluded and if Acehnese data were mined solely for their conservative features) would have made the task immeasurably more difficult. This is especially true of anything affecting the reconstruction of the shapes of presyllables. It is also true that numerous phonological irregularities remain in the forms of Thurgood's Proto-Chamic reconstructions; we simply do not know everything about the phonological history of Chamic languages. Many loose ends still remain at the level of the reconciliation of troublesome facts about individual word histories in these languages (for instance the wide range of disparate and 'irregular' word-final consonants and vowels which Thurgood lists for many of his Proto-Chamic reconstructions).

The degree of morphological diversity among Chamic languages, especially as far as the use of bound inflectional morphology is concerned, is less than that which is found

[17] Part of this lexical self-direction on the part of Acehnese has involved the absorption of Mon-Khmer loans (presumably Aslian ones from languages of eastern Malaya, but maybe also some further Katuic ones) which are not found in other Chamic languages. These loan strata have yet to be identified or worked upon fully, although a good place to start would be among the large number of monosyllabic contentives found in Acehnese which have no PMP, Malay or Chamic parallels. Further attention also needs to be paid to the Mon-Khmer lexical stratum in Malay, which is not inconsiderable in size or in centrality to the everyday Malay vocabulary (though it is especially rich as a source of ecological terms), but which has yet even to be listed comprehensively.

in the Chamic segmental, canonical and other phonological systems, but this apparent uniformity is largely a result of the overall paucity of inflectional morphology in these languages to begin with. The Chamic language which stands out the most from the others in the realm of morphology is Acehnese, which looks incongruous when it is compared with other Chamic languages or with Malay. But this is because in many respects (for example in its possession and productive use of verbal infixation or derivation) Acehnese has been conservative, and as such resembles non-Malayo-Chamic but nonetheless Western Malayo-Polynesian languages such as Ilokano and Tagalog, whereas Malay and the Chamic languages have innovated in shedding this morphology.

The loss of productive use of inflections is a process which has happened extensively but separately in Chamic and in Malay; it naturally dates after the split-up of Malayic and Chamic, and occurred under separate sets of social circumstances and as the result of intense contact from different sets of languages. (Thurgood 1999: 43 finds another structural parallel of this post-split typological difference between very closely related languages within the realm of Western Malayo-Polynesian. He points out that Malagasy, which historically and genetically is a Bornean language of the Southeast Barito subgroup which was removed c. 400 AD from that island to Africa and thereby from the full-scale morphological effects of intensive and submissive contact with Malay[18], preserved the inherited morphological feature of infixation. In contrast, Malagasy's unrelocated relatives among the Southeast Barito languages of Borneo, that is to say languages such as Ma'anyan which were all much more heavily exposed to direct and continued influence from Malay than Malagasy was, eventually lost their infixes and simplified their morphology). A table illustrating this situation and comparing the morphological typologies of a number of relevant South East Asian languages can be found as Table 3.

Some representative scores for the percentage and number of shared forms (of whatever origin) between particular pairs of Chamic languages include the following sets of results:

Malay/Acehnese: 135 items out of the 199 discrete forms which were recorded in the available data (although 4 of the shared forms may actually be loans from Malay into Acehnese),

Phan Rang Cham (henceforth PRC)/PMP: 107/198[19];
PRC/Acehnese: 133/198;
PRC/Standard Malay: 102/198;
PRC/Western Cham: 194/197;
PRC/Jarai: 177/198;
PRC/Northern Roglai: 177/198,
PRC/Tsat 171/198,
PRC/Rade: 168/198,
PRC/Haroi: 171/182,

[18] The lexical impact of Malay upon Malagasy, though, could be found in some surprisingly basic realms of vocabulary, for instance the names given to body-parts (a number of such examples are given in Adelaar 1995).

[19] Despite the fact that we have complete 200-item Blust lists for PMP, Malay, Achenese, PRC, Jarai, Rade, Chru, Tsat and Northern Roglai, the numbers of compared forms add up only to 199 (where Acehnese is involved) or 198 (if any other Chamic language is involved) because of the duplication of certain stems in the system of plural pronouns; we cannot count the same stem twice.

PRC/Chru 187/198.
Western Cham/Malay: 102/198;
Western Cham/Acehnese: 131/198;
Western Cham/Haroi: 158/180;
Jarai/Rade: 175/198;
Rade/Northern Roglai: 173/198;
Haroi/Chru: 178/180;
Chru/Jarai: 180.198;
and Tsat/Northern Roglai: 143/198.[20]

There are also 89 forms out of 200 which meet two conditions: they are shared by Malay and PRC, and they reconstruct back to PMP. 3 such forms show Chamic phonological modification of the original PMP form in a way which is also shared with the cognate form in Malay (for instance we have Acehnese *lumpeuy*, Malay ***men-impi***, PMP *h-in-ipi*, an infixed form of earlier PMP **hipi*, all of these meaning 'to dream'; the Malay form involves the addition of a modern Malay prefix to a stem which includes an infix which is no longer identifiable as such to modern Malay speakers), and 30 forms are Common Malayo-Chamic, inasmuch as they are found in Chamic, and Malay, and sometimes Acehnese, but are not among the PMP forms. I have used PRC data here in this comparison since this is the Indochinese Chamic variety for which my data were fullest and clearest at the time when I first did my calculations. In addition it is the Chamic variety which has strayed the least geographical distance from the historic centre of Champa.

By contrast, there are at least 21 unique items (items with no cognates in any other Chamic language or elsewhere) out of the 200 Tsat forms which were available to me for completion of the Blust list, though rather few of these unique items are taken from a Chinese language (nor, as far as I know, do they derive from Hlai). Indeed the origin of most of these forms which are unique to Tsat is uncertain and there are no clear instances among them of unique retentions, within the range of exemplified Chamic languages, of forms from PMP which have been replaced elsewhere within Chamic by borrowed or innovated forms. One item from Western Cham (the word for 'spider') apparently derives from or is influenced by the form in Khmer (Robert Headley, personal communication).

The number of items that have been retained from PMP in the lists in the various Chamic languages is given in the table below. Cases where a PMP descendant and a form of other origin coexist have been marked as though they were pluses. Cases where the same form is used in more than one gloss (for instance where the same morpheme is used in both the singular and plural pronouns, or cases where 'short' and 'small' are expressed by the same root) are only counted once, however, which explains why some languages with full lists show totals under 200. This is also the practice where the form in question in a particular language is a compound of two elements, both of which are already separately logged in the table. The figures are as follows:

---

[20] The proportion of cognates on these lists which can be found between several of these pairs of languages (which are of course instances of pair-referenced lexicostatistics!) are several percentage points higher than those cited in the Tables 2a and 2b. But since different lists have been used in the present study from the ones used to calculate the percentages in Tables 2a and 2b (which themselves are based on the results gleaned from slightly different lists), no direct comparison of these sets of percentages is possible. In those cases where one language has two equivalents for the same gloss, one cognate with another form and the other not so, the cognate form is the one taken notice of in my calculations.

PMP/Standard Malay: 112/200;
PMP/Acehnese: 114/199 (but this total possibly minus a couple of as yet undetected Malay loans);
PMP/PRC 107/198,
PMP/Western Cham 105/198,
PMP/Haroi: 105/180,
PMP/Chru: 105/198,
PMP/Jarai: 100/198,
PMP/Rade: 98/198,
PMP/Northern Roglai. 98/196,
PMP/Tsat: 101/198.

In all these cases those forms which have been retained from PMP account for over 50% of the forms on each Chamic Blust list.

If we want to track uniquely shared lexical innovations within subgroups of Chamic as a means of ascertaining the scope of any subgroups (say Highland versus Coastal Chamic) we need to assess which forms are common to all the Chamic languages, or which have been replaced by loanwords in one or more cases. We then need to identify and discount these loanwords, and we also need to establish and set aside any bodies of what we may call *uniques*. The number of 'uniques' found in the Blust lists for other Chamic languages is much smaller. By 'unique' I mean what is sometimes (albeit pedantically inaccurately) referred to as a *hapax legomenon*, namely a phonological form which is only found in a single language and which is assumed to be an innovation within that language. For instance the verb 'jump', for which no etymology is known, is a unique within English.

Rade, which stands out from other Chamic languages in a number of linguistic respects, including the phonology of its presyllables, has only five uniques in the list (plus maybe another one), including a form for 'eye' which refers to the yolk of an egg in other Chamic languages (Rade has lost PMP **mata* in the sense of 'eye'), and the number of uniques in the other languages is even smaller. There is only one unique form given for Blust list glosses in the (admittedly incomplete) data for Haroi, for example, there are only two uniques each for the same bodies of data for Phan Rang Cham and Jarai, and there are none for the Blust list items for Chru or for Western Cham. There also do not appear to be any forms on this list which are lexical innovations (rather than borrowings) that are exclusively found in Northern Roglai and Tsat (which has 21 unique forms of its own), though there is an abundance of inherited Proto-Chamic forms which are common to these languages.

Given the understandably large role which Thurgood (1999), a volume with an admittedly comparative approach to Chamic, has played in the assembling of these data, it is probable that, if we had fully-recorded lists for all the above languages, that there would not be an appreciably greater number of shared cognates than we already find, and that consequently the overall percentage of cognate forms between any pairing of two Chamic languages would be lowered accordingly, even though the cognates which have so far been recognised between various Chamic languages would remain.

This leaves open the question of how (if at all) one should interpret the silence of our information on certain languages (especially Haroi and Chru) in regard to the potential existence there of words which are found in many or most other Chamic languages. Since the Haroi and Chru equivalents for many glosses have not been made available in Thurgood's comparative Chamic lists (which is the source for most of my Haroi and Chru

forms), are we to assume that Haroi and/or Chru express each of these ideas by using words which are not found anywhere else in Chamic? Are the words which these languages do use to express these concepts recent borrowings from Mon-Khmer languages such as Vietnamese or Bahnar, or are they sometimes forms whose origins are as yet unknown and which may have originated within the languages themselves? Or is it simply the case that Haroi and Chru cognate forms of well-known Common Chamic words have not been recorded in the materials available to us? It is impossible for us to say, given the information currently available to us. We can only work with what we have and we cannot employ *argumenta a silentio* to help us out.

The lexical forms in Acehnese which are not shared with other Chamic languages mostly fall into two groups. There are those which are similar to forms in Malay and which look as though they may have been borrowed from Malay, and there are those whose origins are uncertain, though some of them may derive from Aslian languages (however, no convincing etymologies for these latter have been found yet). Among the 200 forms on this list only the Acehnese form for 'to swim' preserves a PMP form, in this case a reflex of PMP **naŋuy*, a form which has been by chance replaced (albeit by different words) both in Malay and in other Chamic languages, and which in Acehnese shows the distinctive Chamic change of *n-* to *l-*: Acehnese *languy* 'to swim'. The replacement word for 'swim' in the other Chamic languages is pan-Chamic, and is most probably borrowed from a Mon-Khmer language. The Malay form for 'swim' is pan-Malayic in distribution but I do not know its origin. One further form, a retention from PMP, is shared between Tsat and Acehnese but not with the other Chamic languages.

A rough and ready indication of the relative degree of linguistic diversity in Chamic can be provided by simply counting up the number of different forms used in the aggregation of Chamic languages in expressing the 200 glosses on the Blust list and then expressing it as a ratio. Acehnese apart, eight lists have been used, those for the Phan Rang and Western varieties of Cham, for Jarai, Rade, Northern Roglai, Tsat, Haroi and Chru. Although the Chamic-language material available to me has serious lexical gaps for Haroi (and there are more gaps here for Haroi than there are for the comparable Malayic lists), I have found that the number of forms used for expressing the 200 concepts on the sum total of the Chamic lists, apart from Acehnese, is 300, that is, a ratio of 1.5 different forms per gloss across a sample of eight languages. (A ratio of 1.00 would indicate to us that all the languages were identical isolects with nothing to distinguish one from another; a ratio of 8.00 would highlight to us that all indications suggested that the eight languages were completely unrelated to one another.) I have full lists for Phan Rang Cham, Western Cham, Jarai, Rade, Chru, Tsat and Northern Roglai; the list for Haroi has 17 omissions.

Now these 300 forms cover 1568 filled slots. The number of slots is arrived at as follows: Ideally I would have 200 forms from the 8 sampled non-Acehnese Chamic languages, making 1600 slots on the grid for these languages. But I have 17 gaps on my table for glosses for which I lack a form in one language. Additionally there are gaps in the columns for most Chamic languages for the 3pl pronoun form, since it is identical to the 3sg form in nearly all Chamic languages, and the same is true of most 2pl pronouns, while some languages also use the same form for 'short' and 'small'. These slots could potentially be filled by 1568 different items, if it were the case that the languages in question bore no lexical resemblances between each other whatsoever. But in fact only 300 separate items are used (excluding a handful of cases in which one language uses two different unique forms to express the meaning of a particular gloss – only one unique is

counted for such slots in each case). This makes this a ratio of 5.26667 slots per individual glossed item (for whatever this ratio may actually be worth; please note that this figure is the **reciprocal** of the figure for the average number of cognate sets per word across the eight languages surveyed).

If one adds into this total the forms on the list which are only found in Acehnese among the Chamic languages (whether or not they are also retained from PMP or are also found in Malay, or whether they are innovations within Acehnese), the total of different forms rises to 361 and the ratio of unrelated forms per gloss therefore rises to 1.85 forms per gloss across a sample of nine languages, exhibiting a total of 1756 slots (for we have a full list for Acehnese), making this a ratio of 4.8642659 slots (out of nine slots available for each gloss) which are occupied on average by each individual glossed item.

This overall very high degree of congruity and commonality of the basic lexicon in Chamic is, we should point out, in marked contrast to the very wide degree of phonological variation (if one views the matter diachronically) which is found across these languages and which is even amply exemplified in the various phonological shapes of those forms which have been inherited from PMP, but which is especially vivid in fully-tonal Tsat.

By comparison, the number of different forms which are used for the equivalents on the eight wordlists which were given for various Malay isolects in Blust (1988)[21], a dataset which has fewer overall gaps than the Haroi list has, and one which represents a group of isolects whose genetic unity has never been in doubt, is 490, or 2.45 forms per individual glossed item across a sample of eight isolects. Were Blust's Salako (or Selako) Dayak Malayic list fuller for our purposes (but unfortunately it is not, as it contains only 173 of the Blust list forms out of a target of 200 (though Sander Adelaar has provided me with the Salako forms for the missing entries), while one form is missing from his Ambonese Malay list), the number of discrete items in use here (and the proportion of items to each gloss) would certainly be higher and it might push the average figure for the number of cognate sets per gloss above 2.50. This is because the material which Blust presents, though incomplete, nonetheless shows that Salako is lexically innovative when contrasted with other Malayic isolects.

So what do we find when we look for shared innovations in an attempt to subgroup the Chamic languages? Not a lot, really. We can make a solid start at answering this question, since we know which forms are inherited from PMP or Proto-Malayo-Chamic and which other widespread forms in Chamic are actually innovations within Chamic (including or excluding Acehnese). We also know which forms on the lists are 'uniques' and which forms are loanwords from various sources. There are also a few cases in which one or more language has two forms to express one gloss, and one or both of these forms are uniques, a fact which also inflates the figures slightly. If we subtract these strata, then

---

[21] The Malay lects which are surveyed in that article are Bahasa Indonesia, Banjarese, Medan Malay, Salako, Iban, Minangkabau, Jakartanese (Betawi) and Ambonese Malay (Bahasa Ambon). Adelaar (1991) uses five of these lists and also provides a directly comparable wordlist for the Middle Malay language Seraway, providing equivalents for 188 out of the 200 items on the Blust list. He additionally reconstructs Proto-Malayic forms from this evidence wherever possible. Neither author provides a list for Kerinci, which is usually classified as a phonologically divergent dialect of Minangkabau, although we do know that it retains 100 out of the 200 PMP forms that are used on the Blust list (Blust 2000b: 329). I have only recently had access to Blust's list for Kerinci (Blust et al. 2005) and have therefore not used it in the above work.

what remain should be the clusters of exclusively shared lexical innovations. The problem is that so little material is left to us after these subtractions. Even some assured historical linkages, such as that of Northern Roglai and Tsat, are not manifested by large bundles of shared lexical innovations in our data; Thurgood (1999) shows us that the strong evidence for this history of shared development is actually phonological. In order for us to have strong evidence, from the basic lexicon, of substantive subgroups within Chamic (apart from Acehnese, which stands somewhat on its own, by virtue of having retentions, innovations and numerous loans from Malay) we would need to find clusters of lexical forms which have developed independently among two or more Chamic languages at a period after at least the beginning of dialectal differentiation within Chamic, and this we do not find to any striking extent.

In regard to a possible Highland versus Lowland Chamic division, Jarai and Rade seem to share a couple of forms on their translations of the Blust list which are not also found in PMP, Malay, Acehnese, or Western and Eastern Cham (forms standing for 'dust' and 'to spit', for instance Rade *ɓruih* and *bah* respectively; I cite these forms from Egerod 1978). But this 'Highland Chamic' group is weakly supported, and there is no innovatory lexical evidence in the basic vocabulary for a similar coastal group comprising (say) Chru, Roglai and Haroi.

On a final note, it should be mentioned that the origins of the various forms are not easily stratifiable by form class. Mon-Khmer borrowings into Chamic languages include verbal and several pronominal forms in addition to nouns, while the stratum of forms of uncertain origin also includes some pronouns. The form class which is most homogeneous in terms of its origin is that of the numerals, which are robustly Austronesian or at least Malayo-Polynesian in origin.

6. Summary of findings

The distinction between horizontal and vertical lexicostatistics has been discussed above. In this study both techniques are used, firstly the horizontal and then the vertical, together with norm-referenced lexicostatistics in which the norm used provides the vertical element in the study, and the various techniques tell us different kinds of things. (It is therefore essential to employ the several methods in the correct sequence, otherwise the end result is pseudo-statistical nonsense.) Comparison of lists for various modern Chamic languages is an example of horizontal lexicostatistics, whereas the use (as the cross-referencing 'norm') of Proto-Malayo-Polynesian reconstructions against which to compare the occurrence or non-occurrence of such forms in modern Chamic languages is an instance of vertcial lexicostatistics. The inclusion in the study of tabulated findings from lists from modern Standard Malay and from modern Acehnese allow us to examine diachronic issues which have to do with Chamic languages, and they allow us to appreciate that Malay is the most closely related language grouping to Chamic and that Acehnese is equidistantly similar to all the (more conservative) Chamic languages. This is just what one might expect from a language which derives from a Chamic variety whose speakers left Indochina in a period before the Chamic lects had had opportunity to separate into different languages.

The status of Acehnese as a historical witness is reinforced by the fact that it retains morphological features, for instance the use of productive infixation, which were common to Proto-Chamic and Proto-Malayic and further back in time, to Proto-Malayo-Chamic, but which were subsequently lost (or reduced to lexicalised vestiges) in all the other Chamic languages and in Malayic ones too. Its status as a lexical witness is somewhat diluted by its

wholesale absorption of words from Malay. Malay, Acehnese and the other Chamic languages have all lost many features which their common ancestor had retained and which it had often retained from PMP or even earlier, but they have not always lost the same things.

The use of the informative but still undervalued technique of norm-referenced lexicostatistics makes the degree of similarity between pairs of languages much clearer than the normally-used technique of pair-referenced lexicostatistics does. It also enables us to see what kinds of forms are shared between languages, which other forms differ in any or all languages, and we can also see whether there are any forms which seem to buck otherwise prevalent linguistic trends – the presence of stray retentions from the common proto-language in one language when all other languages in the sample have shared an innovation, for instance. If the norm which is used as the point of comparison with the other languages is an earlier stage of an attested language, or a reconstructed proto-language, **but only if it is a proto-language which has been reconstructed without reference to the particular languages which are under discussion in the sample being examined**, the findings can be far more informative and they may give a much clearer historical picture. Such information, often regarded as too cumbersome to present in part (as I do here with the grid) or in whole (as would be done by reproducing the exact forms) can shed light on what lies behind the bleak tables of unannotated percentages which Dyen and his followers have offered.

This 'criterion of primordial objectivity' is clearly met here, because Proto-Chamic and its descendants have played little or no part in the reconstruction of Proto-Austronesian (PAn) or of its daughter proto-languages, so that the process does not involve an excess of application of circular reasoning. PMP or PAn reconstructions can this be used as much more objective yardsticks to cast certain sorts of light on Chamic linguistic history. The advantage of using forms from an actually-attested, or at least well-reconstructed, earlier stage (a 'parent language') of the languages under examination in such a sampling is that one can mark up the rows on the grid to show which forms are maintained from the parent language and which are replaced by innovations (or borrowings) in each of the 'daughter' languages under observation. Having done that, it is then possible for us to plot patterns of occurrences of these innovations, and to see to what extent these correlate across and within the daughter languages.

Lexical material has been privileged in this study because of the paucity of elements of bound morphology in the Chamic languages, and the list which I have used is one which is supposed to be especially suitable for the exploration of the histories of Austronesian languages. Other language families would require the use of other lists, but there is no reason why norm-referenced lexicostatistics should not be used as part of the battery of tests used to determine the genetic affiliation of 'troublesome' languages, and in the case of the Chamic languages, where the usually diagnostic bound morphology is so sparse (and is sometimes clearly borrowed), it happens to be especially useful.

Bound morphology is the kind of material which would normally be looked upon as providing firmer evidence for the Austronesian affinities of Chamic and for placing Chamic in the right niche within the Austronesian family tree than would normally be provided by lexical evidence. The fact that most of the rather few productive bound morphs in Chamic languages are typologically, semantically, formally and functionally very similar to those found in Mon-Khmer languages, not least the Central Bahnaric ones,

is a factor which has probably supported and assisted their continued use in the structures of Chamic languages.

It is also the case that at least the Indochinese Chamic languages have borrowed (or, more precisely, it is true that the descendants of women who shifted from Mon-Khmer languages to Chamic ones have perpetuated) a very large proportion of their free grammatical morphs from Mon-Khmer languages, a proportion which, viewed crosslinguistically, is unusually high and which includes personal pronouns, semantically-blank adpositions such as *baʔ* 'at', discourse particles, and possibly some negators such as *bɛʔ* 'irrealis negator' (to say nothing of the presence of some very common Chamic verbs which are of Mon-Khmer origin). Some of these forms are included on the Blust list, which in any case was not drawn up primarily with Chamic languages in mind. The same remarks also apply, though to a smaller degree, to the presence in these form-classes of a number of elements of (at present) uncertain origin which are pan-Chamic in distribution and which therefore reconstruct back to Proto-Chamic. In fact, the only form-class in Chamic languages whose contents are purely Austronesian in origin is the system of numerals.

The Proto-Chamic (or at least pan-Chamic) material which the Blust list provides and draws upon can be shown to contain many elements inherited from PMP, a band of elements modified from their PMP prototypes, further elements which are shared with Malayic lects, and yet other elements which are derived from Mon-Khmer languages and some which are pan-Chamic in distribution but of unidentified origin. The contents of these bands rarely overlap, there are rather few cases in which some languages have adopted words for a particular items from one source while other Chamic languages have retained words for the same concepts which were to be found in an earlier historical stage of the language. It can only rarely be shown (for instance in the case of the additional Bahnaric forms which come from Hrê and which occur in Haroi but nowhere else in Chamic) that one Chamic language has taken a greater share or a bigger number of forms from a particular group of Mon-Khmer languages, and this donor group being a group which has provided forms that are reflected throughout the Chamic languages, than any other Chamic language has.

This relative discreteness of the various bands suggests that the contents of each new band of elements had largely consolidated in the period before the next wave of elements entered Proto-Chamic, and this implies that productive contact with each set of donor languages had largely ceased, before the next wave of loans or innovations came in from a different direction. Periods of borrowing, from whatever source, appear to have been succeeded by periods of consolidation of these borrowings (and of other external influences). Proto-Chamic included elements from both North and Central Bahnaric languages, and these spread into the modern languages from Proto-Chamic rather than from fortuitously coincidental borrowing of such forms from adjoining languages.

The main features of the picture are clear enough, and they fully support Thurgood's historical hypotheses in Thurgood (1999). The Chamic languages derive from a Western Malayo-Polynesian language which in origin is very similar to modern Malay, with which it shares a number of phonological and lexical innovations which set them apart from other Malayo-Polynesian languages. (Nonetheless the Chamic and Malayic branches have both subsequently gone their separate ways, and it is evident that they had already done so even at the periods of first attestation of Old Cham and Old Malay.)

Acehnese is clearly a part of the Chamic subgroup, rather than being coordinate with Chamic and Malay.[21] It is coordinate with all the other Chamic languages (which had not differentiated much before the departure of the Acehnese to Kelantan and latterly to Sumatra, and which may in fact have diffused several more innovations and loans among its dialects after Acehnese's departure). Acehnese shares with the other Chamic languages a number of basic loans from Mon-Khmer languages and additionally a number of pan-Chamic words of unknown etymology, although an examination of the contents of its basic Blust list vocabulary indicates that later Acehnese contact with Malay and (probably and to a lesser extent) with other, as yet unidentified, languages has also taken place. The striking parallels (which are caused mostly by shared retentions of Proto-Chamic forms) which are to be found within the basic lexicon of Chamic languages belie the first impressions of immense diversity among them. These first superficial differences have resulted from several different series of phonological changes which have affected particular Chamic languages (or which have sometimes affected groups of them together, we note for instance the development of word-final preploded nasals which is shared by Tsat and Northern Roglai, and which caused Graham Thurgood to link them together historically). Probably none of the Chamic languages discussed here are nowadays interintelligible, but a millennium ago this mutual unintelligibility may not have been the case.

This study also shows that the specifically Chamic affinities of Tsat are similarly historically secure, as Tsat contains elements from Mon-Khmer languages and a portion of the aforementioned lexical 'unknowns', in addition to containing a few later loans from Hlai and an especially large number of loans from Chinese languages, which are not found in other Chamic languages and which mostly do not figure in the materials on the Blust list.[22] The retention in the Tsat lexicon of a number of common Chamic forms, which are

---

[22] Not everyone agrees with this view, and Sidwell (2004) discusses some objections to it. In his view, which draws upon some descriptive and historical work on Moken-Moklen by Michael Larish (Larish 1999), Malayic, Moken-Moklen of the Mergui Archipelago in Burma and coastal Thailand, plus the language of the Orang Laut of eastern coastal Sumatra, plus Acehnese and Chamic, are all members of a genetic subgroup of Malayo-Polynesian that was centred on mainland Southeast Asia, and whose members were strongly influenced (at least at a lexical level) by Mon-Khmer languages and also by other substrate languages which have no known cognates and which have left no other trace, and which I call the 'submerged substrate' language(s). (Part of the evidence for this is that exceedingly few of the alleged Bahnaric loans into Chamic languages are found in West Bahnaric languages; they are much more common in Central and South Bahnaric languages. Consequently, Sidwell suggests that both South and Central Bahnaric and Chamic languages have borrowed these forms from the submerged substrate.) According to this hypothesis, Acehnese, which shares only a small proportion of the Mon-Khmer and 'submerged substrate' elements which are found in all (other) Chamic languages (including the 'submerged substrate' loans which Thurgood (1998) and others took to be loans into Chamic from Bahnaric), has subsequently acquired many features which make it appear Chamic as a result of the migration of many Chams to northern Sumatra in the Middle Ages. Additionally, Acehnese has borrowed massively and often at a very basic level from Malay in the last few centuries, a practice which would have diluted the number of Mon-Khmer loans in the language in any case. The full implications of these controversial claims have yet to be worked out.

[23] It is also feasible that the Utsat, being Muslims, have also borrowed philosophical vocabulary and other Islamically-focussed lexicon from Malay, but I do not know of any such examples in the available Tsat material.

found in Indochinese Chamic languages (and which are often though not always also evident in Acehnese[23]) and which are of Mon-Khmer origin, easily gives the lie to any vain idea that Tsat reflects the evidence of an early separate migration to Hainan which is coeval with the date of dispersal of Acehnese and the other Chamic languages, and which would suggest that Tsat is a primary and primordial subdivision of Chamic, with Acehnese on the one hand and the remaining Chamic languages on the other, comprising the two other branches. But Tsat and Acehnese do not appear to share any special lexical innovations (nor do they preserve many sole lexical retentions, for that matter) against the other Chamic languages which would permit us to unite them in a special subgroup. If they had shared some lexical retentions, then it would most probably be the case that the forms for the relevant glosses in other Chamic languages would be lexical innovations which had passed through Indochinese Chamic languages after the departure of what were to become the Acehnese and Tsat speech communities.

There is little in the way of strong or plentiful innovatory lexical evidence for any particular subgroupings within Indochinese Chamic, although Jarai and Rade seem lexically to be slightly more similar to one another than they are to the other Chamic languages, and they seem to share a few (but only a few) lexical innovations which are unknown elsewhere. This admittedly small degree of shared lexical innovation occurs despite their very different phonological histories, with the relative phonological conservatism of Jarai pitted against the extreme degree of Rade phonological innovation, something which is especially marked in Rade presyllables. But we should always remember that the Jarai and Rade speech communities neighbour one another in the southern Vietnamese highlands and the neighbouring parts of Cambodia. It is also true that they are more similar to one another in most respects of typology and in fabric (that is, in the morphemes which they possess) than they are to any other neighbouring language, so that some words may have diffused from one of the languages to the other one. For the rest, the basic vocabularies and Blust list responses of the two modern Cham (rather than Chamic) languages, and those of Chru, Haroi and Northern Roglai, especially the first two mentioned there, are very similar to one another (at least as far as what we can adduce from what we have available), so that we have to look elsewhere other than the basic lexicon in order to find differences between the languages.

The evidence of Blust-style lexicostatistics supports the picture which Thurgood gives in his book, that of what was originally the northernmost Chamic language (or the northernmost link in the Chamic dialect chain) peeling off, migrating to the south and forming the basis for modern Acehnese. Meanwhile the next most northerly language also moves out of what is now northern Vietnam and its speakers cross to Hainan and form the nucleus of the modern Tsat speech community (a community which is to be expanded with the later arrival of speakers of a more southerly Cham dialect).

We may date the split of Acehnese to the late tenth century AD, with the downfall of the northern Cham empire, and that of Tsat maybe a little later. Jarai and Rade are the next to split away, but if they share a period of unity it is a brief one, with few shared lexical innovations and not many shared phonological ones either. The remaining Cham lects then diffuse along the coast and their speakers go somewhat further into the hinterland, presumably in the course of the first half of the second millennium, while

[24] Some of these forms may once have existed in Acehnese but they may have been replaced (possibly by loans from the more prestigious Malay). Absence of evidence is not evidence of absence.

speakers of Western Cham are parted from their stay-at-home Eastern Cham fellows after 1471 and retreat to Khmer-speaking territories. Acehnese went with its speakers from Indochina to Sumatra as a Western Malayo-Polynesian language which had come under strong influence from Mon-Khmer languages, an influence which were already beginning to reshape its phonology and which had already done this to its lexicon, although its morphological system remained typically and conservatively Austronesian.

The Chamic languages which remained in Indochina gradually absorbed more and more of the areal features of general (and later on, of individual) Mon-Khmer languages, and they absorbed more and more vocabulary from this source too. Much later – mostly within the last century – all of the Indochinese Chamic languages except Western Cham, which has been as strongly influenced by Khmer as the others have been by Vietnamese, have absorbed huge amounts of vocabulary from Vietnamese (and there are even a few of these Vietnamese loans present in Mekong Delta Western Cham too, since the official language there, though not the regional majority language, is Vietnamese; Headley 1991 provides a couple of examples of these loans).

The Chamic languages are unusual among Austronesian languages inasmuch as a high proportion of the elements in the extensive non-Austronesian parts of the basic vocabularies can be etymologised. Furthermore a very high proportion of the inherited elements in the Chamic language lexica that derive from PMP can be found in the most basic strata of the vocabulary (at a rough guess, maybe almost 40% of the inherited morphs in Chamic languages which serve to make these languages lexically Austronesian can be found among the Blust list responses).

Tsat's process of change, which had progressed further from the inherited Western Malayo-Polynesian norm than Acehnese had, was interrupted at a tome when it had already shed such features as infixation (involving a feature and elements which Acehnese never lost) and had absorbed plentiful amounts of Mon-Khmer lexicon. Firstly weak influence from Hlai, then much stronger influence from the Hainanese form of Southern Min (Minnan) and finally two waves of influence from Mandarin, the second much stronger than the first, shaped and shape Tsat. With the very partial exception of Acehnese, all Chamic languages show in every way the marks of profound influences from non-Austronesian languages, but none show these more so than Tsat, where the influence, which comes especially from Chinese but which also includes earlier influence from Mon-Khmer languages, is massively clear at all levels if one knows what to look for.

And the effects of these languages on Tsat's basic vocabulary are no exception to this generalisation. At first glance Tsat appears to be anomalous among the Chamic languages because it contains a higher proportion of forms on the Blust lexicostatistical list which are unique and which are not found elsewhere in Chamic. But some of these un-Chamic forms are simply recent loans that have been acquired from Chinese languages, and the origins of others may yield themselves up to us after further investigations are carried out among the languages spoken in southern China and especially on Hainan Island.[24] Tsat's genetic relationships with other languages can still best be seen by the

[25] The impact of Tsat on the Hlai lexicon is probably not to be underestimated either, although this topic requires further investigation – and yet there is not as much information available on Hlai as one might wish for. It is almost certain that Hlai was the first tonal language with which Tsat-speakers were in contact, and that it was spoken by people who were living around and maybe among Tsat-speakers, and it is further probable that this Tsat-Hlai contact began long before any Chinese language came to be used in that part of Hainan. However, so far only a handful of

etymological examination of its basic vocabulary, since typologically it doers not look Chamic at all, having lost all its distinctive bound inflectional morphology, while in its possession of tones and other features it rather resembles the typology of other languages of Hainan (such as the Tai-Kadai languages Hlai and Ong-Be, and of course the non-Tai Minnan Chinese), whatever the genetic origins of these languages are.

We may note that Thurgood (1999) points throughout his book that there are frequent problems with demonstrating that the various reflexes in the daughter languages of Proto-Chamic forms are perfectly lautgesetzlich. Quite often there are phonological irregularities of realisation simultaneously in the initial, vowel and final phone in some Chamic language's reflex of a particular Proto-Chamic form, even though we can be all but certain (or we put faith in the hope) that the form derives from (or reconstructs back to) Proto-Chamic. And many of the problems which these almost certainly cognate but phonologically aberrant forms present have yet to be solved, just as is the case with many other issues in the internal and external history of the Chamic languages.

References

Adelaar, Karl Alexander. 1991. 'Some notes on Sri Lanka Malay.' *Papers in Western Austronesian languages,* edited by Hein Steinhauer, 23-37. Pacific Linguistics: A-81. Canberra: Research School of Pacific Studies, Australian National University

-----1992. *Proto-Malayic: the reconstruction of its phonology and parts of its lexicon and morphology.* Pacific Linguistics C-119. Canberra: Research School of Pacific Studies, Australian National University.

-----1995. 'Borneo as a crossroads for comparative Austronesian linguistics.' *The Austronesians: historical and comparative perspectives*, edited by Peter Bellwoof, James J. Fox and Darrell T. Tryon, 75-97. Canberra: Australian National University.

-----1996. 'Malay in the Cocos (Keeling) Islands.' *Reconstruction, classification, description: Festschrift in honor of Isidore Dyen,* edited by Bernd Nothofer, 167-198. Hamburg: Abera.

-----2001. 'Malayic, Chamic and Bali-Sasak-Sumbawa: the demise of Malayo-Javanic.' Paper presented to the Fifth International Symposium on Malay/Indonesian Linguistics.

Alieva, Natalia F. 1984. 'A language union in Indo-China.' *Asian and African Studies* [Bratislava] XX: 11-21.

Bakker, Peter, and Maarten Mous (eds.). 1994. *Mixed languages: 15 case studies in language intertwining.* Amsterdam: IFOTT.

Benedict, Paul K. 1941. 'A Cham colony on the island of Hainan.' *Harvard Journal of Asiatic Studies* 4: 129-134.

Bennett, Patrick R. 1998. *Comparative Semitic linguistics: a manual.* Winona Lake, Illinois: Eisenbrauns

Blust, Robert, Russell D. Gray and Simon Greenhill. 2005. *Austronesian Basic Vocabulary Database.* Department of Psychology, University of Auckland. Accessible at: http://language.psy.auckland.ac.nz/index.php

---

loans have been identified as coming from Hlai into Tsat, and a few words have gone the other way, notably a form for 'six' which derives from Chamic **nam*, which presumably replaced an earlier Hlai numeral (Graham Thurgood, personal communication, April 2003).

Blust, Robert A. 1981a. 'Variation in the retention rate in Austronesian languages.' Paper presented at the Fifth International Conference on Austronesian Languages, Denpasar, Bali.

-----1981b. ' The reconstruction of Malayo-Javanic: an appreciation.' *Bijdragen tot de Taal-, Land- en Volkenkunde* 137: 456-469.

-----1990a. 'Malay historical linguistics: a progress report.' *Rekonstruksi dan cabang cabang bahasa Melayu induk*, edited by Moh[amme]d Thain Ahmad and Zaid Mohamed Zaidi, 1-33. Kuala Lumpur: Dewan Bahasa dan Pustaka.

-----1992. 'The Austronesian settlement of mainland Southeast Asia.' Karen L. Adams and Thomas John Hudak (eds.), *Proceedings of the Second Annual Meeting of the South East Asian Linguistics Society*, 25-83. Tempe: Arizona State University Press.

-----1993a. 'Central and Central-Eastern Malayo-Polynesian.' *Oceanic Linguistics* 32: 243-292.

-----2000a. Review of Thurgood (1999). *Oceanic Linguistics* 39: 435-445.

-----2000b. 'Why lexicostatistics doesn't work: the 'universal constant' hypothesis and the Austronesian languages.' *Time depth in historical linguistics*, edited by Colin Renfrew, April McMahon and Larry Trask, 311-331. Cambridge: McDonald Institute for Archaeological Research.

Bochet, Gilbert, and Jacques Dournes. 1953. *Lexique polyglotte: Vietnamien, koho, roglai, français*. Saigon: Editions France-Asie.

Bui Khanh The. 1995. *Tu-Dien Cham-Viet*. Ho Chi Minh City: Nha Xuat Ban Khoa Hoc Xa Hoi.

Collins, [Ira] Vaughn. 1969. 'The position of Atjehnese among Southeast Asian languages.' *Mon Khmer Studies* 3: 48-60.

Daud, Bukhari, and Mark Durie. 1999. *Kamus Basa Acèh - Kamus Bahasa Aceh – Acehnese - Indonesian - English Thesaurus*. Pacific Linguistics: C-151. Canberra: Australian National University.

Dempwolff, Otto. 1934-1938. *Vergleichende Lautlehre des austronesischen Wortschatzes*. Hamburg: Beiheft zur Zeitschrift für Eingeborenensprachen.

Durie, Mark. 1990. 'Proto-Chamic and Acehnese mid-vowels: Towards Proto-Aceh-Chamic.' *Bulletin of the School of Oriental and African Studies* 53: 100-114.

Dyen, Isidore. 1962. 'The lexicostatistical classification of the Malayopolynesian languages.' *Language* 38: 38-46.

-----1965. *A lexicostatistical classification of the Austronesian languages*. International Journal of American Linguistics Memoir 19. Baltimore.

-----1971. 'The Chamic languages.' *Current Trends in Linguistics, volume 8: Oceania*, edited by Thomas A. Sebeok, 110-120.

Egerod, Søren. 1978. 'An English-Rade Vocabulary.' *Bulletin of the Museum of Far Eastern Antiquities* [Stockholm] 50: 47-108.

Fuller, Eugene. 1977. 'Chru phonemics'. Thomas et al (ed., 1977): 77-86.

Grace, George W. 1966. 'Austronesian lexicostatistical classification: a review essay.' *Oceanic Linguistics* 5: 13-31.

Greenberg, Joseph H. 1987. *Language in the Americas*. Stanford: Stanford University Press.

Headley, Robert A. 1976. 'Some sources of Chamic vocabulary.' *Austroasiatic studies*, edited by Philip N. Jenner, Laurence C. Thompson, and Stanley Starosta, 453-476. Honolulu: University Press of Hawai'i.

-----1991. 'The phonology of Kompong Thom Cham.' *Austroasiatic languages: essays in honour of H. L. Shorto*, edited by Jeremy H. C. S. Davidson, 105-121. London: School of Oriental and African Studies.

Hooley, Bruce A. 1971. 'The Austronesian languages of Morobe District, Papua New Guinea.' *Oceanic Linguistics* 10: 79-151

Hudson, Alfred B. 1967. *The Barito dialects of Borneo, a classification based on comparative reconstruction and lexicostatistics*. Ithaca, New York: Cornell University, Department of Asian Studies.

Jacob, Judith M. 1963. 'Prefixation and infixation in Old Mon, Old Khmer and Modern Khmer.' *Linguistic comparison in South East Asia and the Pacific*, edited by H. L. Shorto, 62-70. London: Luzac.

Lafont, Pierre-Bernard. 1968. *Lexique jarai parler de la province de plei ku.* Paris: Publications de l'Ecole Française de l'Extrême-Orient LXIII.

Larish, Michael David. 1999. *The position of Moken-Moklen within the Austronesian language family*. Unpublished doctoral dissertation, University of Hawai'i at Manoa.

Lee, Ernest Wilson. 1998. 'The contribution of Cat Gia Roglai to Chamic.' Thomas (ed., 1998), 31-54.

Matisoff, James A. 2001 'Genetic versus Contact Relationship: Prosodic Diffusability in Southeast Asian Languages', in Alexandra. Y. Aikhenvald and R.M.W. Dixon (eds.), *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*, Oxford University Press, Oxford, 291-327.

Miller, Wick R. 1967. Uto-Aztecan cognate sets. *University of California Publications in Linguistics 46.* Berkeley and Los Angeles: University of California Press.

-----1984. 'The classification of the Uto-Aztecan languages based on lexical evidence.' *International Journal of American Linguistics* 50: 1-24.

-----, James L. Tanner and Lawrence P. Foley. 1971. 'A lexicostatistical study of Shoshoni dialects.' *Anthropological Linguistics* 13: 142-164.

Pang, Kang-Feng. 1998. 'On the Ethnonym 'Utsat.' Thomas (ed.), 55-60.

Ringe, Don, Tandy Warnow and Ann Taylor. 2002. 'Indo-European and computational cladistics.' *Transactions of the Philological Society* 100: 59-127.

Shintani, Tadahiko L. A. 1981. *Boh blu Êdê-Yuan-Za pô nê = Tù v.ung Êdê-Vi.êt-Nhât.* Tokyo: Viên Nghiên Cúu Ngôn Ngự Và Van Hóa á Phi. [Japanese, Rade and Vietnamese dictionary.]

Sidwell, Paul. 2004. 'On the origins of the Proto-Chamic lexicon and significance of Acehnese.' To appear in *Mon-Khmer Studies.*

Swadesh, Morris. 1950. 'Salish Internal Relationships.' *International Journal of American Linguistics* 17: 257-267.

-----1955. 'Toward Greater Accuracy in Lexico-Statistic Dating.' *International Journal of American Linguistics* 21: 121-137.

Tegenhardt-Mundhenk, Alice, and Hella Goschnick 1977. 'Haroi phonemes.' Thomas et al. (ed., 1977): 1-15.

Tharp, James and Y-Bhăm Buôn Yă. 1980. *A Rhade-English dictionary with English-Rhade finderlist.* Pacific Linguistics C-58. Canberra: Australian National University.

Thomas, David D. (ed.) 1998. *Studies in Southeast Asian languages no. 15: Further Chamic studies*. Pacific Linguistics A-89. Canberra: Research School of Pacific Studies, Australian National University.

-----, Ernest W. Lee and Nguyen Đang Liem (eds.). 1977. *Papers in Southeast Asian languages, no. 4. Chamic studies*. Pacific Linguistics A-48. Canberra: Research School of Pacific Studies, Australian National University.

Thurgood, Graham.1996. 'Language contact and the directionality of internal drift: the development of tones and registers in Chamic.' *Language* 71: 1-31.

-----1999. *From ancient Cham to modern dialects: two thousand years of change*. Oceanic Linguistics Special Publication 28. Honolulu: University Press of Hawai'i.

-----To appear a. 'A preliminary sketch of Phan Rang Cham.' 25 pp., to appear in *The Austronesian Languages of Asia and Madagascar*, edited by Alexander Adelaar and Nikolaus Himmelmann. London: Curzon.

-----To appear b. 'Learnability and direction of convergence in Cham: the effects of longterm contact on linguistic structures.' Manuscript, 21 pp.

-----To appear c. 'Crawfurd's 1822 Malay of Champa'. 12 pp., to appear in a Festschrift for P J Mistry.

-----, and Fengxiang Li. 2003. 'Contact-induced variation and syntactic change in the Tsat of Hainan.' David Bradley, Randy LaPolla, Boyd Michailovsky and Graham Thurgood, eds., *Language variation: papers on variation and change in the Sinosphere and in the Indosphere in honour of James A. Matisoff*, 185-200. Canberra: Pacific Linguistics 555.

Tín, Pham Xuan. 1955. *Đạ ngữ tiêu tự diển: Lexique polyglotte*. Dalat [Vietnam]: Langbian. [Vietnamese, Jarai, Chru and French glossary.]

Zheng Yiqing. 1997. *Huihuihua yanjiu*. Shanghai: Shanghai Yuandong Chuban She.

TABLE 1: *A norm-referenced lexicostatistical grid for PMP, Malay and Chamic languages (including Acehnese), with comments.*

(The languages surveyed are: Proto-Malayo-Polynesian, Standard Malay, Acehnese, Phan Rang Cham, Western Cham of the Mekong Delta in Vietnam, Haroi, Chru, Jarai, Rade, Northern Roglai, and Tsat). Comments are provided for some entries.

No.	Gloss	PMP	Mal	Ach	PRC	WeC	HA	CR	JA	RA	RO	TS	Comments
1	Hand	A	A	A, B	A	A	A	A	A	A	A	A	
2	Left	A	A	B	C	C	C	C	C	C	C	C	C<MK
3	Right	A	A	B	C	C	C	C	C	C	C	D	D<UNK
4	Foot	A	A	A	A	A	A	A	A	A	A	A	
5	To walk	A	B	C	A	A	0	A	A	A	A	A	
6	Road	A	A	A+	A	A	A	A	A	A	A	A	A+<MAL
7	To come	A	B	C	A	A	A	A	A	A	A	A	
8	To turn	A	B	C	C	C	C	C	C	D	C	C	C<UNK
9	To swim	A	B	A	C	C	C	C	C	C	C	C, D	C<MK (often compounded with 122), D<UNK
10	Dirty	A	B	B+	C	C, D	D	C	D	E	C	C	B+<MAL
11	Dust	A	B	C	C	C	0	C	D	C, D	C	D	C<SKT
12	Skin	A	A	A	A	A	A	A	A	A	A	A	
13	Back	A	A	B	C	C	C	C	C	C	C	C	C<MK
14	Belly	A	B	B	A	A	A	A	A	A	A	A	B<MK?= Chamic 'guts'

15	Bone	A	A	A	A	A	A	A	A	A	A	A	
16	Guts	A	B	B	B	B	B	B	B	B	B	B	
17	Liver	A	A	A	A	A	A	A	A	A	A	A	
18	Breast	A	A	A	A	A	A	A	A	A	A	A	
19	Shoulder	A	B	A	A	A	A	A	A	A	A	A	The A form occurs with a different sense in Malay
20	To know	A	A	A	A	A	A	A	A	A	A	A	
21	To think	A	B	B	C	C	C	C	D	C	C, E	C?	B<AR
22	To fear	A	A	A	B	B	B	B	B	B	B	C?	
23	Blood	A	A	A	A	A	A	A	A	A	A	A	
24	Head	A	B	C	C	C	C	C	C	C	C	C	B<SKT, C<MK
25	Neck	A	A	B	B	B	B	B	B	B	B	B	B<MK?
26	Hair	A	B	A	A	A	A	A	A	A	A	A	B= 'root' in PMP
27	Nose	A	A	A	A	A	A	A	A	A	A	A	
28	Breath	A	A	A	A	A	A	A	A	A	A	A	
29	To smell	A	B	B	B	B	B	B	B	B	B	B	B<MK
30	Mouth	A	B	A	A	A	A	A	A	A	A	A	
31	Tooth	A	B	B	B	B	B	B	B	B	B	B	
32	Tongue	A	A~	A	A	A	A	A	A	A	A	A	
33	To laugh	A	A	B	C	C	C	C	C	C	C, D	C	C<UNK
34	To weep	A	A	B	C	C	C	C	C	C	C	C	C<MK?
35	To vomit	A	A	A	A, B	B	A	B	A, B	A, B	B	A	B<MK
36	To spit	A	B	A	C	D	C	C	C, D	C, D	C	C, E	C, D<MK
37	To eat	A	A~	A~	A~	A~	A~	A~	A~	A~	A~	A~	
38	To cook	A	A	A, B	A	A	A	A	A	A	A	A, B?	
39	To chew	A	A	A	A	A	A	A	A	A	A	A	
40	To drink	A	A~	A~	A~	A~	A~	A~	A~	A~	A~	A~	
41	To bite	A	B	C	C	C	C	C	C	C	C	D?	C<MK
42	To suck	A	B	C	D	D	D	D	D	D	D	C	D<MK
43	Ear	A	A	A	A	A	A	A	A	A	A/B	A	
44	To hear	A	A	A	B	B	B	B	B	B	B	B	B<UNK
45	Eye	A	A	A	A	A	A	A	A	B	A	A	B= 'eggyolk' in other Chamic lgs
46	To see	A	B	C	D	D	D	D	D	D	D	D	D<MK
47	To yawn	A	A	B	C	C	B	B	B	B	B	C	B<MK, C<UNK
48	Sleep	A	A	B	B	B	B	B	B	B	B	B	B<MK?
49	To lie down	A	B	B	C, D	C, D	0	D	C	C, D	C	D	
50	To dream	A	A~	A~	A~	A~	A~	A~	A~	A~	A~	A~	
51	To sit	A	B	B	B	B	B	B	B	B	B	B	
52	To stand	A	B	C	C	C	C	C	C	C	C	C	C<MK
53	Person	A	B	B	B	B	B	B	B	B	B	B	
54	Male	A	A	A	A	A	A	A	A	A	A	A	
55	Female	A	B	C	A, D	A, D	A, D	A, D	A, D	A	A, D	A, D	D<UNK
56	Child	A	A	A	A	A	A	A	A	A	A	A	
57	Husband	A	A	B	C	D	C	D	C	C	E	A?	C<MK, E= 'master of house'
58	Wife	A	B	A	C	C	C	C	C	C	C	C	B<SKT
59	Father	A	B	C	B	B	B	B	B	B	B	B	
60	Mother	A	B	A	A	A	A	A	A	A	B	A	
61	House	A	A	A	B	B	B	B	B	B	B	B	B in ACH = tent
62	Thatch	A	A	A	B	B	B	B	B	B	B	B	B<UNK
63	Name	A	B	A	A	A	A	A	A	A	A	A	B<SKT
64	To say	A	B	C	D	D	D	D	D	D	D	D	B<SKT, D<MK
65	Rope	A	A	A	A	A	A	A	A	A	A	A	
66	To tie	A	A	A	A	A	A	A	A	A	A	A	
67	To sew	A	A	B	A	A	A	A	A	A	A, C	A	
68	Needle	A	A	A	A	A	A	A	A	A	A	A	
69	To hunt	A	B	B	C	D	0	D	C	D	E	F	C<UNK
70	To shoot	A	A	A	A	A	A	A	A	A	A	A	
71	To stab	A	B	C	D	D	E	E	F	E	E, G	E	F<BAH
72	To hit	A	B	C	C	C	C	C	C	C	C	C	C<MK
73	To steal	A	B	C	C	C	0	C	C	C	C	C	B<SKT, C <MK

74	To kill	A	A	B	C	C	C	D	E	D	C	D	D= 'CAUS + die', E= 'CAUS + ?'
75	To die	A	A	A	A	A	A	A	A	A	A	A	
76	(To) live	A	A~	A	A	A	A	A	A	A	A	A	
77	To scratch	A	B	C	D, E	E	E	F	D	D, F	F	F	D<MK?, F<MK
78	To cut	A	A	A	A	A	A	A	C	A	A	A, B	B, C <UNK
79	Wood	A	A	A	A	A	A	A	A	A	A	A	
80	To split	A	A	A	A	A	A	A	A	A	A	A	
81	Sharp	A	A	A	B	B	0	B	B	C, D	B	E	B<MK?
82	Dull	A	A	A	B	B	0	B	B	B	B	B	
83	To work	A	B	B, C	D	D	D	D	D	D	C	C	
84	To plant	A	A	A~	A	A	A	A	A	A	0	A	
85	To choose	A	A	A	B	B	B	B	B	B	B	B, C	B<MK, C<UNK
86	To grow	A	A	A	A	A	A	A	B	C	A	A	
87	To swell	A	A	A	B	A, B	A, B	A	B	B	A, B	C	B<MK
88	To squeeze	A	A	B	A	B	A	B	A, B	B	B	B	B<MK
89	To hold	A	B	C	D	D	E	D	D, E	E	D, E	D	D<UNK?, E<MK
90	To dig	A	A	B	A	A	A	A	A	A	A	B?	
91	To buy	A	A	A	A	A	A	A	A	A	A	A	
92	To open	A	A	B	C	C	C	C	D	D	D	E	
93	To pound	A	B	C	C	C	C	C	D	E	C	A	C<MK; D, E < UNK
94	To throw	A	B	C	D	E	F	F	D	G	D	H?	
95	To fall	A	B	C	A	A	A	A	A	A	A	A	
96	Dog	A	B	A	A	A	A	A	A	A	A	A	
97	Bird	A	B	C	C	C	C	C	C	C	C	C	C<MK
98	Egg	A	A	A	A	A	A	A	A	A	A	A	
99	Feather	A	A	A	A	A	A	A	A	A	A	A	
100	Wing	A	A	A	A	A	0	A	B	A	C	=1	
101	To fly	A	B	C, D	D	D	D	D	D	D	D	D	D<MK
102	Rat	A	A	A	A	A	A	A	A	A	A	A	
103	Meat	A	A	A	B	B	A	B	A, B	A, B	B	A	B<MK
104	Fat	A	A	A	A	A	A	A	A	A	A	A	
105	Tail	A	A	A	A	A	A	A	A	A	A	A	
106	Snake	A	A	A	A	A	A	A	A	A	A	A	
107	Worm	A	A	A	A	A	A	A	A	A	A	A	
108	Louse	A	A	A	A	A	A	A	A	A	A	A	
109	Mosquito	A	A	A	A	A	0	A	B	C	A, D	A	
110	Spider	A	A~	B	C	D	E	E	E	E	F	A	D<Khmer, E<MK, F<BAH
111	Fish	A	A	A+	A	A	A	A	A	A	A	A	A+<MAL?
112	Rotten	A	A	A	A	A	A	A	A	A	A	A	
113	Branch	A	A	A	A	A	A	A	A	A	A	*	* compound using item 79
114	Leaf	A	A	B	C	C	C	C	C	C	C	C	C<MK
115	Root	A	A	B	B	B	B	B	B	B	B	B	
116	Flower	A	A	A	A	A	A	A	A	A	A	A	
117	Fruit	A	A	A	A	A	A	A	A	A	A	A	
118	Grass	A	B	B	C	C	B, C	B, C	B, C	B, C	B, C	B	C<MK?
119	Earth	A	A	A	B	B	B	B	B	B	B	B	B<UNK
120	Stone	A	A	A	A	A	A	A	A	A	A	A	
121	Sand	A	B	B	C	C	C	C	C	C	C	C	C<MK
122	Water	A	A	B?	B	B	B	B	B	B	B	B	B<PMP?
123	To flow	A	B	C	C	C	C	C	C	C	C	C	C<MK
124	Sea	A	B	B+	A	A	A	A	A	A	A	A	B+<MAL (= PMP 'sea-ward'
125	Salt	A	A	A	A	A	A	A	A	A	A	A	
126	Lake	A	A	A	A	A	A	A	A	A	A	*	*=compound from PMP elements
127	Forest	A	B	B	B, C	C	C	C	C	C	C	C	C<UNK
128	Sky	A	A	A	A	A	A	A	A	A	A	A	
129	Moon	A	A	A	A	A	A	A	A	A	A	A	
130	Star	A	B	A	A	A	A	A	A	A	A	A	
131	Cloud	A	B	B	C	C	C	C	C	C	C	C	B<UNK, C<MK

132	Fog	A	A	B, C	C, D	C, D	0	C, D	C, D	C, D	C, D	E?	C<MK<SKT, D<MK
133	Rain	A	A	A	A	A	A	A	A	A	A	A	
134	Thunder	A	B	C	D	D	D	D	D	D	D	D	B, C, D all similar, possibly cognates
135	Lightning	A	A	A	B	B	0	B	B	B?	B	C	
136	Wind	A	A	A	A	A	A	A	A	A	A	A	
137	To blow	A	A	A	A	A	A	A	A	A	A	A, B	B<UNK
138	Hot	A	A	B	C	C	0	C	C	C	D	C	
139	Cold	A	B	C	C	C	C	C	C	C	C	C	C<MK
140	Dry	A	A	B	B	B	C, D	B	B, C	D	B	B	B<UNK, C<UNK
141	Wet	A	A	A	A	A	A	A	A	A	A	A	
142	Heavy	A	A	A, B	C?	C	C	C	C	C	C	C	
143	Fire	A	A	A	A	A	A	A	A	A	A	A	
144	To burn something	A	B	C	D	D	0	D	D	C	B	C	
145	Smoke	A	B	B	B	B	B	B	B	B	B	B	
146	Ash	A	A	A	A	A	A	A	A	A	A	A	
147	Black	A	A	A	A, B	A, B	B	B	B	B	B	A	B<MK
148	White	A	A	A	A	A	A	A	A	B	B	A	B<UNK
149	Red	A	A	A	A	A	A	A	A	A	A	A	
150	Yellow	A	B	A	A	A	A	A	A	A	A	A	B<BTK
151	Green	A	B	B	B	B	C	C	C	C	C	C	A and C both <PMP
152	Small	A	B	C	B?	B?	D	D	D?	D	E	F	
153	Big	A	B	A, C	A, C	A, C	A, C	C	C	C	C	C	
154	Short	A	B	B	C	C	0	C	D	=152	B	B	
155	Long	A	A	A	A	A	A	A	A	A	A	A	
156	Thin	A	A	A	A	A	A	A	A	A	A	A	
157	Thick	A	A	A?	A	A	A	A	A	A	A	A	
158	Narrow	A	B	C	D	D	D	D	D	D	D	C	D<MK?
159	Wide	A	A	B	C	C?	0	C	C	C	D	C	D<MK
160	Sick	A	A	A	A	A	A	A	A	A	A	A	
161	Shy	A	A	A	A	A	A	A	A	B	A	B	B<UNK, it means 'fear' too in Rade
162	Old	A	A	A	A	A	A	A	A	A	A	A	
163	New	A	A	A	A	A	A	A	A	A	A	A	
164	Good	A	B	C	C	C	C	C	C	C	C	D	C<MK?
165	Bad	A	A	A	A	A	A	A	A	A	A	A	
166	True	A	A	A	B, C	B, C	B, C	B, C	C	C	C	D?	C<MK
167	Night	A	B	B	B	B	B	B	B	B	B	B	B=PMP 'evening'
168	Day	A	A	A	A	A	A	A	A	A	A	A	Form A in Malayic and Chamic is irregular in shape
169	Year	A	A	A	A	A	A	A	A	A	A	A	
170	When?	A	B	C, D	E	E	0	0	C	E	F	E?	
171	To hide	A	A	B, C	D	D	D	D	D	D	D	D	
172	To climb	A	A	A	A	A	A	A	A	A	A	B?	
173	At	A	A	A	A	A	A	A	A	A	A	A	
174	In	A	A	A	A	A	A	A	A	A	A	A	
175	Above	A	A	A	B	B	0	B	C	D	E	D	E<UNK
176	Below	A	A	B	B	B	B	B	B	B	B	B	B<UNK
177	This	A	A	A	A	A	A	A	A	A	A	A	
178	That	A	A~	A~	A~	A~	A~	A~	A~	A~	A~	A~	
179	Near	A	B	C	D	D	D	D	D	D	D	D	D<MK
180	Far	A	A~	B	B	B	B	B	B	B	B	B	
181	Where?	A	B	C	B	B	0	B	D	E	F	F	
182	I	A	A, B	A	A	A	A	A	A	A	A	A	B<SKT 'slave'
183	Thou	A	B	C	D	D	D	D	D	A, D	D	D	
184	S/he	A	B	C	A~	A~	A~	A~	A~	A~	A~	A~	
185	We	A	A	A	A	A	A	A	A	A	A	A	
186	You	A	A, B	=183	=183	=183	=183	=183	=183	=183	=183	=183	
187	They	A	B	C	=184	=184	=184	=184	=184	=184	D	=184	

188	What?	A	A	A	A	B	0	B	C	B	D	A~	
189	Who	A	A	A	A, B	A, B	B	B	B	B	B	B	B<UNK
190	Other	A	A	A, B	A, B	A, B	A	A, B	A, B	B	A	B	B<MK
191	And	A	B	B	C	C	0	C	C	C	C	B	C<MK
192	All	A	B	B	B	B	B	B	B	B	C	B	
193	If	A	B	B	C	D	0	D	E	F	E	D	B<SKT, E<CHI
194	How?	A	B	C	D	E	C	C	C	C	D	D	B< com-pound: TAM+PMP
195	No	A	B	C	C	C	A	A	C	C	C	C	C<UNK
196	To count	A	B	A	A	A	A	A	A	A	A	A	
197	1	A	A~	A	A	A	A	A	A	A	A	A	
198	2	A	A	A	A	A	A	A	A	A	A	A	
199	3	A	B	A	A	A	A	A	A	A	A	A	B<SKT
200	4	A	A	A	A	A	A	A	A	A	A	A	
201	5	A	A	A	A	A	A	A	A	A	A	A	
202	To sing	A	A	B	C	C	C	C	C	C	C	D	C<MK, D<CHI
203	To play	A	A	A	A	A	A	A	A	A	A	A	
204	We incl	A	A	A	A	A	A	A	A	A	A	A	

LEGEND: ACH (= Acehnese), AR(abic), BAH(nar), BTK (= Batak), CAUS(ative), CHI (Min Chinese), MA(lay), MK (Mon-Khmer, usually North or Centtral Bahnaric), SKT (Sanskrit), TAM(il), UNK(nown as to origin but usually reconstructible to an immediate proto-language such as Proto-Chamic). The use of the symbol ~ indicates that the language uses a morphologically aberrant development of a form which is nonetheless cognate with the PMP form. The use of + (in the Acehnese column) indicates that the form is related to the form whose letter it bears, but that it is actually a loan of this form from Malay, rather than being an inherited element. The sign 0 indicates that an equivalent for this gloss and in this language was not available to me. The cognacy of those items which are marked with a letter followed by ? with other items that are marked out with the same letter is indicated as yet being uncertain.

Table 2a. *Dyen's lexicostatistical percentages for selected Indochinese Chamic languages, using the Swadesh 200-item list and horizontal lexicostatistical techniques (Dyen 1971: 111).*

Cham				
73.0	Chru			
68.0	73.0	Roglai		
66.0	71.5	66.5	Jarai	
60.0	68.5	64.5	83.5	Rade

Table 2b. *Lexicostatistal percentages for certain Chamic languages using the Swadesh 200-item list and horizontal lexicostatistics (Thomas 1977: viii).*

Western Cham							
82	Eastern Cham						
75	76	Chru					
77	77	77	Southern Roglai				
71	71	72	71	Northern Roglai			
64	67	69	65	67	Haroi		
62	62	64	60	64	73	Jarai	
61	61	63	59	61	66	72	Rade

Table 3. *Selected morphological properties of Chamic and certain other relevant languages.*

Feature	Tagalog	Proto-Malayo-Chamic	Bahasa Melayu	Aceh-nese	Old/Inscriptional Cham	Written Cham	Phan Rang Cham	Tsat	Modern Chinese	Modern Khmer
Bound inflection	Yes	No?	No/yes	No	No	No	No	No	No	No
Prefixes	Yes	No?	No/yes	No	No	No	No	No	No	No
Infixes	Yes	No?	No	No	No	No	No	No	No	No
Suffixes	Yes	No?	No	No	No	No	No	No	No	No
Bound derivationals	Yes	Yes	Yes	Yes	Yes	Yes	Hardly	No	Emerging	Yes
Prefixes	Yes	Yes	Yes	Yes	Yes	Yes	Not productive	No	No?	Yes
Infixes	Yes	Yes	No	Yes	Yes	Yes	Not productive	No	No	Yes, non-productive?
Suffixes	Yes	No	Yes, but few	No	No	No	No	No	Emerging ?	No
Lexical tones	None	none	none	None	None	none	two	five	Six in Hainanese	none

The Proto-Malayo-Chamic language has not been reconstructed in detail and no descriptions of how it may have looked exist in the linguistic literature. The presence of certain kinds of morphological features in this language is inferred from the evidence of retentions of actual inherited morphemic forms (which are what I call 'fabric') in our records of Old Malay, Old Cham, modern Chamic languages, and in modern Malay and Acehnese. Prefixes and especially infixes were used more productively in Old and Middle Khmer than they are in Modern Khmer, which uses more free grammatical morphemes, though suffixes have never been used in Khmer (this issue is discussed further in Jacob 1963).

Grant, A.P. "Norm-referenced lexicostatistics and Chamic". In Grant, A. and Sidwell, P. editors, *Chamic and Beyond: Studies in mainland Austronesian languages.* **PL-569:105-146.** Pacific Linguistics, The Australian National University, 2005. DOI:*10.15144/PL-569.105*