Aslian linguistic prehistory

A case study in computational phylogenetics

Michael Dunn[a,b], Niclas Burenhult[b,c], Nicole Kruspe[d], Sylvia Tufvesson[b], and Neele Becker[b,e]
[a]Radboud University Nijmegen / [b]Max Planck Institute for Psycholinguistics, Nijmegen / [c]Lund University / [d]University of Melbourne / [e]Johannes Gutenberg Universität Mainz

This paper analyzes newly collected lexical data from 26 languages of the Aslian subgroup of the Austroasiatic language family using computational phylogenetic methods. We show the most likely topology of the Aslian family tree, discuss rooting and external relationships to other Austroasiatic languages, and investigate differences in the rates of diversification of different branches. Evidence is given supporting the classification of Jah Hut as a fourth top level subgroup of the family. The phylogenetic positions of known geographic and linguistic outlier languages are clarified, and the relationships of the little studied Aslian languages of Southern Thailand to the rest of the family are explored.

Keywords: Computational phylogenetics, evolutionary modeling, Orang Asli, Aslian languages, Mon Khmer languages, Austroasiatic languages

1. Introduction

Linguistics and biology have a history of methodological cross-fertilization. In *The Origin of Species*, Charles Darwin used the relationships of languages as a self-evident example of the process of genealogical diversification and drift (Darwin 1859). Most recently however, the locus of innovation in phylogenetics has been biology — an abundance of new methods are becoming available, enabling better quality inferences of ever more complex histories. In this paper we set out to compare genealogies of Aslian languages generated according to several different methods. Aslian languages are a subbranch of the Austroasiatic stock spoken in the Malay Peninsula, and they hold important information for understanding Southeast Asian prehistory. Quantitative methods provide new ways of investigating language variation

Diachronica 28:3 (2011), **291–323**. DOI 10.1075/dia.28.3.01dun
ISSN 0176–4225 / E-ISSN 1569–9714

that bring processes of language change to the fore. This reinvestigation of Aslian genealogical relationships is possible due to extensive word lists, in the most part collected by the authors. Methodological advances — enumerated below — have increased the amount of useful information which can be extracted from these data.

We will focus on two broad methods for estimating phylogenetic relationships. Firstly, we will work through some approaches using **distance methods**. Distance measures subsume what is known in linguistics as **lexicostatistics**, a measure of 'cognate distance'. Lexicostatistics uses an aggregate distance measure (shared cognate percentages), and can provide an effective heuristic for estimating phylogenetic relationships. Linguists have tended to investigate lexicostatistical comparisons by visual inspection of pairwise distance matrices; we will show that considerable additional information can be extracted from these distance matrices by applying tree- and network-drawing algorithms. The tree-drawing algorithm we will use is called **neighbor joining** (Saitou & Nei 1987, Wang 1996, McMahon & McMahon 2003); it provides a rigorous way of converting a matrix of distances into a tree structure. An alternate approach for investigating the same data is the **NeighborNet** method (Bryant & Moulton 2003, Ben Hamed 2005, Ben Hamed & Wang 2006), which can represent 'conflicting signal'. The limitations of lexicostatistics as a method of historical inference have been widely discussed, and we are not committed to the results of our lexicostatistical analysis; rather, we use it to link our work to previous quantitative analyses of Aslian language histories, and to contrast the richer results obtainable with more modern methods.

In contrast to the distance measures, **character-based methods** look at individual linguistic features, rather than treating them as an aggregate. Our method of choice, **Bayesian phylogenetic inference**, is a sophisticated tree-building technique, which models the evolutionary behaviour of individual features. While conceptually challenging, it has a number of advantages which make it worth pursuing. Phylogenetic trees produced by Bayesian phylogenetic inference incorporate a measure of confidence; they give a measure of amount of change on branches; the amount and rate of change can (at least in principle) be calibrated against historical and archaeological dates to estimate a chronology calibrated in years.

In general, these three methods give increasingly good estimates of the phylogeny. Neighbor joining is additionally interesting because it rigorously tests the generalizations that have been made on the basis of lexicostatistical matrices, and provides a useful first approximation of a phylogeny. The NeighborNet graph gives another useful summary of the same information, and provides in addition an estimate of how 'tree-like' the data are. Assuming tree-like data, Bayesian phylogenetic inference gives the best estimation of phylogeny available — and not only tree topology, but also of other useful statistical properties such as quantified uncertainty and quantified amount of change.

1.1 Aslian languages and their speakers

The Aslian languages form a subgroup of the Mon-Khmer language family, which itself is a division of the Austroasiatic stock. The 18 or so Aslian languages are mostly spoken in inland areas of Peninsular Malaysia and extend a little way into the south of Thailand. According to evidence from lexicostatistics (Benjamin 1976) and historical phonology (Diffloth 1975, 1979), the Aslian subgroup is traditionally divided into three major clades, referred to as Northern, Central and Southern (see Figure 1). Aslian is believed to have been introduced to the peninsula by Austroasiatic-speaking immigrants from the Southeast Asian mainland, possibly in connection with the arrival of agriculture some 4000–5000 ybp.

The three Aslian clades have been proposed to coincide broadly with three ethnographically defined subgroupings of indigenous cultures, characterised by distinct societal and economic features (Benjamin 1985). Thus, according to this widely accepted classification, Northern Aslian is by and large associated with nomadic foragers known ethnographically as the Semang; Central Aslian is associated with semi-sedentary swidden horticulturalists referred to as the Senoi; Southern Aslian, finally, is linked to groups of collectors-traders called Aboriginal Malay (and who, in Benjamin's work, show broader Malayic cultural patterns).

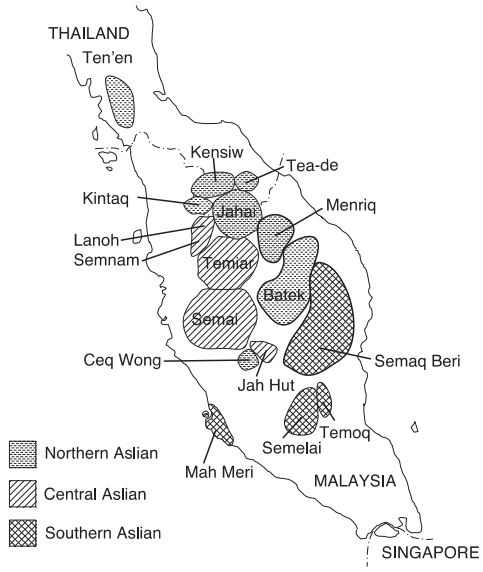

Figure 1. Map of the Malay Peninsula showing the approximate distribution of Aslian languages and subbranches according to the traditional tripartite classification. Language distribution is adapted from Benjamin (1976) and Burenhult (forthcoming).

However, several Aslian ethnolinguistic groups are difficult to classify according to this societal-economic division, either because they display a mix of societal-economic features, or because their linguistic identity does not match the expected cultural one (see further below). The connection between Southern Aslian and the Aboriginal Malay category is particularly problematic in light of ethnographic and linguistic work since Benjamin proposed his classification: Aboriginal Malay is a hybrid grouping consisting primarily of non-Islamic speakers of Malay-dialects who inhabit the southern half of the Malay Peninsula, and some Southern Aslian speakers.

The earliest genetic roots of the populations of the Malay Peninsula can be traced back into the Pleistocene, and perhaps to the initial dispersal of modern humans in Australasia 50,000–60,000 years ago. Secondary genetic colonisation from the north took place around 4000–5000 ybp, coinciding with the arrival of agricultural practices and, most likely, the ancestor of today's Aslian languages. There is genetic evidence that the current Aslian-speaking aboriginal groups represent descendants of both these stocks, but to varying degrees (Hill et al. 2006, 2007). Thus, the locally ancient genetic lineages are most evident in the Semang, who appear to have experienced only minor subsequent gene flow from outside the peninsula.[1] They are also evident in the Senoi, who nevertheless also carry lineages which can be traced to an origin in Indochina within the last 7000 years. The Aboriginal Malay are more diverse, and also display mid-to-late-Holocene lineages originating from Island Southeast Asia, possibly reflecting genetic influence associated with the arrival of Austronesian languages. However, the ancestor of the now dominant Austronesian language in the peninsula, Malay, came with a much more recent colonisation, possibly from Borneo around 1500 years ago (Collins 2006).

Most interpretations of the prehistory of the Malay Peninsula rely on Benjamin's 1985 classification of indigenous societal patterns and the associated Aslian family tree. Models by Rambo (1988) and Fix (1995) propose that Benjamin's categories can be projected into the past and suggest a common genetic, linguistic and cultural origin of all of the peninsula's indigenous groups (cf. Bulbeck 2004). On the basis of the more recent genetic studies cited above, Burenhult (forthcoming) questions the idea of a closely intertwined development of biology, language and culture. He instead elaborates a scenario combining elements of local genetic

1. Observers have long maintained that there is a close connection between the Semang foraging tradition and certain physical features of its bearers. The purportedly short stature, dark skin and curly hair of the Semang led early anthropologists to classify them physically as 'Negritos' (see e.g. Schebesta 1952), a term which to some extent is still used. It is also used to refer to physically similar hunter–gatherer populations in the Andaman Islands and the Philippines.

continuity with demic diffusion and language shift to account for the current distribution of Aslian languages, and genetic diversity of their speakers. According to this hypothesis, the ancestor of the Aslian languages entered the peninsula in connection with the introduction of agriculture. Some local foragers adopted the new economy, intermixing to a greater or lesser degree with the agriculturalist immigrants, but in any case adopting an Aslian language. The hypothesis further proposes that it was in this diverse setting of intermixture that the Aslian subbranches split from the proto-language in distinct situations of cultural and linguistic contact. Alongside this, some foragers retained their economy and nomadic lifestyle (the predecessors of today's Semang). At some point the pre-existing ties between the forager groups and the more settled, intermixed groups led to language shift such that the remaining foraging peoples also came to speak Northern Aslian languages.

Burenhult's hypothesis questions the customary equation of Aslian subbranches with the three cultural subgroupings (especially the wholesale association of Northern Aslian with foraging) and places emphasis on the ethnolinguistic groups which do not match the previously proposed societal-linguistic correlations. Thus, the idea that Northern Aslian crystallised in a non-forager setting is supported by the existence of one distinct and conservative Northern Aslian language spoken by a semi-sedentized group with a mixed economy which does not belong to the Semang forager sphere (Ceq Wong). Another mismatch is a small group of Central Aslian languages spoken by foragers (Semnam and Lanoh): most likely the result of a later language shift from Northern Aslian (Burenhult et al. forthcoming). The rest of the Central Aslian languages, as well as most Southern Aslian languages, are spoken by non-Semang. However, one Southern Aslian language (Semaq Beri, not discussed in Burenhult forthcoming) is spoken by people usually classified as foragers, although they do not belong to the Semang forager grouping (Kruspe forthcoming). Yet another anomalous language (Jah Hut), traditionally considered Central Aslian but difficult to classify with certainty, is spoken by people with mixed societal traditions difficult to assign to the proposed cultural categories.

2. The present study

The present study owes a great deal to the pioneering lexicostatistical classification of Aslian languages by Benjamin (1976, 1997), which has shaped much of the discussion of peninsular ethnography and prehistory. Our own study takes Benjamin's careful, reproducible lexicostatistics as its starting point, and uses recent innovations in phylogenetic inference. Our study also benefits from greatly increased knowledge of several branches and languages of Aslian which were

poorly documented at the time of Benjamin's study (e.g. Kruspe 2004, forthcoming, Burenhult 2005).

2.1 Outstanding questions

In this paper we provide a broad overview of the genealogical relationships between the Aslian languages, and address some outstanding questions in Aslian linguistic prehistory:

1. To what extent can we validate earlier classifications, especially the lexicostatistical classification by Benjamin (1976)?
2. Can we clarify the classification of languages which were considered to be outliers by Benjamin, i.e. languages which were genealogically problematic in his study, such as Ceq Wong, Jah Hut, and Mah Meri?
3. What is the classification of hitherto unclassified Aslian languages of Southern Thailand (varieties of Ten'en), as well as other undocumented varieties in Malaysia (Menriq Rual, Batek Teq), which were not included in Benjamin's study?
4. What can we contribute to the understanding of the population history of Aslian-speaking communities, including issues of demic diffusion, language shift, and local continuity?

2.2 Wordlist

A 146 item wordlist was compiled for 27 linguistic varieties belonging to the Aslian family (see Table 1); this set of data represents the presence of apparent cognate forms sharing particular basic meanings. The authors made all cognacy judgements, and identified Malay loanwords. Many of the meaning sets showed reflexes of more than one cognate set; once loanwords were excluded there were 472 phylogenetically informative cognate sets in the data (another 246 sets had only one member, so did not provide any usable information for grouping languages).

The list of basic meanings used in this study is taken from Benjamin (1976). It is essentially a modified version of the Swadesh 200-word list which he adapted to better suit the Aslian context (Benjamin 1976: 53). While not discussing the individual changes, he removed items from the Swadesh list according to the following criteria:

1. They did not fit with the semantic structure of Aslian languages, e.g. colour terms and some deictics,
2. They were not environmentally or culturally relevant, e.g. 'sea', 'to freeze', and

3. They were considered problematic in terms of the ability to collect accurate data, e.g. 'all'.

A few items of local relevance were also added, e.g. 'dance', 'rice (unhusked)' and 'shoot (blowgun)'.

Benjamin became aware of a Mon-Khmer-based list (Thomas 1960) only after most of his lists had been compiled. He noted that a new study revisiting the data should take advantage of Thomas' list which had been compiled specifically for the Southeast Asian region (Benjamin 1976: 54). Based on our own experience we considered compiling a revised Aslian list for the current study, but decided to retain Benjamin's original list because it would allow for an easier comparison with the original study, and allow for the possibility of incorporating Benjamin's data for languages where we did not have our own data. As it happened, it was not necessary to incorporate any of Benjamin's data except for his Temiar list — based on extensive fieldwork and knowledge of the language.

As noted above, a major impetus for the current work was the fact that we now have detailed knowledge of several more Aslian languages from more branches. This reflects primary research done in the last 10–20 years. When Benjamin did his study there was no expert knowledge available on Southern Aslian and very little on Northern. Eight of our lists are based on long-term work and represent languages of which the researchers have speaking knowledge. Collection of these lists also involved interviewing in the target language rather than Malay (even if the Malay reflexes may sometimes have been used for elicitation), and many of the other lists have been collected using a commonly spoken Aslian variety as contact language. Furthermore, most of the current lists were collected *in situ*, or at least at the current site of habitation of the consultant, rather than relying — as Benjamin did to a great extent — on the availability of consultants at the Orang Asli Hospital in Gombak, outside Kuala Lumpur.

2.3 Language sample

Benjamin's sample of languages was based primarily on the ethnic groups recognised in Malaysian administrative practice, and was supplemented with unrecognised groups deemed linguistically relevant (see Figures 2 and 3 for the languages in his sample, and with his original orthography). Our own somewhat larger sample (Table 1) is based on similar principles but is different in some important respects:

– For some languages we have included more dialect varieties than Benjamin did, e.g. Jahai, Kensiw, Menriq, Semaq Beri, and Temiar.

- Our sample from Batek includes partly different varieties: unlike Benjamin, we do not have data from Batek Nong or Mintil but Batek Teq and Batek Teh instead. Both of these are known but hitherto undocumented varieties. Batek Teq is a moribund variety now spoken by only 5–6 families in the Besut district of northern Terengganu (identified as a distinct ethnic group by Endicott 1979). Batek Teh is claimed to be identical to Menriq rather than a variety of Batek proper (Benjamin 1976: 47 cf. §4)
- Menriq spoken at Rual (Jeli, Kelantan), from which we have a list, is a distinct and previously unrecorded variety of Menriq not analysed by Benjamin. It is sometimes referred to by neighboring groups as Jedek and by themselves as Menrik.
- Benjamin includes four varieties of the cluster known officially as Lanoh: Lanoh Yir, Lanoh Jengjeng, Semnam, and Sabüm. We include two varieties of Semnam and one variety of Lanoh which we label Kertei (the original location of the consultant). The latter most likely corresponds to Benjamin's Yir and is not actively spoken anymore but remembered by a handful of elderly speakers. We have not been able to identify a separate contemporary variety known as Jengjeng; Sabüm is now reported extinct (Burenhult 2006).
- We include two varieties of Ten'en, a geographically isolated group of dialects spoken in the Trang, Pattalung, and Satun provinces of southern Thailand. Benjamin did not have access to data from Thailand (Benjamin 1976: 50,94), and this is the first time this language is compared lexicostatistically to its relatives further south (cf. Diffloth 1975 for a preliminary phonological comparison).
- Our Ceq Wong list is from the eastern dialect, whereas Benjamin's was from the western one; likewise, our lists from Mah Meri and Semaq Beri are from other varieties than Benjamin's.
- Benjamin includes a list from Temoq, an anomalous group closely related to Semelai and Semaq Beri. We did not have access to speakers.
- Like Benjamin, we include two varieties of Semai, but different ones. However, the sample is similar in that both include one lowland and one highland variety.

The present sample is at least as wide-ranging as that of Benjamin, and in some respects it reflects more detail as well as including previously unanalysed varieties. However, it cannot claim to provide full coverage of Aslian. Thus, there are languages and dialects which are not included but which would have added to the study, like the above-mentioned Batek Nong, Mintil, and Temoq. Also, the many diverse dialects of Semai are underrepresented in the current sample. Furthermore, a recently identified variety in Thailand called Tea-De (Phaiboon 2006), presumably closely related to Kensiw, is not represented in our sample.

Table 1. Sources. Asterisk (*) indicates lists of the highest reliability, where the collector has done long term work and has speaking knowledge of the target language.

Variety	Collected by	Location
Ten'en Palian	Becker 2008	Ton Tok, Palian, Trang, Thailand
Ten'en Paborn	Becker 2008	Paborn, Pattalung, Thailand
Kensiw Perak	Burenhult 2005	Sungai Lebey, Hulu Perak, Perak, Malaysia (speaker/s from: Betong, Yala, Thailand)
Kensiw Kedah	Burenhult 2005	Bukit Asu, Hulu Perak, Perak, Malaysia (speaker/s from: Lubok Legong, Baling, Kedah, Malaysia)
Kintaq	Burenhult 2005	Bukit Asu, Hulu Perak, Perak, Malaysia (speaker/s from: Lubok Legong, Baling, Kedah)
Jahai Banun	Burenhult 1998–2008*	Sungai Banun, Hulu Perak, Perak, Malaysia (speaker/s from: Sungai Mangga, Hulu Perak, Perak)
Jahai Rual	Burenhult 2000–06*	Sungai Banun, Hulu Perak, Perak, Malaysia (speaker/s from: Sungai Rual, Jeli, Kelantan)
Menriq Lah	Burenhult 2006, 2008	Kuala Lah, Gua Musang, Kelantan, Malaysia
Menriq Rual	Burenhult 2005	Sungai Rual, Jeli, Kelantan, Malaysia
Batek Teh Taku	Burenhult 2006	Kuala Krai, Kelantan, Malaysia
Batek Teh Lebir	Burenhult 2006	Pos Lebir, Kuala Krai, Kelantan, Malaysia
Batek Teq	Kruspe 2008	Sungai Berua, Hulu Terengganu, Terengganu, Malaysia (speaker/s from: Kampong Sayap, Besut, Terengganu)
Batek Deq Terengganu	Kruspe 2001, 2008	Sungai Berua, Hulu Terengganu, Terengganu, Malaysia (speaker/s from: Kuala Koh, Gua Musang, Kelantan)
Batek Deq Koh	Burenhult 2006	Kuala Koh, Gua Musang, Kelantan, Malaysia
Ceq Wong	Kruspe 2002–06*	Kuala Gandah, Temerloh, Pahang, Malaysia
Semnam Bal	Burenhult 2006–08*	Air Bah, Hulu Perak, Perak, Malaysia
Semnam Malau	Burenhult 2004	Sungai Banun, Hulu Perak, Perak, Malaysia (speaker/s from: Malau, Hulu Perak, Perak)
Lanoh Kertei	Burenhult 2004	Sungai Banun, Hulu Perak, Perak, Malaysia (speaker/s from: Sungai Kertei, Hulu Perak, Perak)
Temiar Kelantan	(Benjamin 1976)*	Perolak, Gua Musang, Kelantan, Malaysia
Temiar Perak	Burenhult 2004	Sungai Banun, Hulu Perak, Perak, Malaysia (speaker/s from: Ulu Grik, Hulu Perak, Perak)
Semai Ringlet	Burenhult 2005	Sungai Banun, Hulu Perak, Perak, Malaysia (speaker/s from: Ringlet, Cameron Highlands, Pahang)

Table 1. (*continued*)

Variety	Collected by	Location
Semai Kampar	Tufvesson 2006–08*	Batu Berangkai, Kampar, Perak, Malaysia
Jah Hut	Kruspe 2002, 2005	Kuala Gandah, Temerloh, Pahang, Malaysia (speaker/s from: Paya Terbol, Temerloh, Pahang)
Mah Meri	Kruspe 2000–06*	Bukit Bangkong, Sepang, Selangor, Malaysia
Semaq Beri Berua	Kruspe 2001, 2007–08*	Sungai Berua, Hulu Terengganu, Terengganu, Malaysia
Semaq Beri Pergam	Kruspe 2007–08*	Sungai Pergam, Kemaman, Terengganu, Malaysia
Semelai	Kruspe 1990–91, 2000–01*	Tasek Bera, Bera, Pahang, Malaysia
Mon	Kruspe 2009	Melbourne, Australia (speaker from: Bilugyn Island, Myanmar)
Khmer Surin	Dunn 2009	Siem Reap, Cambodia (speaker from: Surin, Thailand)
Khmer Siem Reap	Dunn 2009	Siem Reap, Cambodia
Kammu	Burenhult 2009	Lund, Sweden (speaker from: Rmcùal, Louang Namtha, Laos)

2.4 Cognate coding

The analysis draws on the coding of the reflex of each basic meaning as a potential cognate at the Proto-Aslian level. Thus, there is no a priori recognition of the established subbranches, and possible intra-Aslian loans are included in the analysis.

The primary criterion used here for identifying likely cognates targets the consonants of the last syllable of forms. To be coded as cognates, forms need to share the same place of articulation in both the consonant onset and coda of that syllable. It is a principle which can be applied unproblematically to the sample, since the final syllable in most Aslian languages always has the structure /CVC/. The reason for this criterion is that Aslian languages, like other Mon-Khmer languages, seldom have suffixes, and the end of the word is therefore usually part of the root and not affected by synchronic or diachronic morphophonemic processes. Also, the final syllable is the most informative part of a word in that it always receives stress and contains the greatest phonemic variation (Diffloth 1976: 102, p.c.).

There are two types of exceptions to the place-of-articulation criterion. One is where this criterion can be shown to apply too restrictively, i.e., where it excludes cognacy when such cognacy is obvious for other reasons. In the current data set, this exception applies to cases where — e.g. a sound change, or phonotactic

change, or some (perhaps obsolete) morphological operation — can be shown to have affected the consonants of the final syllable. If there is systematic language-internal evidence for the changes in question, forms have been coded as cognates. Examples of identifiable changes include a systematic replacement of the rhotic /r/ with the palatal approximant /y/ in Semnam; a replacement of palatal codas with alveolar ones in one variety of Semaq Beri; absence of some final syllable codas in Kensiw, Kintaq, Mah Meri, Semelai, Semaq Beri, and Ten'en; and occasional instances of final syllable onsets which can be shown to result from infixation of a consonant into the final syllable, e.g. two cases of /m/ infixation in Palian Ten'en.

The second type of exceptions are those where the place-of-articulation criterion would include forms which adhere to it by chance, and not because of cognacy. Such forms have been excluded only if there is clear evidence of chance resemblance, e.g. by the established existence in a language of a more likely cognate which happens not to be represented in the current data set, or if two similar forms are known to derive from different sources, e.g. two different reconstructable proto-forms. In the data, this criterion applies mainly to the forms for "water", where /tɔm/ and /(b)tew/ type words adhere to the place-of-articulation principle but for which two distinct forms can be reconstructed at the Proto-Mon-Khmer level (Diffloth, p.c.).

Forms which can be identified as loans from Malay, as well as occasional forms which can be shown to have been borrowed historically from an Austronesian source other than Malay, were identified and excluded from the analysis.

Our cognate coding criteria are broadly similar to those used by Benjamin. The main differences are: (1) We allow the initial consonant of final syllables to vary as long as place of articulation is the same; with the exception of palatals and liquids, Benjamin requires root-initial consonants to be identical. For example, we code the forms /dɛʔ/ and /tɛ̃ʔ/ "water" as cognates, whereas these do not fulfill Benjamin's requirements; (2) Benjamin allows for cases where alteration of one or two phonetically similar phonemes would change the words into obvious cognates; here we have been more restrictive and insist upon same place of articulation in onset and coda of the final syllable.

3. Methods and results

The basic division into three Aslian groups can be established using the comparative method (Benjamin 1976, Diffloth 1975, 1979). The Northern Aslian group can be distinguished from the other groups by several regular correspondences. Important amongst these is the phonologically unusual correspondence between forms of /i/ and /e/ in Northern Aslian, and /a:/ in Central Aslian (which is retained from Proto-Mon-Khmer, Diffloth 1975: 6).

Table 2 shows cognate sets illustrating the distinction between Northern Aslian and the rest. These terms are all inherited from Proto-Mon-Khmer, reconstructions of Proto-Mon-Khmer forms from Shorto et al. (2006).[2]

Table 2. Cognate sets from Northern, Central and Southern groups (separated by horizontal lines) showing correspondences e.g. between Central/Southern Aslian /a:/~/a/ and Northern Aslian /i/~/e/; between Central/Southern Aslian /s/ and Northern Aslian /h/; and between Central long vowel and Southern/Northern short vowel

	bone	**eat**	**leaf**	**tail**	**tongue**	**hair**
Proto-Mon-Khmer	*cʔaaŋ	*caʔ(a)	*slaʔ(a)	*staʔ(a)	*l(n)taak	*suuk
Ten'en Palian	ʔiyeŋ			hatĩʔ	ltik	sɔk
Ten'en Paborn	ʔiyeŋ		hliʔ	hatiʔ		sɔk
Kensiw Perak	ʔiyeŋ	ciʔ	hliʔ	htĩʔ	lntik	sɔk
Kensiw Kedah	ʔiʔeŋ	ciʔ	hliʔ	htiʔ	ltik	sɔk
Kintaq	ʔiʔeŋ	ciʔ	hliʔ	htiʔ	ltik	sɔk
Jahai Banun	jɾeŋ		haliʔ	hatĩʔ	lntek	sɔk
Jahai Rual	jɾeŋ		slaʔ(b)	hatĩʔ	lntek	sɔk
Menriq Lah	jɾiŋ	ciʔ	haliʔ	hatɛ̃ʔ	lntik	sɔk
Menriq Rual	jɾiŋ	ciʔ	haliʔ	hatɛ̃ʔ	lntik	sɔk
Batek Teh Taku	jɾiŋ	ciʔ	haliʔ	hatɛ̃ʔ	lntik	sɔk
Batek Teh Lebir	jɾiŋ	ciʔ	haliʔ	hatɛ̃ʔ	lntik	sɔk
Batek Teq			haliʔ	hacɛ̃ʔ	lntik	sɔk
Batek Deq Terengganu		ciʔ	haliʔ	hacɛ̃ʔ	ləntik	sɔk
Batek Deq Koh		ciʔ	haliʔ	hacɛ̃ʔ	ləntĩk	sɔk
Ceq Wong	jɾeŋ	cəʔ	haliʔ	hateʔ	latek	sɔk
Semnam Bal	jɾa:ŋ		sla:ʔ	snta:ʔ	lnta:k	
Semnam Malau	jɾa:ŋ		sla:ʔ	snta:ʔ	lntag	
Lanoh Kertei	jɾaŋ		slaʔ	sntaʔ	lntak	sɔok
Temiar Kelantan	jəʔa:k	ca:ʔ	səla:ʔ	sɛnta:ʔ	lɛnta:g	so:g
Temiar Perak	jɾɛ̃:ŋ	ca:ʔ	sla:ʔ	sntaʔ	lntak	so:k
Semai Ringlet	jɾa:k	caʔ	sla:ʔ	sntaʔ	lnta:k	sɔ:k
Semai Kampar	jɾʊ:k	cʊ:ʔ	slʊ:ʔ	snta:ʔ	lnta:k	sʊ:k
Jah Hut	jəʔaŋ	caʔ	hlaʔ(c)	sntaʔ	lntak	sɔk

2. Shorto's comparative dictionary of Mon-Khmer is a posthumous publication of a copious but unfinished work, and many of the reconstructions should be considered to be provisional. It is likely that improved reconstructions will become available (Diffloth 2008).

Table 2. (*continued*)

	bone	eat	leaf	tail	tongue	hair
Mah Meri	jəʔaḳ	naca				sup/suk
Semaq Beri Berua	jəʔaŋ	ɲca	sala	hateʔ(d)		suk
Semaq Beri Pergam	jʔaŋ		salah	hateʔ(d)		suk
Semelai	jʔaŋ	ca				suk

(a) We are unsure whether these reconstructions from Shorto et al. (2006) are correct, since we would expect a long vowel (see fn. 2)
(b) Central/Southern Aslian loanword into Northern
(c) This form shows *s→/h/, like the North, but it retains *a→/a/ like Central and Southern Aslian
(d) Northern Aslian loanword into Southern

3.1 Distance methods

Lexicostatistical comparisons produce crude classifications of languages based on rates of cognacy between all pairs of languages. No reconstruction is involved, so the 'cognates' here are proposed, apparent cognates, not established ones. Once the percentage of apparent cognates has been calculated for every pair of languages in the sample, some kind of tree-drawing algorithm is used to represent these distance relationships.

Lexicostatistical data has commonly been used as part of the method of glottochronology. Many criticisms have been levelled at glottochronology, the major one being that it assumes that the rate of linguistic change is constant. This has long been known to be false. Some of the disrepute of glottochronology has adhered to lexicostatistics as well, perhaps undeservedly. Lexicostatistics are, within their limits, a useful heuristic method for determining the approximate shape of a linguistic family tree, and many elaborate historical linguistic reconstructions originate in the testing and refinement of lexicostatistically generated hypotheses.

The interpretation of lexicostatistical data is usually done impressionistically. This is an unfortunate way to treat numeric measures. We propose using a rigorous tree drawing algorithm such as *Neighbor Joining* (Saitou & Nei 1987, Gascuel & Steel 2006) to summarize the linguistic distances in a principled manner. A neighbor joining tree starts from an unresolved, 'star' phylogeny, and recursively collapses the nearest pairs of nodes into branches until the tree is completely resolved (see Appendix A.1).

Benjamin gives a table of lexicostatistical percentages for Aslian languages, and manually draws trees of the three Aslian subgroups (Benjamin 1976: 58–59). Figure 2 shows these same lexicostatistical percentages drawn as a neighbor joining tree. The details of the subgroup trees are the same for Southern and Central Aslian, but visual inspection was apparently insufficient to detect that Jah Hut was

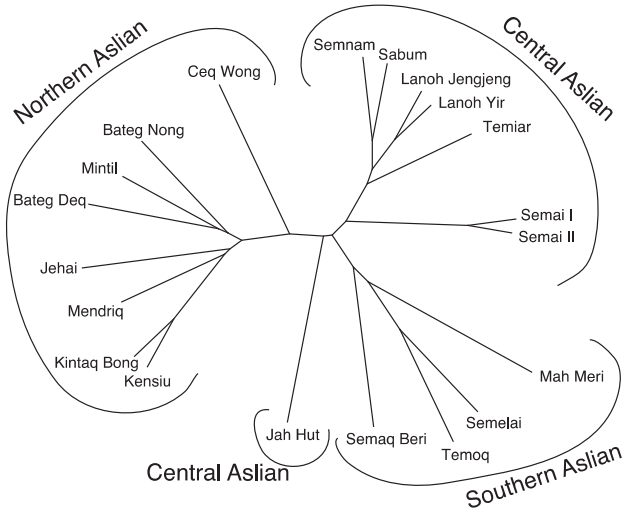

Figure 2. Neighbor-joining tree of Benjamin's lexico-statistical distances.

not subgrouped with the rest of Central Aslian. It is however unclear how well the tree represents the underlying distance data. In an encyclopedia article Diffloth & Zide (1992) claim that Jah Hut forms a fourth top-level subgroup of Aslian. Diffloth (p.c.) bases the reappraisal on patterns of vowel change, and has also considered a closer connection between Jah Hut and Southern Aslian for the same reason.

Figure 3 shows Benjamin's lexicostatistical distances as a NeighborNet splits graph instead of a tree (see Appendix A.2). This clarifies the position of Jah Hut — it is quite divergent from all the Southern and Central languages, but is closer to Semai I and II than the rest. Other interesting aspects of this network: Ceq Wong shows a considerable amount of conflict. Overall however the network is reasonably treelike, with a number of heavily weighted splits separating groups of languages. We will refrain from further discussion of Benjamin's data in favor of our own more extensive language sample.

Figure 4 shows the results of the sample NeighborNet analysis carried out on our own data. This network shows major splits separating the North Aslian group and also the Southern Aslian group; Central Aslian looks more like a residual group. An interesting feature (also present in Figure 3) is the secondary split between the languages of the Semang and non-Semang groups (indicated by the dashed line). This split partially conflicts with the major splits distinguishing the three subgroups, and shows evidence for clustering the three Central Aslian languages spoken by Semang-type foragers with the Semang Northern Aslian languages (that is, Northern Aslian languages apart from Ceq Wong) instead of their genealogical subgroup. This is evidence for contact (Burenhult et al. forthcoming).

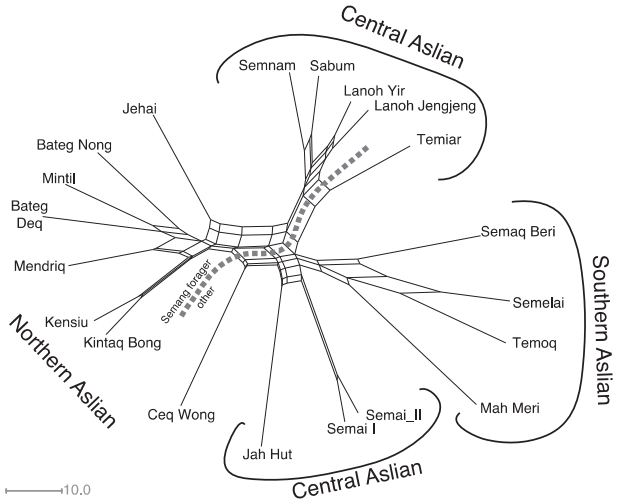

Figure 3. NeighborNet splits graph of Benjamin's lexicostatistical distances showing regions where the distance relationships are not commensurable with the broad three-subgroup classification of Aslian — invisible in the tree representation of the same data in Figure 2.

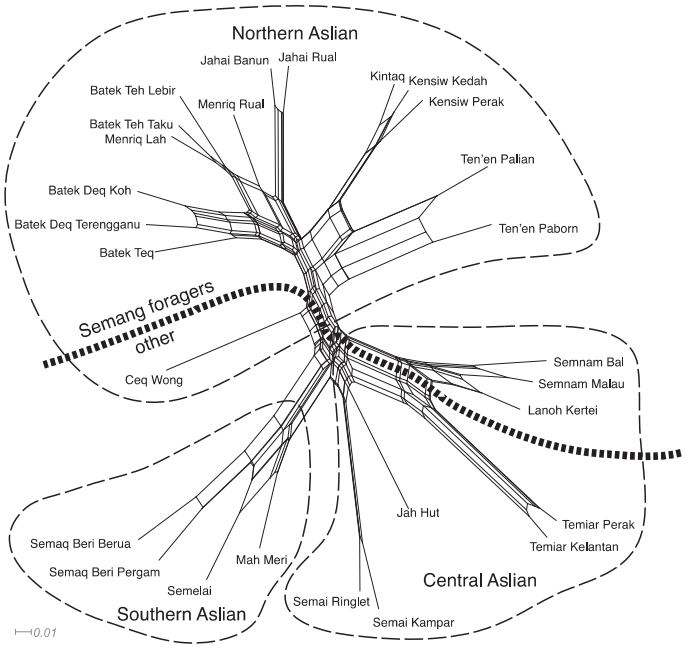

Figure 4. NeighborNet clustering of Aslian basic vocabulary. The network represents binary splits of lexical data in the form of reflexes of 146 basic meanings in 27 Aslian language varieties. The method measures lexical similarity without a priori recognition of established subbranches, and the network should therefore not be regarded as a 'family tree'.

The difference in the position of Jah Hut in Figure 3 and Figure 4 is more apparent than real: in both cases Jah Hut is joined by a long branch (i.e. much divergence) to a central, undifferentiated node of the graph. Lexical distance measures are not very informative about the likely position of Jah Hut.

3.2 Character-based methods

The weakness of any kind of distance measure for phylogenetic inference is that it ignores the content of similarities between languages. A simple distance measure like those discussed only approximates the historical structure of the similarities between languages — we know how different languages are, but a distance measure is not very informative about the processes by which this difference came about. The more effective approach to phylogenetic inference is to use **character-based methods**, that is, methods which infer history on the basis of a model of the behaviour of the individual linguistic features present in the data set. Our method of choice is **Bayesian phylogenetic inference** (see Appendix A.3), a probabilistic, model based approach originally taken from computational biology, but which can be advantageously used with linguistic data to produce family trees with a number of important properties which are absent from trees produced by traditional historical linguistics. The two properties of the greatest interest to us are (i) that each branch of a tree produced by Bayesian phylogenetic inference incorporates a measure of uncertainty, and (ii) that each branch has a measure of the amount of change over that branch (usually indicated by branch length) (Greenhill & Gray 2005, Dunn et al. 2008). 'Amount of change' is not always straightforward to interpret, as it is a function of both rate of change and length of time (e.g. a branch can be long because things were changing fast, or because it represents a long period of changes, or both). With enough historically and archaeologically dated calibration points within the tree, and some assumptions about smoothly varying rates of change, it is possible to convert branch lengths into a real chronology (Gray & Atkinson 2003); unfortunately, there are not enough known calibration points (i.e. dated branching events) to produce a reliable chronology for Aslian languages, so we have chosen to report only the compound measure showing amount of change.

Dunn (2008) gives a non-mathematical introduction to the techniques of Bayesian phylogenetic inference. Following Gray and others, we have employed the *cognate-gain, word loss* model of language change, which treats languages coded for reflexes of cognate classes. We have constructed a comparative word list and coded all the forms for probable cognacy. As stated above, from a list of 146 meanings we found 472 phylogenetically informative,[3] cognate sets. Malay

3. These are sets with more than one member, so that they may be significant for subgrouping.

loanwords had been removed, but intra-Aslian loans — in the cases where we could identify them — were left.[4]

The result of a Bayesian phylogenetic inference is a statistical sample of trees that best explain the observed linguistic data. In cases where a tree model can explain the data well, these trees will be very similar to each other. Figure 5 summarizes the results of the Aslian analysis with a maximum clade credibility (MCC) tree — the tree which has the highest overall probability of all trees in the

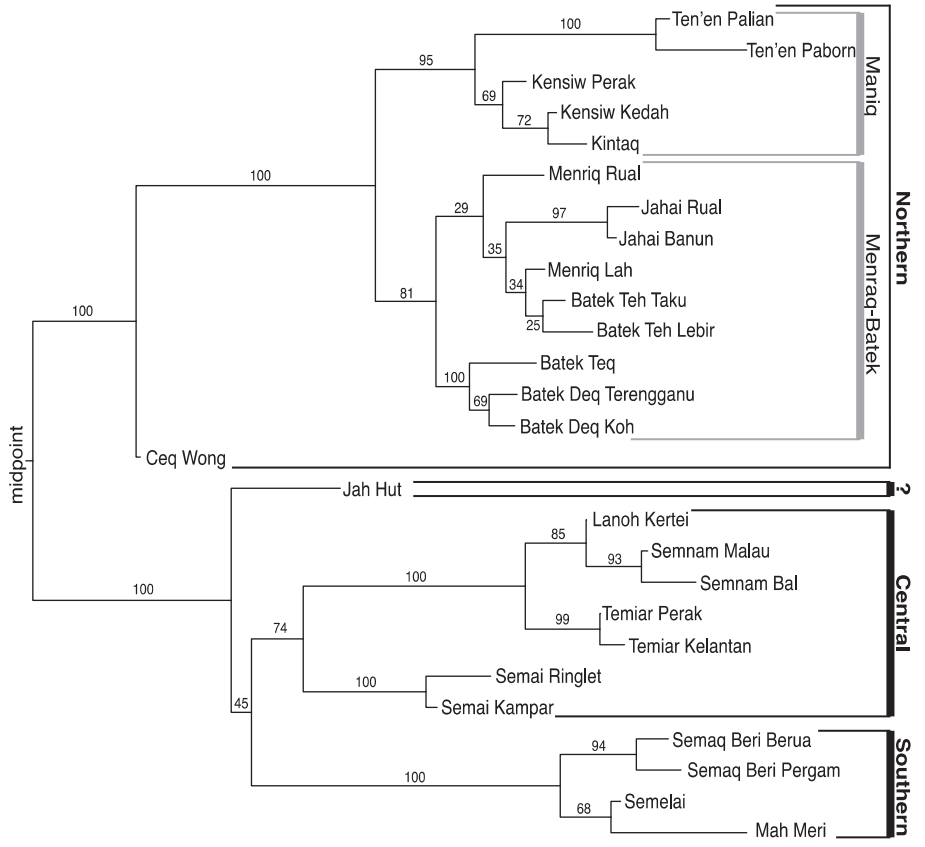

Figure 5. Maximum clade credibility tree of Bayesian tree sample. Branch length conflates rate of change and length of time to indicate the *amount of diversification*; numbers indicate *posterior probability*, the confidence we can have in each branch (expressed as a percentage; thus, Northern and Southern form distinct clades with 100% support, the Central languages form a subgroup with 74% support, and Jah Hut can't be allocated with confidence to any parent clade). This tree is rooted on the midpoint.

4. See for example loans identified by notes (a) and (c) to Table 2.

sample (calculated by multiplying the probability of all the nodes together). The percentage of the tree sample supporting each branch is marked: branches with less than 50% support should not be taken seriously. The MCC tree is congruent with the more certain aspects of the comparative method reconstruction. With the exception of Jah Hut, which is in any case questionable (see §4), the tree confidently recapitulates the comparative method classification into Northern, Southern, and Central subgroups. This tree supports the hypothesis that Jah Hut is a top level subgroup of Aslian; the low support (45%) for the Jah Hut branch indicates that the branching order of Jah Hut and the Southern and Central subgroups can't be given with confidence, and so should best be treated as a trifurcation. The other instances of low support in the tree (<50%) are all within the Menraq-Batek clade. The short branch lengths within this clade indicate that most of these languages are closely related, and suggests that the relationships between them might be better modeled as a dialect continuum rather than a tree.

Note that the MCC tree is only an approximate summary of the true result. The true result of the analysis is the tree sample produced by Bayesian phylogenetic inference, a set of more-or-less equiprobable phylogenetic trees whose variance tells us which aspects of the tree can be identified with certainty, and what the distributions of topologies of the less certain aspects are like (the MCC tree is better at summarizing the former than the latter since, for example, branches with 100% support are present in all trees of the sample).

3.3 Rooting

A non-directional (unidirectional) model of change — where the rate of gain of a trait is not differentiated from the rate of loss — produces an unrooted tree. Without a root the tree lacks any chronological dimension: it is impossible to determine that any node is earlier or later than any other. A bidirectional model does not solve this problem, since it requires a priori specification of the root. In the phylogenetic classification reported above, we chose to use a unidirectional model of evolution rather than the bidirectional one, since the added complexity of the bidirectional model did not provide any advantage in the form of significantly higher likelihood scores (the reason the non-directional model works relies on the presence of a covarion parameter, see §A.3). While from a historical perspective a rooted tree is more interesting, in most forms of phylogenetic analysis the root must be specified in advance. A lot of the interpretation of a tree thus depends on the rooting, and a poor (or unlucky) choice of root can have serious implications for the accuracy of the result. We tested several methods of determining the root and determined that they are in general agreement.

The rooting shown in Figure 5 is a *midpoint* rooting. This means that the root of the tree is drawn midway between the two most distant taxa. This makes a simplifying assumption that the amount of change that these two taxa have undergone is equal. This assumption is unlikely to be true, but it is nevertheless a useful approximation, as we can have some confidence that the lower level clades of a midpoint-rooted tree are monophyletic. The real historical root of the tree is best determined by *outgroup* rooting. An outgroup consists of one or more languages which are known to be a sister group to the languages under investigation (the ingroup). The node of the tree between the outgroup and the ingroup is the ancestor to both, and thus the root. The branch that joins the root to the ingroup determines the ancestral node of the ingroup. Word lists from three potential outgroups were transcribed and coded for possible cognacy according to the same principles used with the Aslian data. These outgroups were *Mon* (Monic branch of Mon-Khmer), *Kammu* (Khmuic branch of Mon-Khmer) and two closely related varieties of *Khmer* (Siem Reap Khmer of Cambodia and Surin Khmer of Thailand, both of the Khmer branch of Mon-Khmer).

Despite the obvious advantages of having a rooted tree, not every choice of outgroup is equally informative, and a bad choice of outgroup can result in a tree which is actively misleading (Sanderson & Shaffer 2002: 61). Furthermore, the particular outgroup selected can affect the topology of the ingroup in unpredictable ways (Milinkovitch & Lyons-Weiler 1998). In the present analysis, using all outgroup languages together (Mon, Kammu and the two Khmer dialects) as outgroup resulted in a tree with ingroup topology very different to the topology of the midpoint rooted tree. The problem seems to be that in using the outgroup, we are in effect rooting the Aslian tree on a hypothetical proto-language — which we could call 'Proto-Mon-Kammu-Khmer' — which was a sister group to Proto-Aslian. There is no reason to think that any such language existed, and it is thus nonsensical to talk about its relationship to Proto-Aslian.

The problem of aggregating different families into the outgroup led us to instead use the different outgroup families in isolation. This gives us three outgroup hypotheses, one for each of the Mon, Kammu, and Khmer outgroups. All three of these outgroup hypotheses agree in putting the root of the Aslian clade between Southern Aslian and Central/Northern (note that the midpoint rooting suggested a root between Northern Aslian and Southern/Central). Trees rooted on these outgroups are shown in Figure 6 and Figure 7. An argument could be made to prefer the trees rooted on Mon or Kammu over Khmer, since Khmer is more divergent, having had a long period as a state language, as well as being strongly influenced by Sanskrit. One might further prefer Mon over Kammu, since Mon is geographically closer to Aslian making it potentially more likely to share a more recent common ancestor. However, the major structure of these trees is congruent.

Table 3. Mean lexical distance (1 — proportion of shared cognates) for languages partitioned by Austroasiatic subgroup. The mean lexical distance within the Aslian group is 0.57.

	Aslian	Kammu	Khmer
Kammu	0.85		
Khmer	0.84	0.81	
Mon	0.84	0.79	0.76

The outgroups are all approximately equidistant from each other (measured in terms of mean proportion of unshared cognates, see Table 3), and none of them are especially closer to Aslian. But in any case, the tree topology inferred in any of these trees should be preferred over the midpoint rooted tree in Figure 5.

The roots of the trees in Figures 6 and 7 occur between Southern Aslian and the Central/Northern clades. The midpoint of the tree (shown in Figure 5) occurs between Southern/Central Aslian and Northern. We can conclude from this that the rates of diversification in the family have been very unequal. The rooted trees show that Northern Aslian is much more divergent than the other two major clades, and that Southern Aslian is the most conservative.

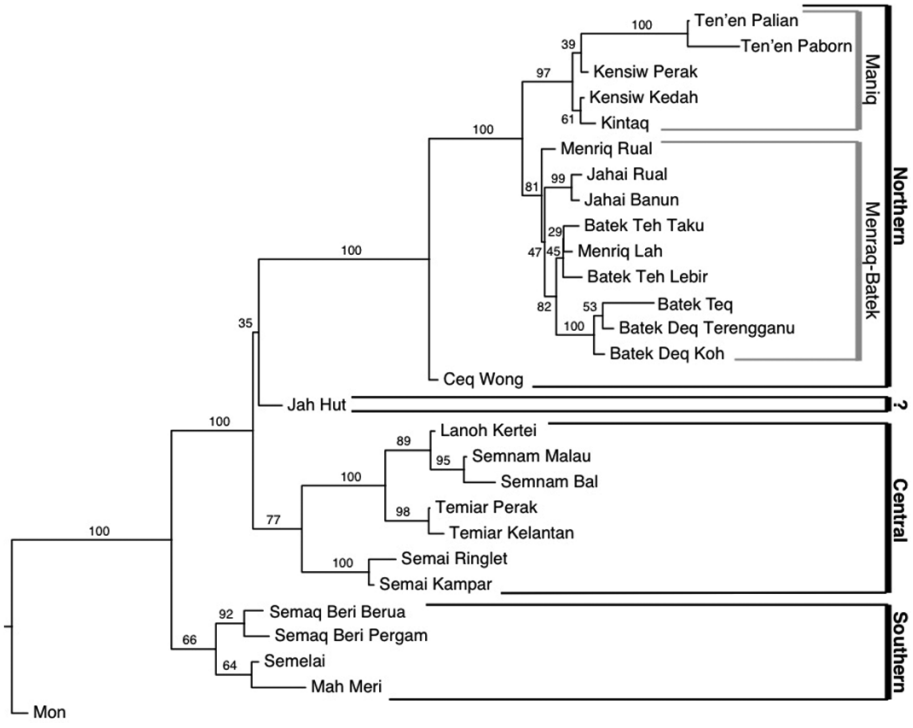

Figure 6. Aslian tree rooted on Mon (Mon-Khmer, Monic).

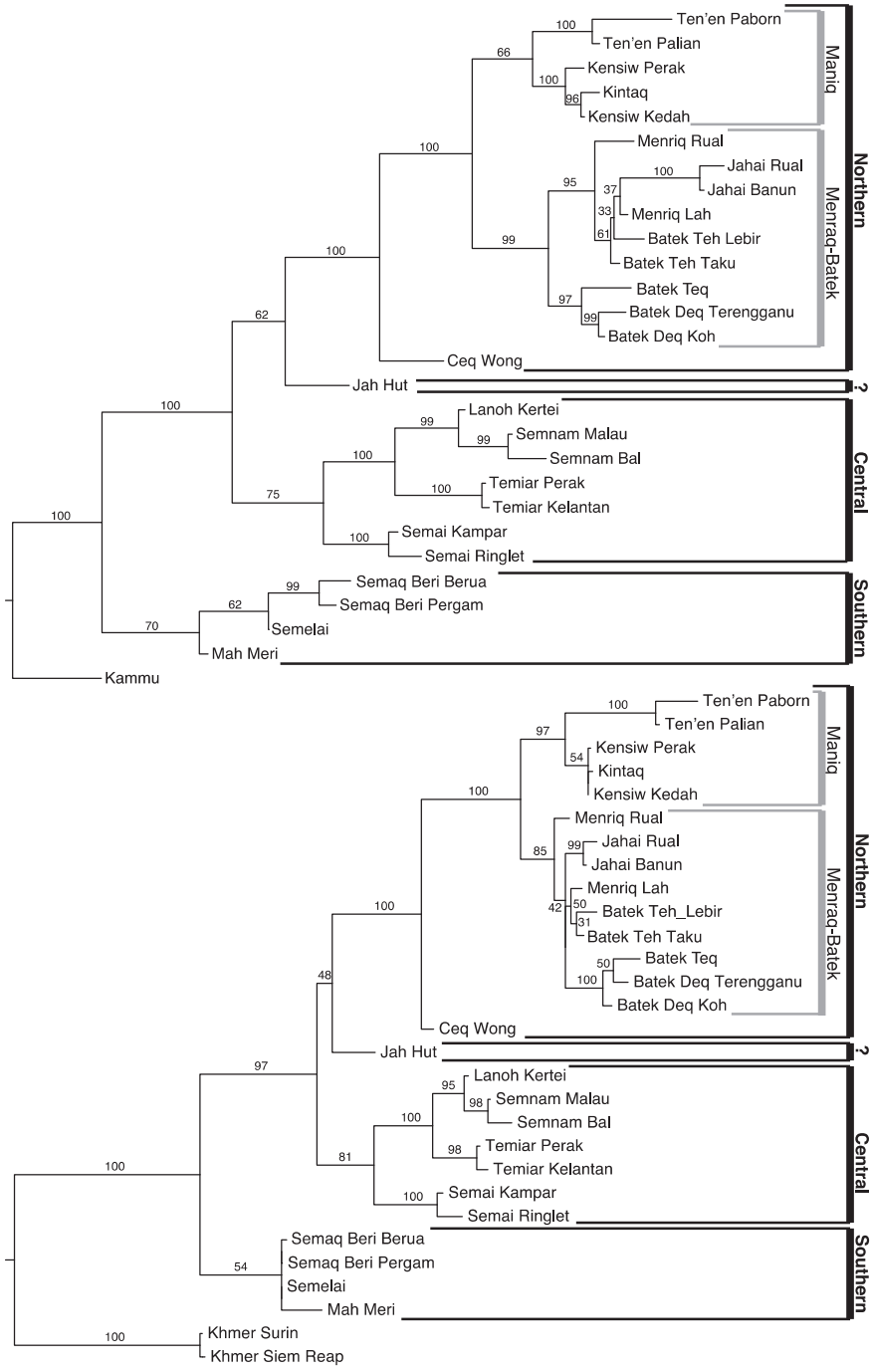

Figure 7. Aslian tree rooted on Kammu (Northern Mon-Khmer, Khmuic), and on Khmer (Eastern Mon Khmer).

4. Discussion and Conclusion

4.1 Previous analyses

The major genealogical analyses carried out on the Aslian language family to date are the lexicostatistical study of Benjamin (1976), and Diffloth's comparative studies (1975, 1979, 2005). The distance-based methods of phylogenetic estimation discussed in §3.1 are essentially elaborations upon Benjamin's methodology, and broadly support the same conclusions. Our reanalysis of Benjamin's own data illustrates the advantages of using tree and network visualization techniques for the interpretation of a lexicostatistical matrix — in this case showing that there were aspects of the data which didn't fit with the subgrouping hypothesis that he was proposing (the position of Jah Hut).

Our results also correspond broadly to those of Diffloth's 1975 comparative method study, but two aspects of his Aslian family tree are worth noting:

1. Diffloth analyses Jah Hut as a Central Aslian (or Senoic, in his terminology) language, and
2. Diffloth's analysis of what we call the Maniq/Menraq-Batek branch of Northern Aslian involves a primary split of Batek vs. the other languages (the latter forming a subgroup which includes Kensiw, Kintaq, Menriq, Jahai, and Ten'en — or Mos or Tonga in his terminology).

4.2 Benjamin's outlier languages

Three of the languages in our sample were considered to be geographical or linguistic outliers by Benjamin (1976).

There is some question as to whether **Jah Hut** is really a Central Aslian language, or whether it forms a fourth subgroup of its own. From the results of the Bayesian phylogenetic inference we would tend to favor the hypothesis that Jah Hut forms an independent fourth subgroup. In the midpoint-rooted tree (Figure 5) Jah Hut is a sister clade to Central and Southern Aslian (the branch putting Jah Hut as an earlier branching to Central and Southern has very low probability, so its exact position is uncertain). The trees rooted by outgroups (Figure 7) put Jah Hut on a similarly low probability branch between Central and Northern Aslian. The outgroup-rooted trees give a more realistic estimate of the position of Jah Hut, leading us to conclude that Jah Hut is most likely a sister clade to Central and Northern Aslian; we can have more confidence in saying that Jah Hut, Central, and Northern Aslian are distinguished on roughly the same phylogenetic level, and remain agnostic about the finer details of subgrouping.

Ceq Wong is a cultural and geographic outlier; it is the only Northern Aslian language not spoken by Semang peoples, and it is also the only Northern Aslian language separated from the main body of Northern Aslian by Aslian languages of another subgroup. As already proposed (Diffloth 1975, Benjamin 1976), Ceq Wong is a first level split of the Northern Aslian clade. Our results confirm this. Ceq Wong is notable for the great distance between it and the rest of the Northern Aslian clade. The position of Ceq Wong at the very root of the Northern Aslian clade is compatible with Burenhult's reappraisal of Northern Aslian prehistory that the forager-Northern Aslian speakers (our Maniq/Menraq-Batek subgroup) acquired their Northern Aslian languages secondarily in a chain language shift. An alternate scenario, in which Aslian languages were acquired by the foraging peoples prior to a sedentization (or, more accurately, departure from the Semang mobile foraging societal mode) of the Ceq Wong ancestors, is harder to justify. It would have to be just coincidence that the language of the sedentary group is not nested within the rest of the Northern Aslian clade. Mah Meri (see below), in Southern Aslian, shows the kind of pattern which would be expected: a language nested within the family, but highly divergent.

Benjamin (1976: 51) cites **Mah Meri** (Kruspe 2010) as another of the outliers. Our analysis does not suggest anything analogous to the situation of Jah Hut or Ceq Wong for Mah Meri. Rather, Mah Meri may be distinctive for the amount of diversification (indicated by branch length in Figure 5) undergone by the terminal node. While Mah Meri is very different from its nearest sister languages, it is nevertheless neatly nested within the clade. The explanation for this is probably geographical — for the other Aslian languages, contact with sister languages inhibits drift. Mah Meri is geographically isolated, and so does not have the same constraints inhibiting drift, and at the same time has a higher degree of contact with Austronesian.

4.3 Aslian languages of Thailand

To date there is no published classification of the Ten'en language,[5] an Aslian language spoken solely in Thailand. While word lists show it clearly to be a Northern Aslian language (Bauer 1991, Bishop & Peterson 2003, Phaiboon 2006), the precise genealogical affiliation of Ten'en is uncertain. It has even been suggested that

5. Ethnonyms of the group of Northern Aslian dialects spoken in the area where the Thai provinces Trang, Pattalung and Satun meet remain a confusing issue. Early authors referred to them as Tonga and/or Mos (see Schebesta 1952, Benjamin 1976), and Bauer (1991) calls them Trang Kensiw. The present paper uses the ethnonym Ten'en on the basis of Phaiboon (2006) and on information from Kensiw consultants in Malaysia (interviewed in 2004).

Ten'en is a variety of Kensiw/Kintaq (Bauer 1991). Another factor of interest is that since Ten'en is geographically isolated from the other Aslian languages, it is presumably subject to different contact influences.

Our analysis of Ten'en data (two varieties) confirms the close relationship between Kensiw/Kintaq and Ten'en. Ten'en has however undergone a much greater degree of diversification, as shown by the greater branch length of the Ten'en Palian/Ten'en Paborn clade. Branch length in a Bayesian phylogenetic tree only shows amount of diversification, which conflates rate of change and time of separation (see Appendix A.3). Since we know that Ten'en has a recent common ancestor with many much less diverse sister languages, we know that time of separation cannot be the major factor: the major factor producing the long branch lengths must be that the rate of language change in Ten'en is much faster than the rate in the other closely related Northern Aslian languages.[6] We can speculate as to the reasons for the greater rate of diversification of Ten'en. There may be factors which have accelerated the rate of change of Ten'en, such as the effect of isolation and contact with non-Aslian languages. Conversely, the difference in rate of change might be a function of intra-subgroup contact between the rest of the Northern Aslian forager languages (MMB subgroup) — that is, that the rate of linguistic change has been retarded for the rest of the Northern Aslian languages, while Ten'en is unconstrained. The very low posterior support for some of the branches of the Menraq-Batek languages suggests a high degree of interaction between them. These two classes of explanation are compatible with each other: the rate of change of the Ten'en branch might be accelerated at the same time that the rate of change of the other branches is slowed.

4.4 Subgrouping within Northern Aslian

A number of features of the trees in Figures 5–7 are noteworthy with respect to the subgrouping of the languages within Northern Aslian.

It has been assumed for a while that there is a split between Batek/Jahai/Menriq on the one hand, and Kensiw/Kintaq/Ten'en on the other. Following Burenhult (forthcoming), we refer to these subgroups of Northern Aslian as Menraq-Batek and Maniq respectively.[7] The trees rooted on the midpoint and on outgroups all reconstruct this split with a reasonable degree of certainty.

6. That is, this includes all the Northern Aslian languages except Ceq Wong; the Northern Aslian languages spoken by Semang foragers.

7. These terms are taken from the characteristic word for 'human' in languages of these groups. Together, we refer to the parent of these two subgroups — the Northern Aslian languages spoken by Semang foragers — as 'Maniq/Menraq-Batek' (MMB).

Likewise, it has been claimed (Benjamin 1976: 47) that the Batek Teh varieties (spoken on the lower Lebir, Kelantan) are closer to Menriq than to Batek proper. While the posterior probabilities are low for the specific ordering of the subbranches, we can state confidently that the Batek Teh varieties are subclassified distinctly from the dialects of Batek proper. Also, the previously unclassified Batek Teq is apparently distinct from Batek Deq but clearly subgrouped together with it.

4.5 Concluding remarks

Statistical, computational and algorithmic work on evolutionary trees is barely 40 years old, and the application of these methods outside biology (e.g. in language and culture) is only in its early stages. These new methods are a useful addition to the scientific toolkit of the historical linguist, allowing modelling of realistic assumptions about the processes of language change, and producing results with a number of statistically interesting/highly informative properties such as quantified uncertainty and relative chronology. The present study has recapitulated the results of traditional historical linguistic approaches, while adding substantially to the information which can be extracted from the linguistic data.

Acknowledgements

The authors would like to thank the agencies which supported this work: Michael Dunn worked within the NWO Program Grant *Structural Traces of the Sahul Past*; Niclas Burenhult and Neele Becker were part of the *Tongues of the Semang* project, funded by the Volkswagen Foundation DoBeS scheme; Sylvia Tufvesson is supported by the Max Planck Institute for Psycholinguistics, Nijmegen; Nicole Kruspe is supported by the DoBeS program, in the project *Hunter-Gatherer Languages in Contact*. Additional support is gratefully acknowledged from the Max Planck Institute for Psycholinguistics (Dunn, Burenhult), a European Community Marie Curie Fellowship (Burenhult), the Swedish Research Council (Burenhult), and the Max Planck Institute for Evolutionary Anthropology (Kruspe), Research Centre for Linguistic Typology, LaTrobe University (Kruspe), and the Hans Rausing Endangered Languages Program (Kruspe). Thanks to Dee Baer, Geoffrey Benjamin, David Bulbeck, Paul Sidwell and Angela Terrill for discussion of the content of this paper and to Stephen Levinson, director of the *Language and Cognition* group at the Max Planck Institute for Psycholinguistics, Nijmegen, for providing the stimulating interdisciplinary research environment that enabled the authors to carry out this collaboration. This paper benefited from discussion at the workshop *Dynamics of Human Diversity in Mainland Southeast Asia* (January 2009), funded by the Wenner-Gren Foundation and hosted by the École Française d'Extrême Orient in Siem Reap, Cambodia; thanks to Roger Blench, Gérard Diffloth, Nick Enfield, Alan Fix, David Gil, Siân Halcrow, Leif Jonsson, Stephen Oppenheimer, Laurent Sagart and Joyce White for their comments. Gérard Diffloth's assistance transcribing the Khmer data is appreciated, as are his critical comments on our methods. We are grateful to the Prime Minister's Department in Putrajaya and the Department of Aboriginal Affairs

(JHEOA) in Kuala Lumpur for granting permission to carry out the research, and to the Aslian communities whose members have generously shared their time and knowledge with us, as well as the individual expatriate speakers who helped us with the outgroups: Mr. Min Soe Naing (Mon), Damrong Tayanin (Kammu), Samruan Wongjaroen (Surin Khmer) and Mr. Loung (Siem Reap Khmer).

5. Appendix: Trees and networks

It is not the intention of the paper to provide a self-standing methodological introduction. However many of the techniques used are unfamiliar in traditional historical linguistic approaches. Below we provide a sketch of the main techniques we use, with references to introductory materials.

5.1 Neighbor Joining

A Neighbor Joining tree is produced by recursively clustering nearest pairs of nodes from an initial star phylogeny. Each pair of nodes is redrawn as a new, combined node with a paired branch emerging from it, as illustrated in Figure 8.[8]

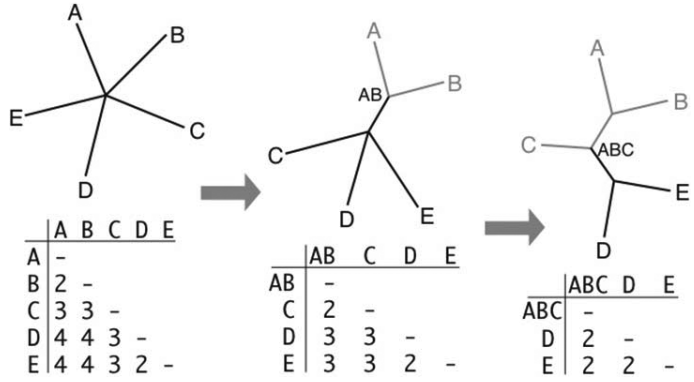

Figure 8. Neighbor Joining algorithm: start from a star phylogeny (left); find the nearest pair of nodes (according to the distance matrix, either of A-B or D-E) (middle); recalculate the distance matrix using the new node (AB); repeat until the tree is fully resolved (right).

8. The distance of the combined node from the root of the star (R) is calculated by averaging the distance from the root to each of the old nodes minus the distance between them, thus in the first panel of Figure 8 this is (DRA+DRB-DAB)/2. The distance from the new node to each of the old nodes is calculated on the basis of the distance between the old nodes and the difference between the distance of each of the old nodes to the root, i.e. for node A it is (DAB+DRA–DRB)/2; for node B it is (DAB+DRB–DRA)/2.

Consider the hypothetical data shown in Figure 9: Language A has a distance of 0.5 (i.e. 50% different) from languages B and D, but languages B and D have a distance of 1.0 from each other. Language C has a distance of 1.0 from A, but like A, it also has a distance of 0.5 from each of B and D. The left panel of Figure 9 is a tree representation of these distances. This tree cannot capture some important relationships in the data: the pairs A-C and B-D should be maximally distinct, as indeed they are, but the other pairs (A-B, B-C, C-D and D-A) should all show a closer relationship. Instead of this, the pairs B-C and D-A are represented as being just as distant as A-C and B-D. Only a small subset of possible matrices of pairwise distances can be represented precisely by a tree; in most cases the relationships between the taxa are somewhat distorted.

Gascuel & Steel (2006) gives the state-of-the-art on neighbor joining trees.

Lang	Cog1	Cog2	Cog3	Cog4
A	0	0	0	0
B	0	0	1	1
C	1	1	1	1
D	1	1	0	0

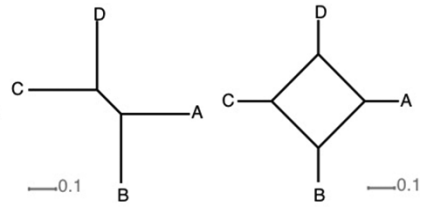

Figure 9. Matrix of hypothetical cognate coding for four languages: A, B, C and D (left), represented as a distance tree (middle) and a splits-graph (right).

5.2 NeighborNet

There are better methods for representing Complex distances for the cases such as the hypothetical example above. The **NeighborNet** method can capture conflicting evidence for clustering. Bryant et al. (2005) have published a good introduction to how NeighborNet works, and how to interpret it, aimed at a linguistic audience. In brief, however, the NeighborNet algorithm works like the Neighbor Joining algorithm, except that instead of collapsing the two closest nodes, it finds the *three* closest nodes, and delays collapsing them until the next cycle of distance calculations so as to preserve (some of) the conflicting evidence in the data.

In a NeighborNet 'splits graph', sets of parallel lines indicate bifurcations in the same way as a single line splits the taxa in a tree into two. Figure 10 shows a set of conflicting bifurcations of a tree, listed under splits. Splits with higher weight (equivalent to branch length on a tree) are longer. In Figure 10, there is strong evidence for clustering with an (A, B) and a (D, E) clade, weaker evidence for clustering the (A, C) and (C, E) clades, and no evidence for a (B, D) clade.

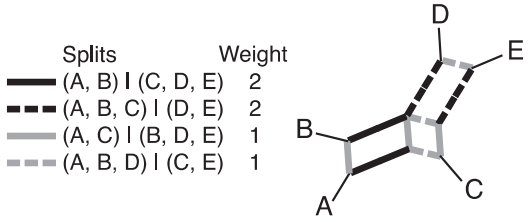

Figure 10. Splits graph, as used by NeighborNet.

5.3 Bayesian phylogenetic inference

Bayesian phylogenetic inference is a model-based method of inferring history. In the model used in this paper we assume that a language can be represented by a matrix encoding the presence of reflexes of a cognate set, and that the historical behaviour of these linguistic features can be characterized by a mathematical formula expressing transition probabilities — in our case, the probabilities of reflexes of a cognate candidate set being gained and lost. In addition to the model we have a set of observations of the feature values (cognate set membership) for a set of taxa (languages). Bayesian phylogenetic inference is a method of finding the most probable set of evolutionary parameters to have produced this observed variation. The evolutionary parameters here include both tree topologies, and transition probabilities. Finding the most likely tree according to a set of observations and a model is an extremely difficult computational task, and one which is intractable for all but the smallest data sets. Bayesian phylogenetic inference estimates the solution to this problem by searching a 'space' of all possible combinations of parameter values (tree topologies and transition probabilities, and randomly sampling the parameters from the region of highest probability. The result of a Bayesian phylognetic analysis is a random sample of the most probable trees. The proportion of the tree set containing a particular node gives us a measure of confidence that we can have in that node.

There are many possible models of evolution that can be used in a Bayesian phylogenetic analysis. The fit of a model and set of parameters to a set of data can be measured. This is called a likelihood score. The likelihood scores of a set of equiprobable trees are — as you would expect — roughly equal. However, the likelihood for the same data using a different model stabilize around different harmonic mean. There are statistical techniques to help decide whether the fit of one model is significantly better than the fit of another. Ideally one wishes to find the simplest possible model that produces the highest likelihood results. The tradeoff between small increases in likelihood and great increases in model complexity can lead to a situation of 'over-fitting', where a model is so tightly tuned to a particular set of data that it loses predictive power.

For this study we tested all combinations of a number of different model parameters: uni- and bi-directional models, where probability of gain is or is not equal to probability of loss of a feature; gamma distribution models, where rates of change are sorted into rate-classes; covarion models, where change of rate is itself a parameter to be estimated over the tree. In common with our experience and the experience of our colleagues, the best performing model (highest likelihood) for lexical cognate data was a covarion model. The performance of the unidirectional model was slightly inferior to that of the bidirectional model, but the difference was not enough for us to prefer the more complex (i.e. bidirectional) model over the simpler one. The covarion method of dealing with rates changes performed better than the gamma model; using both models in combination did not provide any additional advantage.

Dunn (2008) gives a non-mathematical description of Bayesian phylogenetic inference — including model selection and searching the parameter space — with a linguistic slant. Huelsenbeck et al. (2001) is a good introductory technical overview from a more general perspective.

References

Bauer, Christian. 1991. "Kensiw: A Northern Aslian language of Southern Thailand". *Preliminary Report of Excavations at Moh-Khiew Cave, Krabi Province, Sakai Cave, Trang Province and Ethnoarchaeological Research of Hunter-gatherer Group, socall* [sic] *'Sakai' or 'Semang' at Trang Province* ed. Surin Pookajorn, 310–335. Bangkok: Silpakorn University.

Ben Hamed, Mahé. 2005. "Neighbour-nets portray the Chinese dialect continuum and the linguistic legacy of China's demic history". *Proceedings of the Royal Society B — Biological Sciences* 272.1015–1022.

Ben Hamed, Mahé & Feng Wang. 2006. "Stuck in the forest: Trees, networks and Chinese dialects". *Diachronica* 23:1.29–60.

Benjamin, Geoffrey. 1976. "Austroasiatic subgroupings and prehistory in the Malay Peninsula". *Austroasiatic Studies, Part I* ed. by Philip N. Jenner, Laurence C. Thompson & Stanley Starosta, 37–128. Honolulu: The University of Hawai'i Press.

Benjamin, Geoffrey. 1985. "In the long term: Three themes in Malayan cultural ecology". *Cultural Values and Human Ecology in Southeast Asia* ed. by Karl L. A. Hutterer, Terry Rambo & George Lovelace, 219–278. University of Michigan: Center for South and Southeast Asian Studies.

Benjamin, Geoffrey. 1997. "Issues in the ethnohistory of Pahang". *Pembangunan Arkeologi Pelancongan Negeri Pahang* ed. by Nik Hassan Shuhaimi bin Nik Abdul Rahman, Mokhtar Abu Bakar, Ahmad Hakimi Khairuddin & Jazamuddin Baharuddin, 82–121. Pekan: Muzium Pahang.

Bishop, Nancy M. & Mary M. Peterson. 2003. *Northern Aslian Language Survey: Trang, Satul and Phatthalung Provinces, Thailand*. Bangkok: Thammasat University.

Bryant, David, Flavia Filimon & Russell Gray. 2005. "Untangling our past: Languages, trees, splits and networks". *The Evolution of Cultural Diversity: Phylogenetic approaches* ed. by Ruth Mace, Clare Holden & Stephen Shennan, 67–84. London: UCL Press.

Bryant, David & Vincent Moulton. 2003. "Neighbor-Net: An agglomerative method for the construction of phylogenetic networks". *Molecular Biology and Evolution* 21:2.255–265.

Bulbeck, David F. 2004. "An integrated perspective on Orang Asli ethnogenesis". *Southeast Asian Archaeology: Wilhelm G. Solheim II Festschrift* ed. by Victor Paz, 366–399. Quezon City: The University of the Philippines Press.

Burenhult, Niclas. 2005. *A Grammar of Jahai*. Canberra: Pacific Linguistics.

Burenhult, Niclas. 2006. "Tongues of the Semang: Annual interim report I". Nijmegen: Max Planck Institute for Psycholinguistics.

Burenhult, Niclas. Forthcoming. "Foraging and the history of languages in the Malay Peninsula". *Historical Linguistics and Hunter-gatherer Populations in Global Perspective* ed. by Tom Güldeman, Patrick McConvell & Richard Rhodes.

Burenhult, Niclas, Nicole Kruspe & Michael Dunn. Forthcoming. "Language history and culture group among Austroasiatic-speaking foragers of the Malay Peninsula". *Dynamics of Human Diversity in Mainland Southeast Asia* ed. by Nick Enfield & Joyce White. Canberra: Pacific Linguistics.

Collins, James T., ed. 2006. *Borneo and the Homeland of the Malays*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Darwin, Charles. 1996 [1859]. *On the Origin of Species by Natural Selection*. Oxford: Oxford University Press.

Diffloth, Gérard. 1975. "Les langues mon-khmer de Malaisie: Classification historique et innovations". *Asie du Sud-Est et Monde Insulinde* 6:4.1–19.

Diffloth, Gérard. 1976. "Jah-Hut: An Austroasiatic language of Malaysia". *South-East Asian Linguistic Studies,* Vol. 2 ed. by Nguyen Dang Liem, 73–118. Canberra: Pacific Linguistics.

Diffloth, Gérard. 1979. "Aslian languages and Southeast Asian prehistory". *Federation Museums Journal* 24.3–16.

Diffloth, Gérard. 2005. "The contribution of linguistic palaeontology to the homeland of Austro-Asiatic". *The Peopling of East Asia: Putting together archaeology, linguistics and genetics* ed. by Roger Blench, Laurent Sagart, & Alicia Sanchez-Mazas, 79–82. London & New York: Routledge Curzon.

Diffloth, Gérard. 2008. "Review of: A Mon-Khmer comparative dictionary, by Harry Shorto, ed. by Paul Sidwell". *Diachronica* 25:1.137–142.

Diffloth, Gérard & Norman Zide. 1992. "Austro-Asiatic languages". *International Encyclopedia of Linguistics,* Vol. 1 ed. by William Bright, 137–142. New York & Oxford: Oxford University Press.

Dunn, Michael. 2008. "Contact and phylogeny in island Melanesia". *Lingua* 119:1664–1678.

Dunn, Michael, Stephen Levinson, Eva Lindström, Ger Reesink & Angela Terrill. 2008. "Structural phylogeny in historical linguistics: Methodological explorations applied in island Melanesia". *Language* 84:4.710–759.

Endicott, Kirk. 1979. *Batek Negrito Religion: The world-view and rituals of a hunting and gathering people of peninsular Malaysia*. Oxford: Clarendon Press.

Fix, Alan G. 1995. "Malayan paleosociology: Implications for patterns of genetic variation among the Orang Asli". *American Anthropologist* 97:2.313–323.

Gascuel, Olivier & Mike Steel. 2006. "Neighbor-Joining revealed". *Molecular Biology and Evolution* 23:11.1997–2000.

Gray, Russell D. & Quentin D. Atkinson. 2003. "Language-tree divergence times support the Anatolian theory of Indo-European origin". *Nature* 426.435–439.

Greenhill, Simon J. & Russell D. Gray. 2005. "Testing population dispersal hypotheses: Pacific settlement, phylogenetic trees and Austronesian languages". *The Evolution of Cultural Diversity: A phylogenetic approach* ed. by Ruth Mace, Clare J. Holden & Stephen Shennan, 31–52. London: UCL Press.

Hill, Clare, Pedro Soares, Maru Mormina, Vincent Macaulay, Dougie Clarke, Petya B. Blumbach, Matthieu Vizuete-Forster, Peter Forster, David Bulbeck, Stephen Oppenheimer & Martin Richards. 2007. "A mitochondrial stratigraphy for Island Southeast Asia". *The American Journal of Human Genetics* 80:1.29–43.

Hill, Catherine, Pedro Soares, Maru Mormina, Vincent Macaulay, William Meehan, James Blackburn, Douglas Clarke, Joseph Maripa Raja, Patimah Ismail, David Bulbeck, Stephen Oppenheimer & Martin Richards. 2006. "Phylogeography and ethnogenesis of Aboriginal Southeast Asians". *Molecular Biology and Evolution* 23:12.2480–2491.

Huelsenbeck, John P., Fredrik Ronquist, Rasmus Nielsen & Jonathan P. Bollback. 2001. "Bayesian inference of phylogeny and its impact on evolutionary biology". *Science* 294.2310–2314.

Kruspe, Nicole. 2004. *A Grammar of Semelai*. Cambridge: Cambridge University Press.

Kruspe, Nicole. 2010. *A Dictionary of Mah Meri, as Spoken at Bukit Bangkong. (=Oceanic Linguistics Special Publications.)* Honolulu: University of Hawai'i Press.

Kruspe, Nicole. Forthcoming. "Hunter-gatherer languages in contact: Semaq Beri of Terengganu".

McMahon, April & Robert McMahon. 2003. "Finding families: Quantitative methods in language classification". *Transactions of the Philological Society* 101:1.7–55.

Milinkovitch, Michel C. & James Lyons-Weiler. 1998. "Finding optimal ingroup topologies and convexities when the choice of outgroups is not obvious". *Molecular Phylogenetics and Evolution* 9:3.348–357.

Phaiboon, Duangchan. 2006. "Glossary of Aslian languages: The Northern Aslian languages of Southern Thailand". *Mon-Khmer Studies* 36.207–224.

Rambo, Terry A. 1988. "Why are the Semang? Ecology and ethnogenesis of aboriginal groups in Peninsular Malaysia". *Ethnic Diversity and the Control of Natural Resources in Southeast Asia* ed. by Terry A. Rambo, Kathleen Gillogly & Karl L. Hutterer, 19–35. Ann Arbor: University of Michigan Press.

Saitou, Naruya & Masatoshi Nei. 1987. "The neighbor-joining method: A new method for reconstructing phylogenetic trees". *Molecular Biology and Evolution* 4:4.406–425.

Sanderson, Michael J. & H. Bradley Shaffer. 2002. "Troubleshooting molecular phylogenetic analyses". *Annual Review of Ecological Systems* 33.49–72.

Schebesta, Paul. 1952. *Die Negrito Asiens: Geschichte, Geographie, Umwelt, Demographie und Anthropologie der Negrito*. Vienna: St. Gabriel Verlag.

Shorto, Harry, Paul Sidwell, Doug Cooper & Christian Bauer. 2006. *A Mon-Khmer Comparative Dictionary*. Canberra: Pacific Linguistics.

Thomas, David D. 1960. "Basic vocabulary in some Mon-Khmer languages". *Anthropological Linguistics* 2:3.7– 11.

Wang, William S.-Y. 1996. "Linguistic diversity and language relationships". *New horizons in Chinese Linguistics* ed. by James Huang & Audrey Li, 235–267. Dordrecht: Kluwer Academic Publishers.

Zusammenfassung

Dieser Aufsatz befaßt sich mit neu gesammelten lexikalischen Daten aus 26 Sprachen der Asli-an-Untergruppe der austro-asiatischen Sprachfamilie. Die Daten wurden mit Hilfe von computergestützten phylogenetischen Methoden ausgewertet. Wir präsentieren die wahrscheinlichste Topologie des Stammbaums der aslianischen Familie und diskutieren Verwurzelungsmodelle für diese Familie sowie ihre möglichen externen Beziehungen zu anderen austro-asiatischen Sprachen. Belege werden präsentiert, die die Klassifikation des Jah Hut als einen vierten übergeordneten Sprachzweig der Familie unterstützen. Die phylogenetische Position bekannter geographischer und linguistischer "outlier"-Sprachen werden verdeutlicht und das Verhältnis der bisher nur wenig erforschten Sprachen Südthailands zu den übrigen Sprachen der Familie wird untersucht.

Résumé

Cet article se sert de méthodes de calcul phylogénétique afin d'analyser des données lexicales tirées récemment de 26 langues du sous-groupe aslien de la famille linguistique austroasiatique. Nous montrons ce qu'est la topologie la plus vraisemblable de l'arbre généalogique du groupe aslien, calculons ses embranchements et évaluons ses relations avec d'autres langues austroasiatiques ; puis nous comparons les rythmes de diversification entre ses différentes branches. Nous soutenons que la langue jah hut constituerait un quatrième membre au niveau supérieur de l'arbre. Nous examinons les positions phylogénétiques de certaines langues dont on sait qu'elles sont isolées géographiquement et linguistiquement (outliers). Enfin, nous explorons les relations des langues asliennes de la Thaïlande méridionale, mal connues, avec les autres langues constituant l'autroasiatique.

Authors' addresses

Michael Dunn
Radboud University Nijmegen
Centre for Language Studies
P.O. Box 9103
6500 HD Nijmegen, The Netherlands

Max Planck Institute for Psycholinguistics
P.O. Box 310
6500 AH Nijmegen, The Netherlands

Michael.Dunn@mpi.nl

Niclas Burenhult
Max Planck Institute for Psycholinguistics
P.O. Box 310
6500 AH Nijmegen, The Netherlands

Lund University
Centre for Languages and Literature
Box 201
221 00 Lund, HT: 20, Sweden

Niclas.Burenhult@mpi.nl

Nicole Kruspe
University of Melbourne
School of Languages and Linguistics
Babel Building
Parkville, Victoria 3010, Australia

nkruspe@unimelb.edu.au

Sylvia Tufvesson
Max Planck Institute for Psycholinguistics
P.O. Box 310
6500 AH Nijmegen, The Netherlands

Sylvia.Tufvesson@mpi.nl

Neele Becker
Max Planck Institute for Psycholinguistics
P.O. Box 310
6500 AH Nijmegen, The Netherlands

Johannes Gutenberg University Mainz
Department of English and Linguistics
Jakob-Welder-Weg 18
55128 Mainz, Germany

Neele.Becker@uni-mainz.de