Systematic Bayesian phylogenetics reveal dialect chain breakup in the Timor-Alor-Pantar language family

Gereon A. Kaiping^{1,2} and Marian Klammer¹

¹Leiden University Centre for Linguistics, Universiteit Leiden, NL

²Geographic Information Science Center, Universität Zürich, CH

November 25, 2021

This paper refines the subgroupings of the Timor-Alor-Pantar (TAP) family of Papuan languages, using a systematic Bayesian phylogenetics study. While recent work indicates that the TAP family comprises a Timor (T) branch and an Alor Pantar (AP) branch (Holton et al. 2012, Schapper et al. 2017), the internal structure of the AP branch has proven to be challenging. Our Bayesian inference study is based on extensive TAP lexical data from the online LexiRumah database (Kaiping et al. 2019b, Kaiping & Klammer 2018). Systematically comparing different analytical models and tying them back to the historical linguistics, we arrive at a subgrouping structure of the TAP family based on features of the phylogenies shared across the different analyses. Our TAP tree differs from all earlier proposals by inferring the East-Alor subgroup as an early split-off from all other AP languages, instead of the most deeply embedded subgroup inside that branch. The evidences suggests that dialect cluster effects played a major role in the formation of today's Timor-Alor-Pantar languages.

1. Introduction

The Timor-Alor-Pantar (TAP) languages are a family of some 25 non-Austronesian or “Papuan” languages spoken on the islands of Timor, Alor and Pantar, as well as on neighbouring smaller islets, located some 1 000 kilometres west of New Guinea (Holton et al. 2012, Schapper et al. 2017, Holton & Robinson 2017, Klammer 2017). The TAP family is split into three branches: Bunak, the East Timor group comprising the languages Makasae, Fataluku, and Oirata, and the Alor-Pantar (AP) branch, which comprises the remaining languages. For an overview over the TAP languages and their locations, see Fig. 1 and Table 1.

Bunak, Makasae and Fataluku are spoken on the island of Timor, the language Oirata is spoken on the neighboring island of Kisar, the languages Hamap, Adang, Kabola, Klon, Kafoa, Kui, Kiraman, Abui, Kamang, Kula, Sawila, Wersing, and one dialect of Reta are spoken on Alor; Kaera, Klamu, Deing, Teiwa, Western Pantar, Sar and most Blagar varieties are spo-

Figure 1: Map of the Timor-Alor-Pantar languages

ken on the island of Pantar, while Blagar-Pura and Reta are spoken on the island of Pura in the Straits between Alor and Pantar.

Most of the TAP languages have speaker communities ranging between 1 000 and 10 000 speakers, see Table 1. For the entire TAP family, the average number of speakers per language is 12 023 due to the three big languages Bunak (80 000 speakers), Fataluku (30 000), and Makasae (70 000), spoken on Timor. For the AP branch, the average number of speakers per language is much lower (4 510). On the low end of the scale, Klamu has only 200 speakers (Holton 2004), while Sar had only one active speaker left in 2018 (Klamer & Sir 2018) who deceased in 2019.

The relatively small size of most of these speaker communities and the limited geographical area in which they live has resulted in a situation where traditionally neighbouring groups had much contact through (barter) trade, cultural exchanges and intermarriage. Contact with one or more Austronesian donor language(s) goes back a long way, as evidenced by the Austronesian loans that were borrowed into the AP proto-language before it split up, perhaps 3 000 years ago (see Section 2.5; Holton et al. (2012)). The ancient contact between TAP and Austronesian languages continued through pre-modern times (Klamer to appear) and has resulted in Austronesian loan words in the individual TAP languages.

Three modern donor languages of Austronesian loan words are Indonesian, the dominant national language of Indonesia and/or its basilect Alor Malay; Alorese, the lingua franca used on (parts of) Pantar and West Alor before the advent of Malay/Indonesian; and Tetun, one of the national languages of Timor Leste.

Over the last 15 years, a body of work on TAP languages has appeared, including grammatical descriptions, typological comparisons and historical comparative work. Historical

reconstructions used the traditional comparative method based on regular sound changes, first demonstrating the existence of an Alor-Pantar group (Holton et al. 2012) and an East Timor group (Schapper et al. 2012), followed by a demonstration that the groups together form the TAP family (Schapper et al. 2017).

If and how the TAP family relates to other families is examined by Holton & Robinson (2014) who conclude that there is currently no lexical evidence to support an affiliation with any other family in the world, including the Trans New Guinea family, in contrast to what had been assumed previously (Wurm et al. 1975, Ross 2005).

A tree reflecting the internal structure of the AP branch was proposed in Holton et al. (2012), see Fig. 2. Robinson & Holton (2012a) compared this tree with a Bayesian phylogenetic inference tree, see Fig. 3. Both trees were created using data collected up to 2009. Since then, the sample of TAP language varieties has increased from 12 to nearly 50. The new word lists have been combined with the earlier body of data and made available in the LexiRumah database (Kaiping et al. 2019b, Kaiping & Klamer 2018). On the methodological side, recent years have seen a significant improvement of computational tools, practices, models and algorithms for inferring language histories from word lists. For example, there are now new methods for automatic cognate detection using state-of-the-art pairwise phonetic alignment algorithms (List 2012a,b, List & Greenhill & Tresoldi, et al. 2018, List et al. 2017, Jäger et al. 2017) that reach nearly 90% accuracy (B-cubed F-score) in determining the cognate sets to be used in phylogenetic analyses (Rama et al. 2018). Bayesian phylogenetic inference research has empirically shown which models are useful for lexical data (Chang et al. 2015, Kolipakam et al. 2018). In order to apply phylogenetics to linguistic data, tools such as BEASTling (Mau-

Language	ISO	Population	Dialect	Island	Abbreviation	Glottocode
Abui	abz	17000	Atimelang	Alor	Ab	atim1239
			Fuimelang	Alor		abui1241
			Papuna	Alor		abui1241
			Petleng	Alor		abui1241
			Takalelang	Alor		abui1241
			Ulaga	Alor		abui1241
Adang	adn	7000	Lawahing	Alor	Ad	adan1251
			Otvai	Alor		adan1251
Blagar	beu	10000	Bakalang	Pantar	Bl	baka1276
			Bama	Pantar		blag1240
			Kulijahi	Pantar		blag1240
			Manatang	Pantar		blag1240
			Nule	Pantar		blag1240
			Tuntuli	Pantar		blag1240
			Warsalelang	Pantar		blag1240
			Pura	Pura		pura1258
Bunak	bfm	80000	Bobonaro	Timor		buna1278
			Maliana	Timor		buna1278
			Suai	Timor		buna1278
Deing	twe	1000	Deing	Pantar		dein1238
Fataluku	ddg	30000	Fataluku	Timor		fata1247
Hamap	hmu	1300	Moru	Alor		hama1240
Kabola	klz	3900	Monbang	Alor		kabo1247
Kaera	jka	5500	Abangiwang	Pantar	Ke	kaer1234
Kafoa	kpu	1000	Probur	Alor		kafo1240
Kamang	woi	6000	Kamang-Atoitaa	Alor	Km	kama1365
			Tiyei	Alor		tiay1238
			Suboo-Apui	Alor		sibo1242
			Suboo-Atiibaai	Alor		sibo1242
Kiraman	kvd	1900	Kiraman	Alor		kira1248
Klamu	nec	*200	Klamu	Pantar	Nd	nede1245
Klon	kyo	5000	Bring	Alor	Kl	kelo1247
			Hopter	Alor		kelo1247
Kui	kvd	1900	Labaing	Alor	Ki	kuii1253
Kula	tpg	5000	Lantoka	Alor		kula1280
Makasae	mkz	70000	Makasae	Timor		maka1316
Oirata	oia	1220	Oirata	Kisar		oira1263
Reta	ret	*2500	Hula	Alor		rett1240
			Pura	Pura		rett1240
			Ternate	Ternate		rett1240
Teiwa	twe	4000	Adiabang	Pantar	Tw	teiw1235
			Nule	Pantar		teiw1235
			Lebang	Pantar		leba1239
Sawila	swt	3000		Alor	Sw	sawi1256
Sar		*1		Pantar		sarr1247
Wersing	kwv	3700	Maritaing	Alor	We	wers1238
			Taramana	Alor		wers1238
Western Pantar	lev	10300	Tubbe	Pantar	WP	lamm1241

Table 1: Languages of the TAP family with the island where they are spoken, and number of speakers. Speaker numbers are taken from Klammer (2017: Table 1), except (marked by a star) for Reta, which is taken from Willemsen (2020), Klamu from Holton (2004) and Sar from Klammer & Sir (2018).

Figure 2: Subgroupings tree based on shared phonological innovations, according to and reproduced from Holton et al. (2012)

Figure 3: Non-ultrametric tree from Bayesian phylogenetic inference with a stochastic Dollo model, as described in and reproduced from Robinson & Holton (2012a)

rits et al. 2017) or Lexedata (Kaiping & Steiger 2021) provide ways to transform datasets conforming to cross-linguistic standards (Forkel 2017, Forkel et al. 2018) into the data structures used by phylogenetic software.

The earlier reconstructions of the TAP family (Schapper et al. 2017, Holton et al. 2012, Robinson & Holton 2012a) conclusively show the relatedness of the languages. However, they leave large gaps in our understanding of the linguistic history of the TAP family, especially in its subgrouping. In this paper we re-address the subgrouping of the TAP family, with a particular focus on the internal structure of the AP subfamily, while making all the analytical steps involved as visible and explicit as possible. We reduce the subjectiveness of the subgrouping task by applying recently developed computational tools and models. Our data is the now extensive lexical dataset of TAP languages that is publicly available in LexiRumah 3.0.0 (Kaiping et al. 2019b). We focus on the lexicon as a data type that is quantitatively well-understood and for which models of evolution exist.

These computational models of lexical evolution currently rely on cognate-coded word

list data. Each language variety is represented by its forms (words) that best represent a list of pre-defined meanings or concepts. Forms are grouped in cognate classes according to common origin. An example for a list with the four given concepts ‘arm’, ‘arrow’, ‘bamboo’, and ‘to blaze; to shine’ in two language varieties can be found in Table 2.

For this article, we use state-of-the-art automatic cognate detection algorithms to approximate the incremental process of selecting cognate sets and identifying regular sound changes, see Section 2.2. We systematically exclude loan words from Malay/Indonesian and Tetun from our data set, and we compare the results of each method. The weighting of the subgrouping evidence is done through explicit, stochastic models for the evolution of the lexicon. At their core, these models assume that a meaning keeps being expressed by words derived from a particular root for some time, but that a stochastic process removes existing associations between meanings and roots and creates new associations. In order to be computationally tractable, the models make different assumptions about the independence of the stochastic processes. For example, they generally assume that the words expressing different concepts evolve independently, which is a reasonable assumption for forms with separate meanings (eg. ‘to sit’ and ‘sacrifice’), but a necessary simplification for forms that express semantically similar meanings (eg. ‘to sit’ and ‘to dwell’). We explicitly compare two such models.

All our parameter values and the implementations of our methods in unambiguous computer algorithms are provided open for inspection, available as online supplementary material and through https://osf.io/h4nxk/?view_only=8dd7b81f03304d5085191424bd709517. The robustness of all our results is investigated under several alternative models

and/or parameter choices, each with different random seeds.

2. Data and Methods

Running a Bayesian phylogenetic inference procedure on lexical data has the following requirements: (1) The lexical data must be in a comparable format (phonetic or phonemic), with clean meaning-form mappings, and (2) coded for cognacy; (3) loan words must be handled such that the phylogenetic signal is not dominated by the contact signal of the borrowings; (4) a computational model of lexical evolution must be specified; and (5) the posterior probability of the possible evolutionary histories in the model, given the cognate-coded data, must be calculated. In the following subsections we will discuss each of these steps.

The result of a Bayesian inference is a posterior probability distribution of trees often characterized by a DensiTree visualization (Bouckaert 2010, Bouckaert & Heled 2014) of trees sampled from the distribution, or by a consensus tree summarizing properties of the trees in the distribution.

For example, Fig. 4 shows the probability distribution corresponding to “80% certainty that language A and language B are more closely related to each other than to language C, and no other information” (cf. Fig. 4a), and visualizes it using DensiTree (Fig. 4b) and a maximum clade credibility summary tree (Fig. 4c). All prior knowledge and all results of Bayesian phylogenetics are in principle such quantitative certainty statements. The posterior distribution is generated as a long sequence of trees, each with frequency according to its certainty (Fig. 4d).

Concept English Gloss	Abui-Fuimelang		Blagar-Bakalang	
	C	Form	C	Form
arm	1	tatarj	1	atarj
arrow	2	kaike	3	bulit
bamboo	4	mai	5	petuj
	4	maiya		
to blaze; to shine	6	hiede	7	bolor
			6	ede

Table 2: Example word list for two varieties and four concepts, with forms coded for cognacy (columns C). Cognate forms are given the same code and the same color.

Figure 4: Various representations of the tree distribution “80% certainty that A and B are more closely related to each other than to C, and no other information”

2.1. Data

As mentioned above, the data for this study is taken from the open-access LexiRumah database (Kaiping & Klamer 2018). The most recent version v3.0.0 (Kaiping et al. 2019b) of the database contains 357 language varieties (or “lects”), 48 of which are extant Timor-Alor-Pantar languages and three comparative reconstructions of TAP proto-languages, namely proto-Timor-Alor-Pantar, proto-Alor-Pantar and proto-East Timor-Bunak.

The forms of the proto-languages are not independent of the other data points (because some of them have been used in the reconstruction procedure), so we did not include them in our analyses. The remaining 48 word lists of TAP languages have been aggregated from various recent sources and are not all equally long. Originally, their sources may have used different glossing conventions, but LexiRumah provides manually normalized Concepticon mappings (List et al. 2016) for all its concepts. In this way, meanings that are glossed differently across different sources (e.g. the forms glossed as ‘saya’ by Keraf (1978) and as ‘I’ in Blust (1999)) are linked to the same underlying concept (in this case ‘1sg; <https://concepticon.clld.org/parameters/1209>’). Each lexeme in the database is presented with a transcription as supplied by the original source, and a normalized, segmented phonetic transcription in IPA (using diacritics only in those cases where that level of detail is available in the original source).

Some words occur in sets that are highly correlated in shape. For example, in a language with a decimal numeral system, the number words above ‘ten’ will typically have forms reflecting the morpheme for ‘ten’; and when the form for ‘ten’ has been replaced in a language (by e.g. a borrowing), it is very likely that the forms for ‘twenty’, ‘thirty’, etc.,

or the forms for ‘nineteen’, ‘eighteen’, etc., also change. So we excluded the 25 decimally composite numeral concepts from the word lists in our sample. Some but not all of the TAP languages have quinary numeral systems, so that the numbers ‘six’ to ‘ten’ are multi-morphemic in some languages but not in others. Therefore we kept the numeral concepts up to ‘ten’ throughout.

After removing 25 numeral concepts, each TAP lect in our data is attested through a word list containing between 99 (Abui-Petleng variety) and 553 (Fataluku) forms spanning 582 concepts. The mean number of forms for each TAP word lists is 462.5, representing on average 416.9 different concepts, due to an average synonym count of 1.1 forms per filled concept-language slot. Apart from the very short word list of Abui-Petleng, any two word lists in our database share at least 114 concepts.

2.2. Cognate coding

In the context of this paper, we use ‘cognates’ as a shorthand to mean ‘words of shared etymological origin with the same meaning’ (with the additional caveat that our shared etymological origins are not posited by a human linguist with background knowledge in the languages, but by a computer program following heuristics to approximate systematic sound changes, to be explained below). In this definition, we do in principle include borrowed items. While the notion of cognacy is not entirely clear, see List (2016) for an in-depth discussion, this is an extension from the usual meaning of ‘cognate’ even in linguistic phylogenetics. An explanation of this choice, largely following Chang et al. (2015), is given in ??.

In order to get full control over the establishment of cognate classes and to investigate the robustness of results under different

methods, we ran different existing automatic cognate detection algorithms on our data. An natural-language description of the underlying processes and intuitions can be found in Appendix appendix B.

In most of our analyses, we use the LexStat algorithm (List 2012a), which has been shown to perform very well for the purpose of cognate coding (List et al. 2017), and is easily available in the LingPy software package for historical linguistics (List & Greenhill & Tresoldi, et al. 2018). LexStat requires at least 100 shared concepts to work well (List 2014), so we expect the Abui-Petleng wordlist, which is below this number, to not be coded particularly well. As sound class model for LexStat, we use ASJP-transcribed data (analysis D1) in addition to our Baseline, which uses SCA sound classes (List 2012b).

We follow List et al. and cluster forms into cognate classes using Infomap (Rosvall & Bergstrom 2008) on an unweighted graph connecting pairs of forms that have a ‘cognateness distance’ less than a threshold θ : In general, we use the threshold of $\theta = 0.55$ that performed best in List et al.’s analysis. However, for comparison we also present results generated from a threshold based on an approach that is more “splitting” (D2: $\theta = 0.35$) and more “lumping” (D3: $\theta = 0.75$).

An alternative to LexStat is the OnlinePMI method published by Rama et al. (2017). OnlinePMI performed better than LexStat in identifying cognate classes (Rama et al. 2017), but it gave worse results when reconstructing trees using Bayesian phylogenetics on the bases of these classes (Rama et al. 2018).¹

¹It is not particularly surprising that an algorithm may perform better in one of these tasks, but worse in the other task. The measure used for assessing the quality of cognate coding may, for example, barely punish a small systematic bias in the cognate classes, which however is important when using these classes to reconstruct trees.

Compared to LexStat, the OnlinePMI method is better able to handle data with lower numbers of shared concepts. This comes at the cost of a significant random component, making it harder to optimize OnlinePMI for a particular application.

Due to these advantages and disadvantages, we chose LexStat as our Baseline method, but included OnlinePMI for a comparative investigation (D4). In that analysis we cognate-coded our data using Online PMI, with parameters $\alpha = 0.75$, initial cut off $c = 0.5$ and batch size $m = 256$, following the general observations by Rama et al. (2017). We use the implementation available from https://github.com/evolaemp/online_cognacy_ident in the version from commit 3b998ae.

2.3. Lexical borrowing

As mentioned above, the history of the Timor-Alor-Pantar languages is strongly influenced by language contact, and multi-linguality is usual in the region (Klamer 2017, Holton & Klamer 2017). At present, standard Indonesian and local Malay have a strong influence on the vocabulary of all the TAP languages, and Tetun has influenced those spoken in Timor Leste.

Some studies have argued that the results of phylogenetic analyses will be reasonably accurate even if 15% of the vocabulary constitute undetected loanwords (Greenhill et al. 2009, Chang et al. 2015). Tree models can by their nature not encompass borrowing, so that undetected loans and chance resemblances in the data are problematic for them. Therefore, some studies exclude loans by representing a loan word that is connected to a specific meaning as a ‘missing’ or ‘unknown’ form. However, as Chang et al. (2015) point out, loans represent one case of exactly the type of lexical replacement that the phylogenetic inference models purport to model, so it

would be ill-advised to treat them as missing or unknown data when applying such models.

All languages, however, have loans that result from borrowing events that occurred so recently that the loans entered the languages as independent events, i.e. they were not inherited from any level of common ancestor. For example, most languages in LexiRumah express the concept ‘church’, a concept which became widely known in the region only in the last 300 years, with forms like [gered̥ʒa] or [igred̥ʒa], which are obvious recent Indonesian viz. Portuguese loans. Such recent loans do not carry the same phylogenetic signal as the rest of the data. Yet, the presence of a similar word makes the donor language and all the recipient languages appear closer to each other than their actual genealogical relationship would suggest. Therefore, it makes sense to treat each of these recently borrowed forms as an isolate cognate class, distinct from each other and any other cognate class in the data.

In our binary model of cognate evolution (see Section 2.4), in which each pair of a cognate class and a meaning is coded separately as either present or absent, this is functionally equivalent to only providing presence/absence features for non-loanword cognate classes, and coding all cognate classes for a concept expressed using a recent loan as “absent”. We do this as follows.

LexiRumah contains word lists for Indonesian and different Tetun varieties. Indonesian and the locally used variety of Malay are lexically very similar, and both are the source for the majority of recent loans in the TAP languages (Klamer 2017: 11-12). In order to exclude recent loans from our dataset, we therefore set all cognate classes to be “absent” for concepts in languages where the only form appears to be cognate with an Indonesian or Tetun form.

Gray et al. (2010) have suggested Q-residuals in order to assess the ‘binary

tree’-like nature of phylogenetic data. Holman et al. (2011) argue (albeit on ASJP data) that the δ score is a better measure of actual language contact events, so high Q-residuals at the same time as low δ scores may be indicative of dialect chain breakup.

We computed both δ scores and Q-residuals for our data using the *phylogeometric* package by Greenhill (2016). The results for the TAP languages are listed in Table 3. For comparison, the means for the well-studied and more tree-like Indo-European language family (using data by Chang et al. (2015)) and varieties from the Chinese dialect cluster (using data from List et al. (2014)) are also provided. Because of a bug in software used to compute some of these metrics as reported in the literature (Greenhill, p. c.), we also re-computed the δ scores and Q-residuals for the two other data sets, instead of referring to literature values.

The mean Q-residuals for AP are three times as high as for Indo-European, whereas the difference between their respective mean δ scores is quite small. This suggests that dialect chain formation and break up may be a factor in the TAP language history.

2.4. Model of language evolution

The basic principle of Bayesian phylogenetic inference is the application of Bayes’ Theorem

$$P(\text{Tree} \mid \text{Data}) = \frac{P(\text{Data} \mid \text{Tree}) \cdot P(\text{Tree})}{P(\text{Data})} \quad (1)$$

to calculate a posterior probability distribution over all possible language trees. Trees have a higher posterior probability when they are more compatible with either (i) the data (that is, trees with a higher likelihood $P(\text{Data} \mid \text{Tree})$) or (ii) our knowledge about how language phylogenies look in general

Language	Q residual	δ score
Abui, Atimelang	0.00438	0.238
Abui, Fuimelang	0.00457	0.242
Abui, Petleng	0.00504	0.280
Abui, Takalelang	0.00545	0.249
Abui, Ulaga	0.00644	0.258
Adang, Lawahing	0.00454	0.288
Adang, Otvai	0.00464	0.270
Blagar, Bakalang	0.00401	0.215
Blagar, Bama	0.00421	0.215
Blagar, Kulijahi	0.00395	0.217
Blagar, Manatang	0.00568	0.243
Blagar, Nule	0.00392	0.216
Blagar, Pura	0.00481	0.239
Blagar, Tuntuli	0.00454	0.217
Blagar, Warsalelang	0.00407	0.218
Bunak, Bobonaro	0.00419	0.270
Bunak, Maliana	0.00393	0.282
Bunak, Suai	0.00420	0.270
Deing	0.00468	0.264
Fataluku	0.00367	0.268
Hamap, Moru	0.00525	0.293
Kabola, Monbang	0.00435	0.283
Kaera	0.00798	0.292
Kafoa	0.00663	0.334
Kamang, Atoitaa	0.00358	0.244
Kiraman	0.00447	0.304
Klamu	0.00496	0.298
Klon, Bring	0.00576	0.335
Klon, Hopter	0.00481	0.320
Kui, Labaing	0.00420	0.313
Kula, Lantoka	0.00369	0.256
Makasae	0.00399	0.281
Oirata	0.00421	0.303
Papuna	0.00420	0.238
Reta, Hula	0.00739	0.259
Reta, Pura	0.00549	0.242
Reta, Ternate	0.00534	0.265
Sar	0.00403	0.238
Sawila	0.00421	0.266
Suboo	0.00341	0.235
Suboo, Atiibaai	0.00433	0.250
Teiwa, Adiabang	0.00476	0.257
Teiwa, Lebang	0.00583	0.255
Teiwa, Nule	0.00493	0.258
Tiyei	0.00422	0.259
Wersing, Maritaing	0.00374	0.255
Wersing, Taramana	0.00411	0.266
Western Pantar	0.00375	0.300
TAP mean	0.00468	0.264
IE mean	0.00128	0.201
ZH mean	0.00462	0.298

Table 3: Q residuals and Delta scores for all lects in the dataset used, plus the mean values for Indo-European (IE) (Chang et al. 2015) and Chinese (ZH) (List et al. 2014) for comparison.

(that is, trees with a higher prior probability $P(\text{Tree})$). We therefore need to (i) assume a stochastic model that describes the evolution of cognate classes on any given tree and (ii) specify our prior knowledge about the shape of the Timor-Alor-Pantar tree.

The stochastic model will be evaluated many millions of times, so it needs to be computationally very simple, while still reflecting the evolution of the forms, through the perspective of meaning-annotated cognate classes as well as possible. From a linguistic perspective, we might expect that ideally, a cognate class for a particular meaning should come into existence only once in history, and every attested instance of such a cognate class should be explainable as derived from that single origin through inheritance or borrowing. In the phylogenetics literature, this is known as the Dollo assumption (after Dollo (1893)).

In linguistic reality, however, this idealisation does not hold. In addition to chance similarities and borrowings, as discussed above, there are universal patterns in lexical semantics (List & Greenhill & Anderson, et al. 2018, Zalizniak 2018) and universal tendencies in how forms change their lexical semantics over time (Traugott & Dasher 2001, Heine & Kuteva 2005). As a result, it is possible to find formally related words filling one and the same meaning slot in different languages, even though their most recent common ancestor may have used an unrelated form for that meaning. This effect is known as homoplasy in biology (Jäger & List 2018).

Stochastic Dollo models are computationally intensive, because they cannot be computed for part of a tree independent of every other part. Combining this with the effects of imperfections in cognate coding and homoplasy implies that stochastic Dollo models cannot be usefully applied to lexical data in most cases (but see Bowerman & Atkinson (2012),

Figure 5: Possible transitions in the binary covarion model of cognate evolution with their parameterized transition rates. Letters in brackets mark the transitions discussed in the text.

Kolipakam et al. (2018), Michael et al. (2015) for cases where they can be applied).

Besides stochastic Dollo models, several alternative models of cognate class evolution are used in studies of Bayesian phylogenetics. Here we focus on binary models that take the presence or absence of a particular cognate class for a meaning into account. (Where no form is attested for a meaning slot, usually because the concept list used by the reference does collect that meaning, the data for all classes in that meaning slot is coded as ‘unknown’.) The options thus excluded are multi-state models, which have several issues due to opening a larger parameter space and unclear procedures for including synonyms, and binary models acting on the presence/absence of reflexes of a common root, independent of the meaning of such a reflexes. Such models are potentially advantageous (Chousou-Polydouri et al. 2016), but require vastly more data and manual processing than available for

this study.

In this remaining class of binary models, we consider the general binary model and the binary covarion. In both models, the presence or absence of every cognate-class-meaning pair is assumed to evolve independently. In the general binary model, jumps between presence and absence of cognate classes occur randomly with fixed rates.

The binary covarion model allows the 8 transitions shown in Fig. 5, with rates parameterized as shown. A particular pair of a cognate class and a meaning may originally have been present but be lost over time, or it was originally absent but was innovated. Like in the general binary model, such transitions between a cognate class-meaning pair being present and absent are possible directly. However, this happens at a very low (“cold”) rate, see transitions A and B in Fig. 5. In addition, a second “hot” path allows the transition via two much more transient intermediary states. The hot/cold metaphor here comes from models of physical processes, where increasing the temperature de-stabilizes a system (eg. melting, de-magnetizing); the covarion model’s “cold absent” and “cold present” states are each very stable, whereas transitions between “hot absent” and “hot present” can be relatively fast and frequent. Studies comparing different binary models with cognate class-meaning data found that in most cases (Gray et al. 2009, Lee & Hasegawa 2011, Kolipakam et al. 2018) the best performing model is binary covarion. Sometimes, though, eg. by Bown & Atkinson (2012), it is found that binary covarion performs second-best after a Dollo model. The general binary model has generally been found to be outperformed by both these models. We test the simpler general binary model to generate a baseline, compare it with the Binary covarion model, and use the better fitting model in subsequent model comparisons.

A simplified example from our data to illustrate some of the transitions in Fig. 5 would be the following. The proto-TAP form **lamV* ‘walk’ was inherited in proto-AP as **lam(ar)*, and a modern reflex of it is Klon *lam* ‘walk’. Teiwa also has a word *lam*, but it has shifted its meaning to ‘follow (a path/river)’, and the word *tewar* is now the word used for ‘walk’. Thus, the pair *<*lam(ar)-‘walk’>* did not undergo a transition in Klon, but it has become absent in Teiwa, thus changing from “cold present” to “cold absent”, cf. transition A in Fig. 5.

However, the model does not only account for transitions between present and absent at a very low rate; it also contains two transitory states: “hot present” and “hot absent”. Applied to our example, the pair *<*lam(ar)-‘walk’>* would go from “cold present” to “hot present” (Fig. 5, C) in a language when the word is still in use but speakers also start to use another word as an alternative or synonym for it. If over time, this synonym replaces the *<*lam(ar)-‘walk’>* form, the latter will fall into disuse and will only be known by some (older) speakers as the archaic form for ‘walk’. This change would illustrate a transition from “hot present” to “hot absent” (Fig. 5, D). The next stage might be that the word completely disappears, thus switching to the “cold absent” state (Fig. 5, E). Transitions between the intermediate states of “hot present” and “hot absent” are assumed to occur at a much higher rate than any of the other transitions. So, the most likely path for the loss of a cognate class does not go from “present” to “absent” but from “present” to “hot present” to “hot absent” to “absent”. Similarly the most likely path for an innovation would not be to simply go from “absent” to “present” but first to become “hot absent” (e.g. an innovative word introduced by a few speakers), then “hot present” (the new word used in variation with the exist-

ing word), and then become “present”. In the model, the switches between “hot” and “cold” states can reflect different speeds of evolution of cognate-class-meaning pairs in different parts of the tree.

The technical details are as follows. The transition rates are chosen such that the marginal probability to be in a “present” state is π_1 and the equilibrium probability for “absent” is $\pi_0 = 1 - \pi_1$. The vertical transition rates between hot and cold states are balanced such that the system is expected to be “hot” and “cold” each half of the time. The switch rate s is inferred using a Γ prior with parameters $\alpha = 0.05$ and $\beta = 10$, the slower rate α is distributed uniformly between 10^{-4} and 1, and the underlying present/absent frequencies π_0 and π_1 are inferred assuming a flat Dirichlet prior. Rates are normalised to an expected 1 present/absent switch per time unit, and re-scaled with the log-normally distributed clock rate, and a Γ -distributed rate for each cognate set. This Γ rate variation reflects that not all concepts change forms at equal speeds; instead, the replacement speeds between different concepts vary widely (Greenhill et al. 2010, 2017, Chang et al. 2015). Efficient algorithms for per-site (i. e. per-cognateset) rate variation are available, while per-concept rate variation is computationally more intensive, so we use per-site variation.

2.5. Tree prior, clocks and calibrations

In addition to deciding on the evolutionary model, we also need to specify our prior knowledge of the TAP tree. One type of prior knowledge involves the rate in which languages split from their immediate ancestor. A *uniform* tree prior assumes that the internal nodes of the tree have ages that are uniformly distributed between the tips and the

root of the tree (i.e., that we have absolutely no prior information on the age of any subgroups). The advantage of such a prior is that it is computationally very simple, but the disadvantage is that it does not have any basis in how languages evolve in reality (Barido-Sottani et al. 2018, du Plessis 2018). In addition, a uniform tree prior may bias the inference procedure towards more extreme clade probabilities (Yang & Rannala 2005): A subgrouping that is compatible with the data, but at low probability, may be suppressed and not show up in the posterior sample.

In biology, a well-established prior on trees is the coalescent tree prior. It describes the possible ancestries of two genes in a large mixing population: Tracing the gene back in the family tree of different individuals from the population, when do the ancestries of the gene coalesce to a single gene in a single ancestor? Due to the assumption of a large, mixing population, this tree prior does not appear particularly suited to our data.

A third type of tree prior is the Fossilized Birth-Death (FBD) prior, which assumes that languages split from their ancestor at a constant birth rate, and go extinct at a different, but also constant death rate, while leaving fossils behind to the present day with a low chance. (This chance is 0 for our data, which does not contain independent evidence for historical languages.)

Rama (2018) has compared the uniform, coalescent and FBD priors on Indo-European data, and found that the trees inferred using the FBD prior and the uniform prior both closely match the traditionally established Indo-European tree, while the coalescent prior performs much worse. Rama concludes that “any future phylogenetic experiment *should test uniform tree prior* as a Baseline before testing more parameter-rich priors such as FBD or coalescent priors.” (Rama 2018). In our analysis, we thus test both the

FBD and the uniform tree prior.

Our prior knowledge about the TAP tree also contains information about the fact that certain subgroups split, and about times when that happened. While dating the nodes in the tree is not our primary goal here, it is known that including multiple calibration dates in a tree can influence its topology, that is, change its subgrouping. Since we are interested in the subgrouping of TAP, we investigate the effects of including calibration dates on subgrouping in the TAP tree.

The migration of speakers of the ancestors of today’s Oirata speakers from east Timor to the adjacent island of Kisar, i.e. the date when Oirata split from its relatives Makasae/Makalero and Fataluku, can be dated with high confidence to the year 1721 or shortly before that (Hägerdal 2012: 337). We therefore include a normally distributed calibration on the split that is directly ancestral to Oirata in all our inferences with 95% probability (2σ) on the 7 years until 1721, a distribution with mean 282.5 years Before Present (yBP) and standard deviation 1.75 years.

Calibration dates for other splits in the tree are based on circumstantial evidence. One observation that has been made is that the level of (lexical and) grammatical similarity within the AP sub-branch does not support an age of more than a few millennia (Klamer 2017: 10). The reconstructed vocabulary of proto-AP appears to contain Austronesian (Malayo-Polynesian, MP) loan words such as proto-AP *baj ‘pig’, cf. proto-MP *babuy; and proto-AP *bui ‘betelnut’, cf. proto-MP *buaq ‘fruit’ (Holton et al. 2012). These ancient MP loans are found across the Alor-Pantar branch, and they follow regular sound changes, which indicates that they were borrowed at the level of proto-AP. We use them to approximately date when proto-AP split from proto-TAP. Proto-MP has been dated at 4 000 yBP (Pawley 2002), and speakers of MP languages are commonly

assumed to have arrived in the Timor Alor Pantar area around 3 000 yBP (Pawley 2005, Spriggs 2011). The MP borrowings in proto-AP would thus give the AP family a maximum age of around 3 000 years.

More precisely, human genetic studies support a connection between populations of the Lesser Sundas with Papuan populations of New Guinea and Austronesian-Malayo-Polynesians from Asia (Lansing et al. 2011, Xu et al. 2012). The admixture between Papuan and ('Melanesian')-Asian is estimated to have begun about 5 000 yBP in the western part of eastern Indonesia, decreasing to 3 000 yBP in the eastern part. This associates the Papuan-Asian admixture with Austronesian-MP expansion (Xu et al. 2012). This is the only date range with quantitative confidence, and gives the confidence interval used in our analysis. The 90% confidence interval between 2 870 and 3 405 yBP corresponds to a normal distribution with mean 3 137.5 yBP, and a standard deviation of 162.6 years. The dating is thus consistent with the scarce linguistic and archaeological evidence for the contact time frame around 3 000–3 500 yBP.

Unlike proto-AP, proto-TAP does not have any reconstructible MP borrowings. There are significant lexical and grammatical differences between the languages of the AP branch and those of Timor (Holton et al. (2012: 115) and references cited there), suggesting a long period of separation for the branches. This is unsuitable for deriving calibrations, but for the technical reason to guarantee a proper prior on the tree – even for the uniform tree prior with very little data – we assume that the age of the root is less than 18 000 years (using a uniform prior).

As such, our data permits only two actual calibration points. We therefore expect that we do not have sufficient data to distinguish between a strict clock, which implies the same speed of evolution to be the same throughout

the whole tree, and a relaxed clock, which allows the rate of evolution to vary between different parts of the tree by assigning a different relative rate of evolution to each branch in the tree.

However, as with the previous model choices, we first implement the simpler choice (the strict clock) and then compare it to the more elaborate and potentially more realistic model (the relaxed clock), and keep the better one for our conclusive analysis. Our relaxed clock uses a log-normal distribution with mean 1 for its per-branch clock rates, and the σ of the log-normal distribution is drawn using from an improper $[0, \infty)$ uniform prior (starting at $\sigma = 0.2$), to allow the data to inform us how clock-like the evolution is.

2.6. Inference procedure using MCMC

The prior, the likelihood, and the data together are in theory sufficient to know the posterior distribution of trees compatible with the data. There are, however, $2.98744 \cdot 10^{72}$ different tree topologies for our 48 TAP languages, and only a very small proportion of these have any relevant posterior probability. The number of trees is so immense that to compute them all is impossible. To avoid this, we resort to the Markov Chain Monte Carlo (MCMC) sampling method and make use of the fact that similar trees have somewhat similar probabilities. Instead of drawing trees entirely at random and calculating their probabilities (a process known as Monte Carlo sampling), the MCMC sampling method constructs a random walk through the space of trees. It has a memory of one previous state (which is the characteristic property of a Markov chain) and uses small random modifications, such as swapping two sub-trees, to explore the search space of all trees in a more

structured manner. If the modification leads to a more probable tree, the algorithm accepts it and takes it as its new state. If the posterior probability of the new state is worse than the posterior probability of the old state, there is still a chance that the algorithm accepts it and moves to it, otherwise it is rejected, and a different change away from the same original location is tried. Trees that are in the state more often are thus exactly the more probable trees. The rejection probability is calculated such that by tracking the state of the chain at regular intervals, we derive a sample of trees that reflects the posterior probability distribution.

The BEAST2 software package, in its most recent version 2.6.4 (Bouckaert et al. 2014, Drummond & Bouckaert 2015, Bouckaert 2018a), provides an extensible implementation of MCMC for tree inferences.

We run each inference in 4 independent chains in chunks of 22:00 hours on a dedicated computing cluster ², logging results every 5000 steps to avoid a high correlation between subsequent samples. The first samples reflect the starting point of the analysis and not its posterior distribution, so we discard the first 10% of samples, and we check using Tracer 1.7.1 (Rambaut et al. 2020, 2018) whether the chain appears to have converged and whether the resulting samples are dissimilar enough of each other to reflect an effective sample size of at least 200, for every parameter sampled. If that is not the case, we resume the chain from its last state. The run time, depending on the model complexity, may thus range from about a day to a week.

Marginal likelihood values are calculated in each case using path sampling analysis from the BEAST2 package MODEL-SELECTION. We

²https://docs.s3it.uzh.ch/cluster/quick_start/ – The BEAGLE library for BEAST was compiled without GPU support, so the chains were run for 88:00 CPU hours at a time

run the stepping stone analysis with 100 steps, where we choose the internal chain length of each step such that the total number of samples from the stepping stone analysis is bigger than the sample size of the converged posterior chains. In step 0 of the stepping stone analysis, the likelihood has full weight, while the other steps interpolate between posterior and prior by decreasing the weight of the likelihood. In order to avoid problems with burn-in and lack of convergence in the initial stepping stone step, we start the path sampling analysis from the final state of the first of our four independent MCMC chain.

2.7. Summary of model choices

To summarize, our starting analysis (S) uses LexiRumah TAP data in a comparable format and with clean meaning-form mapping; the data is automatically cognate-coded using LexStat to obtain similarity scores between forms, from which disjoint cognate classes are generated by running the infomap clustering algorithm on the graph of all pairs of forms with a distance less than $\theta = 0.55$. We treat all forms that appear to be cognate with Indonesian or Tetun as recent loans and exclude them from our data set. We infer trees from this data set (using BEAST2 for MCMC sampling) by means of Bayesian phylogenetic inference, with a general binary model with Γ rate variation between different cognate sets and trees following a Birth-Death tree prior, calibrated using a strict clock at the ancestors of Oirata (282.5 ± 1.75 yBP), the ancestor of all AP languages (3137.5 ± 162.6 yBP) and the TAP root (younger than 18 000 yBP). The only monophyly constraint in the analysis is on the Alor-Pantar languages as a whole, mandated by the calibration on that branch.

We then compare marginal model likelihoods, estimated using stepping stone anal-

ID	Cognate Coding	No. C	Evolutionary Model	Tree Model	Clock model
S	LexStat-Infomap ($\theta = 0.55$)	9274	Binary	Birth-Death	Strict clock
T	(as S)	9274	(as S)	Uniform	(as S)
M	(as S)	9274	Binary covarion	(as S)	(as S)
C	(as S)	9274	(as S)	(as S)	Relaxed clock
MT	(as S)	9274	Binary covarion	Uniform with	(as S)
MC	(as S)	9274	Binary covarion	(as S)	Relaxed clock
TC	(as S)	9274	(as S)	Uniform	Relaxed clock
X	(as S)	9274	Binary covarion	Uniform	Relaxed clock
D1	LexStat-Infomap ($\theta = 0.55$), but with ASJP instead of SCA as sound class model	9268	(as S)	(as S)	(as S)
D2	LexStat-Infomap ($\theta = 0.35$)	14596	(as S)	(as S)	(as S)
D3	LexStat-Infomap ($\theta = 0.75$)	5259	(as S)	(as S)	(as S)
D4	Online-PMI	9268	(as S)	(as S)	(as S)

Table 4: Summary of the phylogenetic analyses considered in this article. ID = names of models of analysis as referred to in the text. No. C = the resulting number of binary characters included in the analysis. All analyses use per-character Γ rate variation and split calibrations on TAP, AP, and the parent of Oirata, and exclude Indonesian and Tetun loans.

ysis from the MODEL-SELECTION package for BEAST2 (Bouckaert 2018b), to see whether the fit of the model is improved by:

1. replacing the strict clock with an uncorrelated relaxed clock,
2. replacing the binary substitution model with a binary covarion substitution model, and
3. replacing the birth-death tree prior with a uniform tree prior.

We test all combinations of two of these variants, and would test the combination of all three variants if any variant improves upon the starting model. (That is, a variant needs to underperform twice to be rejected.) We check the robustness of the results from the best model by testing different cognate coding methods (S compared with D1-D4).

3. Results

All configurations and results of the phylogenetic inferences are available in the supplementary material.

Model Selection Even after running several chains with optimized operators (Douglas et al. 2021) and Metropolis-coupling (Müller & Bouckaert 2019, Altekari et al. 2004) for 200 000 000 steps in total, the relaxed clock variant (C) persistently showed bad mixing between what appeared to be two likelihood regimes. The results of combining the relaxed clock with any other modification (MC, TC) did not seem to show improved convergence. It would thus be difficult to assess the quality of the MCMC even if the summary statistics suggested convergence (Brown & Thomson 2018). The log files of these unconverged chains can be found in the supplementary material. Because the chains consistently find these two

Analysis	Marg. log-lik	Δ to best
S	-49315	
T	-49342	-27
M	-56944	-7629
C	Did not converge	
MT	-57431	-8116
MC	Did not converge	
TC	Rejected due to previous convergence issues	
X	Rejected as per a priori model selection decision	

Table 5: Estimated marginal log-likelihood of the different analyses. A higher marginal log-likelihood implies that the data is evidence in favour of that model. The Δ values are the logarithmic Bayes factors in favour of the best model.

regimes and are able to jump between them, and no other likelihood regimes show up, there is very weak indication that properties shared between both regimes might be indicative of the posterior space. Of interest in this context is the variability of the per-branch clock rates, which constitutes the main difference between the relaxed clock and the strict clock. In the regime with higher clock variance, the standard deviation in log space of the clock rate has a mean of 0.935 with a maximum 1.4337. 95% of the samples (thus likely also a high probability mass in the converged posterior) of the coefficient of variation for the relaxed clock lie between 0.7 and 1.6. Such numbers would suggest that the relaxed clock, if it could be trusted, would be a better fit than the strict clock, but the phylogenetic signal is indeed more clock-like than non-clock-like (Drummond & Bouckaert 2015: 144). Because of the lack of convergence, we therefore consider only the strict clock in our following results, instead of comparing it to the relaxed clock or resorting to models without clock, and caution against giving interpretative weight to the node ages we infer.

The remaining four models (S, M, T, and MT) show apparent convergence for all parameters, with the longest run having 40 000 000 MCMC steps, so we choose 100 000 000 steps

for the sum of chain lengths for model selection. Using BEAST2’s MODEL-SELECTION 1.4.0 package (Bouckaert 2018b), we run a stepping-stone analysis (Xie et al. 2011, Lartillot & Philippe 2006) with 100 steps for each analysis, each running for 1 000 000 internal steps. This method estimates the marginal likelihood of a model, or, equivalently, the evidence in favour of that model. The marginal likelihoods of our analyses are summarized in Table 5. Surprisingly, the starting model (S) has a higher model likelihood than all the available variant models (M, T, and MT). For the path sampling analyses, the Path Sample Analyser tool accompanying BEAST2’s MODEL-SELECTION add-on reports effective aggregated sample sizes between 14 078 (MT) and 28 698 (S).

Posterior tree sample of the model with highest marginal likelihood

The tree sample from our MCMC S is summarized in Fig. 6. The mean root ages for the four topologies all lie between 5979 and 6089 yBP. Of all trees in this posterior sample, 95% have root ages between 4816 and 7636 years.

Visualizations of all other posterior samples can be found in the Appendix appendix C. The most deviating tree samples come from the extreme cognate coding methods (D2, and

Figure 6: Summary of the posterior tree sample S (general binary model, strict clock, birth-death tree prior), showing all topologies with at least 5% support, with darkness proportional to support (45%, 22%, 13%, and 6%, respectively). The age of each node is its mean age in that topology. The order of lects, from top to bottom, roughly follows the locations of languages from east (Timor) to west (Pantar).

ever more so D3). The standard cognate coding methods (S, D1, D4) are overall very similar, despite the differences in sound classes and coding procedure, respectively.

4. Discussion

In our results, we found that a standard model outperforms the alternatives in terms of marginal likelihood. This is surprising and not reflected in the literature on linguistic phylogenies, where the uniform tree prior had been found to compete with the more mechanistic birth-death tree prior (Rama 2018), and the binary covarion model usually greatly outperforms the simpler binary Markov chain model (Gray et al. 2009, Kolipakam et al. 2018).

We showed that rough cognate coding can greatly influence the higher-order splits inferred, by lumping forms that are likely unrelated, or splitting cognate sets that do indicate relatedness. However, using up-to-date cognate coding methodology with proper parameter values, optimized on other language families, gives reliable tree estimates independent of the specific coding choice.

Our analysis shows a very strong signal of subgrouping of the Timor-Alor-Pantar family. It shows an early split of East Alor, and the remaining Nuclear AP group splits into Central Alor (with Abui, Kamang, Kui, Kiraman, Subo, Tiyei, and probably Kafoa) and West Alor-Straits-Pantar (WASP), with the latter splitting into West Alor (Hamap, Kabola, Adang, and maybe Kafoa), and a Pantar-Straits group. Whether Klon is part of the West Alor clade or a first-order member of WASP is unclear. The Straits-Pantar group clearly shows a split between the languages of the Straits (Blagar, Pura and Reta) plus Kaera versus Western Pantar versus the Pantar languages (Klamu, Teiwa, Deing, Sar).

In the following, we will consider the spe-

cific findings from our analysis concerning the internal structure of the Timor-Alor-Pantar languages in light of the linguistic evidence for the different subgroupings suggested in earlier work.

The differences between Robinson & Holton (2012a) and our core analysis S are both due to the quantity and quality of the data as well as the chosen methodology. A major difference is that the tree in Fig. 3 was improperly rooted, using a reconstructed proto-Alor-Pantar lexicon as outgroup. This aggravated biases already present in the rather scarce data set. In addition to this rooting issue, the differences largely affect the affiliation of the Pantar languages, which also proved difficult to classify in Holton et al. (2012), and continue to resist a clear binary classification.

The traditional linguistic evidence for a tree of AP languages is given in Holton et al. (2012), which describes systematic phonological innovations in 12 AP languages, see Fig. 2. This tree is also the one represented by Glottolog (Hammarström et al. 2020). It has the East Alor languages all fairly deeply embedded in the tree and shows relatively little resolution and no delineation of a Pantar subgroup, so that all Pantar languages are considered as primary split-offs from proto-AP. East Pantar is thus suggested as the location of origin and original break-up of the AP family (Holton et al. 2012: 118).

There are thus two major differences between our analysis and earlier topologies, (i) in our topology the languages of East Alor are a high-level subgroup where they were previously considered to be embedded deeply in the tree; (ii) in our topology the languages of Pantar are a low-level subgroup inside WASP, where previously they were considered first-order members of AP (cf. Figs. 2 and 3).

The differences between our analysis and this earlier application of the classic compara-

Sound changes	Languages
*b > f	Teiwa, Klamu, Abui (in Teiwa and Klamu only non-initially)
*b > p	Kamang, Sawila, Wersing (in Kamang not medially)
*d > r	Abui, Kui (in Kui only finally)
*g > ʔ	Blagar, Adang
*k > ∅	Blagar, Adang
*q > k	W Pantar, Blagar, Adang, Klon, Kui, Abui, Kamang, Sawila, Wersing (Adang ʔ < k < *q)
*s > h	Blagar, Adang, Klon
*s > t	Abui, Sawila, Wersing
*h > ∅	everywhere but Teiwa and W Pantar
*m > ŋ / _#	W Pantar, Blagar, Adang
*n > ŋ / _#	Klamu, Kaera, W Pantar, Blagar, Adang, Abui, Kamang, Sawila, Wersing
*l > i / _#	Teiwa, Kaera, Adang, Kamang
*l > ∅ / _#	Klamu, W Pantar, Abui
*r > l / V_V	Klamu, W Pantar, Adang, Kamang
*r > i / _#	Blagar, Kui, Abui

Table 6: Sound changes in Alor Pantar languages observed by Holton et al. (2012: 113).

tive method of (Holton et al. 2012) may be due to the paucity of TAP language data at the time (short word lists of only 12 lects), as well as the methodology of manual comparison.

The early East Alor split we observed in our analysis is supported by regular innovations, of which we will mention three types in the following (there may be more). Proto-Nuclear Alor Pantar saw a lexical innovation, adding a lexeme *-om ‘inside’ to the expression used for ‘(be) inside’, while proto-East-Alor retained the original form reflecting proto-TAP *mi ‘be in, at’; which is also found in reconstructions for the Timor branches (Klamer 2018), see Appendix Section D.1

Here we also mention two phonological innovations that support the split. Proto-TAP intervocalic *d changed into a liquid *r in proto-Nuclear AP, but was retained in proto-East-Alor; see Appendix Section D.2, where this is illustrated for proto-TAP *hada ‘fire’.

(Other cognate classes showing this pattern are ‘sugarcane’, ‘tongue’ and ‘new’; see the data in Kaiping et al. (2019b)). The opposite pattern of innovation cum retention is found for proto-TAP intervocalic *b. This obstruent became voiceless *p in proto-East-Alor but was retained in proto-Nuclear AP, see Section D.3, where this is illustrated for proto-TAP *habi ‘fish’. Other cognate classes showing this pattern are ‘bat’ and ‘sun’ (see the data in Kaiping et al. 2019b).

The early split-off of East Alor is thus conceivable from the perspective of historical linguistics, even though the systematic sound changes do not give unanimous evidence for it, cf. Table 6. Long-lasting and intensive language contact between neighbouring communities can explain the discrepancy. Forms that are borrowed may look similar without showing the regular sound correspondences dividing East Alor from Nuclear AP languages. As

such, they would be grouped together only under a high cognacy threshold θ , so our analysis D3 corroborates that hypothesis. Phylogenetic network models, which explicitly take borrowing into account, may be used to assess this hypothesis once they become tested and feasible for large data sets such as ours: We expect that forms grouped together under a higher threshold θ will not be purely noise that decreases model likelihood, but in part confirm lateral connections inferred also for a lower θ .

Early split-offs are often taken to reflect the greatest age, so that the location of such a split may point to an area where the proto-language began to diversify, c.f. Sapir (1916: 79f)’s “centre of gravity principle”. An early East-Alor primary subgroup of AP may suggest that the proto-AP speakers first arrived on the (south-)eastern coast of Alor and then spread westwards over the course of subsequent centuries. This scenario contrasts with earlier proposals where the homeland of AP was hypothesized to be on Pantar, with a subsequent direction of movement of the languages from west to east (Holton et al. 2012); or where the homeland was on an island in the Straits between Alor and Pantar (Robinson & Holton 2012b: 143), with languages moving west to Pantar, and east to Alor. A phylogeographic inference, which takes the cultural adaptation necessary to move between different biomes (eg. from coast to mountains) into account, may be able to give deeper insight into the migration history of the region, in particular if it was informing and informed by language contact in a network phylogeny.

Checking the defining phonological changes against our current data set, we find that the most widespread of these changes is the one involving the uvular stop /q/. Although a uvular stop is found in only a few of the modern languages, the reconstruc-

tion of pAP *q is supported by a number of correspondence sets Holton et al. (2012: 102–103), Holton & Robinson (2017: 52). The sound change *q > k occurs in all languages of Alor and the Straits, but not in the Pantar languages. (The few existing irregularities may be explained by diffusion.) Cognate sets showing this regular change are ‘spear’, ‘ten’ and ‘two’. This change results in a merger of *k and *q in virtually all of the languages of Alor and the Straits (Holton et al. 2012: 113) and is therefore incompatible with both major topologies found in our lexical dataset: the first order split of East Alor languages and the low-level grouping of the Pantar languages.

The sound change *s > h (in initial position of ‘chicken’ and final position of ‘mat’) is evidence for an innovation-defined subgroup that includes the languages of West Alor and Straits (Holton et al. (2012: 113)), namely Adang, Hamap, Kafoa, Klon and Blagar. Additional innovations uniting this group are the innovation of a unique (subtractive) morphological shape for the numerals ‘seven’, ‘eight’ and ‘nine’ (Schapper & Klamer 2017: 292f.) and the West Alor-Straits lexical innovation *dol ‘mountain’, where proto-AP has *buku. This set of changes may be seen as further evidence that the lexical tree shown in Fig. 6, which does not group West-Alor with Straits excluding Pantar, is not reflecting all elements of the complex history of the TAP languages. Note also that there are two languages in Fig. 6 that are not monophyletic, namely Adang, with Kabola and Hamap; and Teiwa, with Sar. This shows that even today, the patterns of linguistic signal intersect non-trivially with speakers’ perceptions of which other lects belongs to their dialect chain.

We speculate that the low performance of the uniform tree prior (compared with our expectation set by Rama (2018)) is also related: Following Yang & Rannala (2005), the

long branch lengths of the uniform tree prior favour extreme clade probabilities, so the posterior tree distribution under the uniform prior (T, MT) is unable to account for the highly conflicting signal in the data.

The conflicting subgrouping evidence in the earlier literature and the different tree structures discussed here may point to a historical scenario where proto-Alor-Pantar split into proto-Nuclear Alor Pantar and proto-East Alor. Proto-Nuclear Alor Pantar likely formed a dialect chain (linkage), allowing overlaps in lexical material and sound changes, and innovations shared by neighbouring languages. It then broke up into proto-Central Alor and proto-WASP. The contact between the initially mutually understandable neighboring dialects of proto-Nuclear Alor Pantar persisted while the linkage developed the less porous boundaries of today's Alor-Pantar languages. The tree corresponding to this synthesis shown in Fig. 7.

The hypothesis that the history of the Alor Pantar languages may have involved a dialect chain that broke up is supported by evidence for regional networks of trade and marriage between the various speech communities. Networks that have been reported include alliances between groups on West Alor (Adang, Kabola, Kui) and the Straits (Blagar) (Wellfelt 2016: 228; 277), connections between coastal people and people in the interior of East Alor (e.g. Wellfelt 2016: 73), connections between people in the Central Alor and South Alor (Wellfelt 2016: 177), and connections between groups in East Alor and East Timor (Wellfelt 2016: 99). Regional networks like these must have also shaped the linguistic landscape.

5. Conclusions

In conclusion, our analysis indicates the subgrouping structure of the Timor-Alor-Pantar language family shown in Fig. 7.

The major features of this tree are consistently inferred using various state-of-the-art methods of cognate detection and tree inference. The newly found early split of the Alor-Pantar branch into an East-Alor subgroup and a Nuclear Alor-Pantar subgroup robustly appears from all our sensible inferences (all models, and all non-extreme coding procedures.) This early split can be supported by lexical and phonological innovations. It suggests that the Alor-Pantar speakers arrived in the east of Alor, followed by a subsequent westward movement of the languages.

However, our findings also indicate that proto-AP consisted of a chain of related dialects, and the history of most AP languages is characterized by high levels of mutual interaction and mobility. We speculate that this situation could arise either because different bands of proto-AP speakers arrived around the same time and settled in different locations in, or around the Straits; or because bands of proto-AP speakers left a single original point of arrival relatively shortly after they had arrived, but stayed in contact afterwards, or a combination of both. A similar dispersal scenario has been proposed for Alorese, an Austronesian language that arrived in the Straits around 1300–1400 AD with speakers that had various different settlements on both Pantar and West Alor (Klamer 2011: 10–12).

We also found that computational accounts which assume a binary tree model lead to an inaccurate historical picture, due to the lack of models that explicitly deal with dialect chain data. This is why we made an effort to check the adequacy of the different models against each other, and to tie the derived phylogenies back to the linguistic ev-

Figure 7: Subgrouping structure of the TAP language family.

idence regarding the phonological and lexical innovations in the languages investigated. While computational models for punctuated contact events are slowly becoming feasible (Kelly & Nicholls 2016, Vaughan et al. 2017, Zhang 2018), none of them are available for the inference of large-scale language contact, and inference procedures taking into account the overlapping isoglosses from dialect chains (François 2014, Kalyan & François 2018) have not been proposed yet.

6. Acknowledgements and Contributions

GAK and MK designed the study. GAK implemented the steps of the computational analysis. MK investigated the classical historical signal indicated by the analysis. GAK and MK wrote and revised the article. The authors declare that they do not have competing interests.

We would like to thank two anonymous reviewers, who considered the data and evidence in this article and the preliminary analysis presented in Kaiping & Klammer (2019) in great depth, and provided very helpful feedback on improving the article and the analyses therein. We thank Luke Maurits and Natalia Chousou-Polydouri for helpful advice on apparent incongruencies in earlier versions of the results, and Nico Neureiter, Denise Kühnert, and Natalia Chousou-Polydouri for useful discussions concerning the practical aspects of phylogenetic analyses in BEAST. This research was mostly supported by the VICI research project ‘Reconstructing the past through languages of the present: the Lesser Sunda Islands’ at Leiden University, awarded to Marian Klammer and funded by the Netherlands Organisation for Scientific Research (NWO), project number 277-70-012. GAK was partly funded by the NCCR “Evolving

Language” of the Swiss National Science Foundation (SNF).

Supplementary Material

Data and methods used to produce the results shown in this paper are available as online supplementary material at

https://osf.io/h4nxx/?view_only=8dd7b81f03304d5085191424bd709517.

References

- Altekar, Gautam & Dwarkadas, Sandhya & Huelsenbeck, John P. & Ronquist, Fredrik. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20(3). 407–415. <https://doi.org/10.1093/bioinformatics/btg427>.
- Baird, Louise. 2008. *A grammar of Klon: a non-Austronesian language of Alor, Indonesia* (Pacific Linguistics 596). Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University. 242 pp.
- Baird, Louise. 2017. Kafoa. In Antoinette Schapper (ed.), *The Papuan Languages of Timor, Alor and Pantar: volume 2*, 55–108. Berlin / New York: De Gruyter Mouton.
- Barido-Sottani, Joëlle & Bošková, Veronika & Plessis, Louis Du & Kühnert, Denise & Magnus, Carsten & Mitov, Venelin & Müller, Nicola F. & Pečerska, Jūlija & Rasmussen, David A. & Zhang, Chi & Drummond, Alexei J. & Heath, Tracy A. & Pybus, Oliver G. & Vaughan, Timothy G. & Stadler, Tanja. 2018. Taming the BEAST—A Community Teaching Material Resource for BEAST 2. *Systematic Biology* 67(1). 170–174. <https://doi.org/10.1093/sysbio/syx060>.
- Blust, Robert. 1999. Proto-Austronesian. In Robert Blust & Russell D. Gray & Simon J. Greenhill & Cordelia Nickelsen (eds.), *Austronesian Basic Vocabulary Database*. <https://abvd.shh.mpg.de/austronesian/language.php?id=280> (28 January, 2019).

- Bouckaert, Remco. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* 26(10). 1372–1373. <https://doi.org/10.1093/bioinformatics/btq110>.
- Bouckaert, Remco. 2018a. BEAST 2.5.0. Version 2.5.0. <https://github.com/CompEvol/beast2/releases/tag/v2.5.0>.
- Bouckaert, Remco. 2018b. BEAST2-Dev/Model-Selection. BEAST2-Dev. <https://github.com/BEAST2-Dev/model-selection> (15 February, 2019).
- Bouckaert, Remco & Heled, Joseph. 2014. DensiTree 2: seeing Trees Through the Forest. *bioRxiv*. 012401. <https://doi.org/10.1101/012401>.
- Bouckaert, Remco R. & Heled, Joseph & Kühnert, Denise & Vaughan, Tim & Wu, Chieh-Hsi & Xie, Dong & Suchard, Marc A. & Rambaut, Andrew & Drummond, Alexei J. 2014. BEAST 2: a Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput Biol* 10(4). e1003537. <https://doi.org/10.1371/journal.pcbi.1003537>.
- Bowern, Claire & Atkinson, Quentin. 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* 88(4). 817–845. <https://doi.org/10.1353/lan.2012.0081>.
- Brown, Jeremy M & Thomson, Robert C. 2018. The Behavior of Metropolis-Coupled Markov Chains When Sampling Rugged Phylogenetic Distributions. *Systematic Biology* 67(4). 729–734. <https://doi.org/10.1093/sysbio/syy008>.
- Chang, Will & Cathcart, Chundra & Hall, David & Garrett, Andrew. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244. <https://doi.org/10.1353/lan.2015.0005>.
- Chousou-Polydouri, Natalia & Birchall, Joshua & Meira, Sérgio & O'Hagan, Zachary & Michael, Lev. 2016. A Test of Coding Procedures for Lexical Data with Tupí-Guaraní and Chapacuran Languages. In *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. Tübingen: Philosophische Fakultät. https://bibliographie.uni-tuebingen.de/xmlui/bitstream/handle/10900/68644/Chousou-Polydouri_Birchall_Meira_O'Hagan_Michael.pdf?sequence=1&isAllowed=y (9 August, 2017).
- Dollo, Louis. 1893. The laws of evolution. *Bulletin de la société Belge de géologie, de paléontologie & d'hydrologie* 7. 164–166.
- Douglas, Jordan & Zhang, Rong & Bouckaert, Remco. 2021. Adaptive dating and fast proposals: revisiting the phylogenetic relaxed clock model. *PLOS Computational Biology* 17(2). e1008322. <https://doi.org/10.1371/journal.pcbi.1008322>.
- Drummond, Alexei J. & Bouckaert, Remco R. 2015. *Bayesian Evolutionary Analysis with BEAST 2*. Cambridge University Press. 263 pp.
- du Plessis, Louis. 2018. FBD-Tutorial: Dating Species Divergences with the Fossilized Birth-Death Process. Taming the BEAST. <https://github.com/Taming-the-BEAST/FBD-tutorial> (17 July, 2018).
- Forkel, Robert. 2017. CLDF: cross-linguistic Data Formats. *Zenodo*. <https://doi.org/10.5281/zenodo.835502>.
- Forkel, Robert & List, Johann-Mattis & Greenhill, Simon J. & Rzymiski, Christoph & Bank, Sebastian & Cysouw, Michael & Hammarström, Harald & Haspelmath, Martin & Kaiping, Gereon A. & Gray, Russell D. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5. 180205. <https://doi.org/10.1038/sdata.2018.205>.
- François, Alexandre. 2014. Trees, waves and linkages: models of language diversification. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge Handbook of Historical Linguistics*, 161–189. New York: Routledge.
- Gray, Russell D. & Bryant, David & Greenhill, Simon J. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 365(1559). 3923–3933. <https://doi.org/10.1098/rstb.2010.0162>.
- Gray, Russell D. & Drummond, Alexei J. & Greenhill, Simon J. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323(5913). 479–483. <https://doi.org/10.1126/science.1166858>.
- Greenhill, Simon J. 2016. Phylogetic: a Python library for calculating phylogenetic network

- metrics. *Journal of Open Source Software* 1(2). 28. <https://doi.org/10.21105/joss.00028>.
- Greenhill, Simon J. & Atkinson, Q. D. & Meade, A. & Gray, R. D. 2010. The shape and tempo of language evolution. *Proceedings of the Royal Society of London B: Biological Sciences* 277(1693). 2443–2450. <https://doi.org/10.1098/rspb.2010.0051>.
- Greenhill, Simon J. & Currie, Thomas E. & Gray, Russell D. 2009. Does horizontal transmission invalidate cultural phylogenies? *Proceedings of the Royal Society of London B: Biological Sciences* 276(1665). 2299–2306. <https://doi.org/10.1098/rspb.2008.1944>.
- Greenhill, Simon J. & Wu, Chieh-Hsi & Hua, Xia & Dunn, Michael & Levinson, Stephen C. & Gray, Russell D. 2017. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences* 114(42). E8822–E8829. <https://doi.org/10.1073/pnas.1700388114>.
- Haan, Johnson Welem. 2001. *The grammar of Adang: a Papuan language spoken on the island of Alor East Nusa Tenggara - Indonesia*. Sydney: University of Sydney dissertation. <http://www-personal.arts.usyd.edu.au/jansimps/haan/>.
- Hägerdal, Hans. 2012. *Lords of the land, lords of the sea; Conflict and adaptation in early colonial Timor, 1600-1800*. Brill. https://doi.org/10.26530/OAPEN_408241.
- Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2020. *Glottolog 4.2.1*. Jena. <https://doi.org/10.5281/zenodo.3754591>.
- Heine, Bernd & Kuteva, Tania. 2005. *Language contact and grammatical change*. Cambridge: Cambridge University Press.
- Holman, Eric W. & Walker, Robert & Rama, Taraka & Wichmann, Søren. 2011. Correlates of Reticulation in Linguistic Phylogenies. *Language Dynamics and Change* 1(2). 205–240. <https://doi.org/10.1163/221058212X648072>.
- Holton, Gary. 2004. *Report on Recent Linguistic Fieldwork on Pantar Island, Eastern Indonesia*. https://scholarworks.alaska.edu/xmlui/bitstream/handle/11122/6808/holton-2004-pantar_report.pdf.
- Holton, Gary. 2014. Western Pantar. In Antoinette Schapper (ed.), *Papuan languages of Timor, Alor and Pantar: sketch grammars*, vol. 1, 23–96. Berlin: Mouton de Gruyter.
- Holton, Gary & Klamer, Marian. 2017. The Papuan languages of East Nusantara and the Bird's Head. In Bill Palmer (ed.), *The Languages and Linguistics of the New Guinea Area*. 569–640. Berlin/New York: Mouton de Gruyter.
- Holton, Gary & Klamer, Marian & Kratochvíl, František & Robinson, Laura C. & Schapper, Antoinette. 2012. The historical relations of the Papuan languages of Alor and Pantar. *Oceanic Linguistics* 51(1). 86–122.
- Holton, Gary & Robinson, Laura C. 2014. The linguistic position of the Timor-Alor-Pantar languages. In Marian Klamer (ed.), *The Alor-Pantar languages: history and typology*, 155–198. Berlin: Language Science Press.
- Holton, Gary & Robinson, Laura C. 2017. The linguistic position of the Timor-Alor-Pantar languages. In Marian Klamer (ed.), *The Alor-Pantar languages: history and typology (2nd ed.)* Second edition., 155–198. Berlin: Language Science Press.
- Huber, Juliette. 2011. *A grammar of Makalero : a Papuan language of East Timor*. Utrecht: LOT Publications: Leiden University dissertation.
- Huber, Juliette. 2017. Makalero and Makasae. In Antoinette Schapper (ed.), *The Papuan languages of Timor, Alor, and Pantar: volume 2*, 268–351. Berlin: Mouton de Gruyter.
- Jäger, Gerhard & List, Johann-Mattis. 2018. Using ancestral state reconstruction methods for onomasiological reconstruction in multilingual word lists. *Language Dynamics and Change* 8(1). 22–54. <https://doi.org/10.1163/22105832-00801002>.
- Jäger, Gerhard & List, Johann-Mattis & Sofroniev, Pavel. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: volume 1, Long Papers*, vol. 1, 1205–1216.
- Jäger, Gerhard & Sofroniev, Pavel. 2016. Automatic cognate classification with a Support Vector Machine. In *Proceedings of the 13th Conference on Natural Language Processing*, vol. 16, 128–134.

- Kaiping, Gereon A. & Edwards, Owen & Klamer, Marian (eds.). 2019a. *LexiRumah 2.1.2*. <https://doi.org/10.5281/zenodo.2540954>.
- Kaiping, Gereon A. & Edwards, Owen & Klamer, Marian. 2019b. *LexiRumah 3.0.0*. Zenodo. <https://doi.org/10.5281/zenodo.3537977>.
- Kaiping, Gereon A. & Klamer, Marian. 2018. LexiRumah: an online lexical database of the Lesser Sunda Islands. *PLOS ONE* 13(10). e0205250. <https://doi.org/10.1371/journal.pone.0205250>.
- Kaiping, Gereon A. & Klamer, Marian. 2019. Subgrouping the Timor-Alor-Pantar languages using systematic Bayesian inference. <https://doi.org/10.31235/osf.io/9s5hj>.
- Kaiping, Gereon A. & Steiger, Melvin. 2021. *Lexedata*. In collab. with Natalia Chousou-Polydouri. Version v1.0.0-alpha1. Zürich: Universität Zürich. <https://github.com/Anaphory/lexedata/releases/tag/v1.0.0-alpha1>.
- Kalyan, Siva & François, Alexandre. 2018. Freeing the Comparative Method from the Tree Model: a Framework for Historical Glottometry. In Ritsuko Kikusawa & Lawrence A. Reid (eds.), *Let's Talk about Trees: genetic Relationships of Languages and Their Phylogenic Representation* (SENRI ETHNOLOGICAL STUDIES 98), 59–89. National Museum of Ethnology.
- Kelly, Luke J. & Nicholls, Geoff K. 2016. *Lateral Transfer in Stochastic Dollo Models*. <http://arxiv.org/abs/1601.07931>.
- Keraf, Gregorius. 1978. *Morfologi dialek Lamalera*. Jakarta: Universitas Indonesia Jakarta dissertation. 610 pp.
- Klamer, Marian. 2011. *A short grammar of Alorese (Austronesian)*. Munich: Lincom Europe.
- Klamer, Marian. 2017. The Alor-Pantar languages: linguistic context, history and typology. In Marian Klamer (ed.), *The Alor-Pantar languages: history and typology*, 2nd edn. (Studies in Diversity Linguistics 3), 1. Berlin: Language Science Press.
- Klamer, Marian. 2018. Typology and grammaticalization in the Papuan languages of Timor, Alor, and Pantar. In Heiko Narrog & Bernd Heine (eds.), *Grammaticalization from a Typological Perspective*, 235–262. Oxford: Oxford University Press.
- Klamer, Marian. to appear. Traces of pre-modern contacts between Timor-Alor-Pantar and Austronesian speakers. In Marian Klamer & Francesca Moro (eds.), *Traces of contact in the lexicon: austronesian and Papuan studies*. Leiden: Brill.
- Klamer, Marian & Sir, Amos. 2018. Bahasa Sar. In *Workshop on Documenting Minority Languages in Nusa Tenggara Timur Indonesia*, 97–103. Kupang, Indonesia: Universitas Kristen Artha Wacana. https://vici.marianklamer.org/uploads/1/2/4/7/124768088/kupang_workshop_booklet_smaller.pdf (16 December, 2019).
- Kolipakam, Vishnupriya & Jordan, Fiona M. & Dunn, Michael & Greenhill, Simon J. & Bouckaert, Remco & Gray, Russell D. & Verkerk, Annemarie. 2018. A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science* 5(3). 171504. <https://doi.org/10.1098/rsos.171504>.
- Kratochvíl, František. 2007. *A grammar of Abui: a Papuan language of Alor*. Utrecht: LOT. 142 pp.
- Kratochvíl, František. 2014. Sawila. In Antoinette Schapper (ed.), *Papuan Languages of Timor, Alor and Pantar: sketch Grammars*, vol. 1, 351–438. Berlin: De Gruyter Mouton.
- Lansing, John Stephen & Murray, P. Cox & De Vet, Therese A. & Downey, Sean S. & Hallmark, Brian & Sudoyo, Herawati. 2011. An ongoing Austronesian expansion in island Southeast Asia. *Journal of Anthropological Archaeology* 30(3). 262–272. <https://doi.org/10.1016/j.jaa.2011.06.004>.
- Lartillot, Nicolas & Philippe, Hervé. 2006. Computing Bayes factors using thermodynamic integration. *Systematic Biology* 55(2). 195–207. <https://doi.org/10.10635150500433722>.
- Lee, Sean & Hasegawa, Toshikazu. 2011. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. *Proceedings of the Royal Society of London B: Biological Sciences*. rspb20110518. <https://doi.org/10.1098/rspb.2011.0518>.
- List, Johann-Mattis. 2012a. LexStat: automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH (EACL 2012)*, 117–125. Stroudsburg, PA, USA: Association for Computational

- Linguistics. <http://dl.acm.org/citation.cfm?id=2388655.2388671> (12 August, 2015).
- List, Johann-Mattis. 2012b. SCA: phonetic Alignment based on sound classes. In *New Directions in Logic, Language and Computation*, 32–51. Springer.
- List, Johann-Mattis. 2014. Investigating the impact of sample size on cognate detection. *Journal of Language Relationship • Вопросы языкового родства* 11. 91–101. <https://cyberleninka.ru/article/n/investigating-the-impact-of-sample-size-on-cognate-detection-1>.
- List, Johann-Mattis. 2016. Beyond cognacy: historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution* 1(2). 119–136. <https://doi.org/10.1093/jole/lzw006>.
- List, Johann-Mattis & Cysouw, Michael & Forkel, Robert (eds.). 2016. *Concepticon*. Jena: Max Planck Institute for the Science of Human History. <https://doi.org/10.5281/zenodo.51259>.
- List, Johann-Mattis & Greenhill, Simon & Anderson, Cormac & Mayer, Thomas & Tresoldi, Tiago & Forkel, Robert (eds.). 2018. *CLICS: database of Cross-Linguistic Colexifications*. Jena: MPI for the Science of Human History. <https://clics.clld.org/>.
- List, Johann-Mattis & Greenhill, Simon J & Tresoldi, Tiago & Forkel, Robert. 2018. *LingPy. A Python Library for Quantitative Tasks in Historical Linguistics. Version 2.6.4*. Zenodo. <https://doi.org/10.5281/zenodo.1544172>.
- List, Johann-Mattis & Greenhill, Simon J. & Gray, Russell D. 2017. The Potential of Automatic Word Comparison for Historical Linguistics. *PLOS ONE* 12(1). e0170046. <https://doi.org/10.1371/journal.pone.0170046>.
- List, Johann-Mattis & Shijulal, Nelson-Sathi & Martin, William & Geisler, Hans. 2014. Using phylogenetic networks to model Chinese dialect history. *Language Dynamics and Change* 4(2). 222–252. <http://booksandjournals.brillonline.com/content/journals/10.1163/22105832-00402008> (13 August, 2015).
- Maurits, Luke & Forkel, Robert & Kaiping, Gereon A. & Atkinson, Quentin D. 2017. BEASTling: a software tool for linguistic phylogenetics using BEAST 2. *PLOS ONE* 12(8). e0180908. <https://doi.org/10.1371/journal.pone.0180908>.
- Michael, Lev & Chousou-Polydouri, Natalia & Bartolomei, Keith & Donnelly, Erin & Meira, Sérgio & Wauters, Vivian & O’Hagan, Zachary. 2015. A Bayesian Phylogenetic Classification of Tupí-Guaraní. *LIAMES: Línguas Indígenas Americanas* 15(2). 193–221. <https://doi.org/10.20396/liames.v15i2.8642301>.
- Müller, Nicola F. & Bouckaert, Remco. 2019. Coupled MCMC in BEAST 2. *bioRxiv*. 603514. <https://doi.org/10.1101/603514>.
- Pawley, Andrew. 2002. The Austronesian dispersal: people, language, technology. In Peter Bellwood & Colin Renfrew (eds.), *Examining the language/farming dispersal hypothesis*. 251–271. Cambridge: McDonald Institute for Archaeological Research.
- Pawley, Andrew. 2005. The chequered career of the Trans New Guinea hypothesis: recent research and its implications. In *Papuan pasts: cultural, linguistic and biological histories of Papuan-speaking peoples*. Pacific Linguistics. <https://openresearch-repository.anu.edu.au/handle/1885/84324> (6 October, 2017).
- Rama, Taraka. 2018. Three tree priors and five datasets: a study of Indo-European phylogenetics. *Language Dynamics and Change* 8(2). 182–218. <https://doi.org/10.1163/22105832-00802005>.
- Rama, Taraka & List, Johann-Mattis & Wahle, Johannes & Jäger, Gerhard. 2018. *Are Automatic Methods for Cognate Detection Good Enough for Phylogenetic Reconstruction in Historical Linguistics?* <http://arxiv.org/abs/1804.05416> (13 July, 2018).
- Rama, Taraka & Wahle, Johannes & Sofroniev, Pavel & Jäger, Gerhard. 2017. *Fast and Unsupervised Methods for Multilingual Cognate Clustering*. <http://arxiv.org/abs/1702.04938> (13 July, 2018).
- Rambaut, Andrew & Drummond, Alexei J. & Xie, Dong & Baele, Guy & Suchard, Marc A. 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology* 67(5). 901–904. <https://doi.org/10.1093/sysbio/syy032>.
- Rambaut, Andrew & Xie, Wangang & Baele, Guy & Suchard, Marc A. & Drummond, Alexei J. 2020.

- Tracer: Posterior Summarisation in Bayesian Phylogenetics*. Version 1.7.1. <https://github.com/beast-dev/tracer> (14 August, 2020).
- Robinson, Laura C. & Holton, Gary. 2012a. Internal Classification of the Alor-Pantar Language Family Using Computational Methods Applied to the Lexicon. *Language Dynamics and Change* 2(2). 123–149. <https://doi.org/10.1163/22105832-20120201>.
- Robinson, Laura C. & Holton, Gary. 2012b. Re-assessing the wider genetic affiliations of the Timor-Alor-Pantar languages. In <https://scholarworks.alaska.edu/handle/11122/1053> (6 October, 2017).
- Ross, Malcolm. 2005. Pronouns as a preliminary diagnostic for grouping Papuan languages. In Andrew K. Pawley & Robert Attenborough & Jack Golson & Robin Hide (eds.), *Papuan pasts: cultural, linguistic and biological histories of Papuan-speaking peoples* (Pacific Linguistics 572), 15–65. Canberra: Research School of Pacific & Asian Studies, Australian National University. http://www.academia.edu/download/31189925/Ross_Papuan_Pasts.pdf.
- Rosvall, Martin & Bergstrom, Carl T. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105(4). 1118–1123. <https://doi.org/10.1073/pnas.0706851105>.
- Sapir, E. 1916. *Time Perspective in Aboriginal American Culture: A Study in Method*. 90. <https://doi.org/10.4095/103486>.
- Schapper, Antoinette & Hendery, Rachel. 2014. Wersing. In Antoinette Schapper (ed.), *Papuan languages of Timor, Alor and Pantar: sketch grammars*, vol. 1, 439–504. Berlin: Mouton de Gruyter.
- Schapper, Antoinette & Huber, Juliette & van Engelenhoven, Aone. 2017. The relatedness of Timor-Kisar and Alor-Pantar languages: a preliminary demonstration. In Marian Klamer (ed.), *The Alor-Pantar languages: history and typology*, 2nd edn., vol. 3, 99–154. Berlin: Language Science Press. <https://epub.uni-regensburg.de/31377/> (7 November, 2017).
- Schapper, Antoinette & Huber, Juliette & van Engelenhoven, Aone. 2012. The historical relation of the Papuan languages of Timor and Kisar. *Language and linguistics in Melanesia*. 192–240.
- Schapper, Antoinette & Klamer, Marian. 2017. Numeral systems in the Alor-Pantar languages. In Marian Klamer (ed.), *The Alor-Pantar languages: history and typology* (2nd ed.) Second edition, 277–328. Berlin: Language Science Press.
- Spriggs, Matthew. 2011. Archaeology and the Austronesian expansion: where are we now? *Antiquity* 85(328). 510–528. <https://doi.org/10.1017/S0003598X00067910>.
- Traugott, Elizabeth Closs & Dasher, Richard B. 2001. *Regularity in Semantic Change* (Cambridge Studies in Linguistics 96). Cambridge: Cambridge University Press. 363 pp. <https://doi.org/10.1017/CB09780511486500>.
- Vaughan, Timothy G & Welch, David & Drummond, Alexei J & Biggs, Patrick J & George, Tessy & French, Nigel P. 2017. Inferring Ancestral Recombination Graphs from Bacterial Genomic Data. *Genetics* 205(2). 857–870. <https://doi.org/10.1534/genetics.116.193425>.
- Wellfelt, Emilie. 2016. Historyscapes in Alor : approaching indigenous history in Eastern Indonesia. <http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-56533> (4 May, 2019).
- Wichmann, Søren & Brown, Cecil H. & Holman, Eric W. (eds.). 2016. *The ASJP Database*. Jena: Max Planck Institute for the Science of Human History. <http://asjp.cllld.org/> (6 October, 2017).
- Willemsen, Jeroen. 2020. Reta. In Antoinette Schapper (ed.), *The Papuan languages of Timor, Alor and Pantar*, vol. 3 (Pacific Linguistics 660), 187–266. Berlin: De Gruyter Mouton. <https://www.degruyter.com/document/doi/10.1515/9781501511158/html>.
- Windschuttel, Glenn & Shiohara, Asako. 2017. Kui. In Antoinette Schapper (ed.), *The Papuan Languages of Timor, Alor and Pantar: volume 2*, 109–183. Berlin / New York: De Gruyter Mouton.
- Wurm, Stephen A. & Voorhoeve, C. L. & McElhanon, Kenneth A. 1975. The trans-New Guinea phylum in general. In Stephen A. Wurm (ed.), *New Guinea area languages and language study*, vol. 1 (Pacific Linguistics C 38), 299–322. Canberra: Research School of Pacific & Asian Studies, Australian National University.
- Xie, Wangang & Lewis, Paul O. & Fan, Yu & Kuo, Lynn & Chen, Ming-Hui. 2011. Improving marginal likelihood estimation for Bayesian

- phylogenetic model selection. *Systematic Biology* 60(2). 150–160. <https://doi.org/10.1093/sysbio/syq085>.
- Xu, Shuhua & Pugach, Irina & Stoneking, Mark & Kayser, Manfred & Jin, Li & The HUGO & Pan-Asian Consortium. 2012. Genetic dating indicates that the asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. *PNAS* 109(12). 4574–4579.
- Yang, Ziheng & Rannala, Bruce. 2005. Branch-Length Prior Influences Bayesian Posterior Probability of Phylogeny. *Systematic Biology* 54(3). 455–470.
- Zalizniak, Anna A. 2018. *Database of Semantic Shifts in Languages of the World*. DatSemShift. <http://datsemshift.ru/shifts> (11 February, 2019).
- Zhang, Chi. 2018. *Speciesnetwork: Bayesian Inference of Species Networks from Multilocus Sequence Data (a Package for BEAST2)*. <https://github.com/zhangchicool/speciesnetwork> (10 July, 2018).

Appendix A Comparison of our results with Fig. 3

The tree of Fig. 3 without branch lengths and with our language names.

The summary tree of S (Fig. 6) restricted to the languages from Fig. 3.

Appendix B Details on Cognate Coding

Automatic cognate detection algorithms calculate a score of formal similarity or “cognate-ness” for each pair of word forms that have the same meaning. The currently most established type of scoring functions converts the phonetic or phonemic transcriptions of forms into broader sound classes, aligns them and calculates a distance score based on that alignment. (Promising new scoring functions have been suggested (Jäger & Sofroniev 2016, Jäger et al. 2017), but they are not yet practically available.) Here we describe Lexstat (List 2012a) and OnlinePMI Rama et al. (2017), which have both shown good performance (List et al. 2017, Rama et al. 2017).

Lexstat and is easily available in the LingPy software package for historical linguistics (List & Greenhill & Tresoldi, et al. 2018). It requires at least 100 shared concepts for each pair of

Figure 8: Network of cognate-ness for the forms meaning ‘sacrifice’, calculated using LexStat with the SCA sound class model. Thicker, darker lines represent a smaller inferred distance between forms. Different colours of the forms indicate different inferred cognate sets, see Table 7

languages to work well (List 2014).

In order to reduce the complexity and therefore the number of parameters it needs to estimate, LexStat does not operate directly on the IPA-transcribed forms, but instead uses transcriptions that follow a sound class model that groups similar sounds together into a small number of classes. The SCA (List 2012b) sound class model has been established as useful for automatic cognate detection, but the ASJP sound classes from the Automatic Similarity Judgement Program database (Wichmann et al. 2016) might be a useful alternative.

LexStat (List 2012a) first groups together very similar forms that express the same concept across different languages. Then it calculates systematic sound correspondences by using these very similar forms to bootstrap a table of scores for systematic sound correspondence between each pair of languages. To account for random similarities, these scores are normalized against random correspondences obtained from comparing the words for random concepts in the word lists. The cognate-ness distance of two forms can then be thought of as the effective number of changes (discounting systematic sound changes and over-counting unexpected correspondences, as derived from the bootstrap step) to transform one form into the other, normalized by the chance for random similarities and the length of the forms.

Where LexStat uses three passes over the whole dataset to find systematic sound correspondences in the data, the OnlinePMI method updates its similarity scores continuously (“online” in the machine learning jargon) by going through the dataset in small batches and adapting the scores used for future alignments after each such batch. It uses a statistical measure for the co-occurrence of sound segments known as ‘Pointwise Mutual Information’ to generate these scores.

Computing these similarity scores using either method gives rise to a network, in which the

Language	Form	$\theta = 0.55$	$\theta = 0.35$	$\theta = 0.75$
Abui-Takalelang	dɛi.'puŋ	1	1	1
Abui-Takalelang	dɛi.'pü	1	1	1
Abui-Ulaga	mini	2	2	2
Blagar-Manatang	pemberiaŋ	3	3	3
Blagar-Pura	ameaŋ	4	4	4
Bunak-Bobonaro	kasaʔ gume	5	5	5
Bunak-Bobonaro	kasaʔ goro	5	6	5
Bunak-Suai	kasaʔ gume	5	5	5
Bunak-Suai	kasaʔ goro	5	6	5
Fataluku	ku'lupa	6	7	6
Kaera-Abangiwang	goŋ 'waŋ samuŋ	7	8	7
Klon-Hopter	go'luk go'həl	8	9	5
Kula-Lantoka	'gwajam 'pajam	9	10	8
Makasae	kulu'ba:	6	7	6
Reta-Hula	to:-mamo:-liŋ	10	11	9
Reta-Hula	ma genəŋ	11	12	10
Reta-Hula	sadaka	12	13	11
Reta-Pura	sadaka	12	13	11
Sawila	pəmbə'rian	13	14	3
Teiwa-Lebang	'derma	14	15	12
Teiwa-Lebang	'sembahan	15	16	13
Wersing-Maritaing	kərban	16	17	6
Indonesian	korban	16	18	6
Tetun-Suai	hasae	17	19	14

Table 7: Results of automatic cognate detection on the forms meaning ‘sacrifice’, for the networks shown in Figs. 8, 9a and 9b

connections between the forms (“edges” in the graph literature) are “shorter” (representing a stronger connection) for similar forms and “longer” (representing weaker connection) for different forms. Figure 8 shows the resulting network for the forms meaning ‘sacrifice’. Note that the length of the connections in the figure is only an approximation of the calculated distances, made necessary by embedding the graph in a 2-dimensional plane. The actual value for the cognate-ness distance is instead reflected in the colour and thickness of the lines.

A second part of the algorithm then tries to find clusters in this weighted network. That is, it tries to find non-overlapping sets of forms that all have a short distance to each other, but have a long cognate-ness distance to words outside the set.

We follow List et al. and cluster forms into cognate classes using Infomap (Rosvall & Bergstrom 2008) on an unweighted graph connecting pairs of forms that have a ‘cognate-ness distance’ less than a threshold θ : In general, we use the threshold of $\theta = 0.55$ that performed best in List et al.’s analysis. Infomap is a graph community detection algorithm which has been shown to outperform other methods for cognate class clustering (List et al. 2017). The resulting cognate classes can be represented in tabular format like Table 7.

Figure 9: Networks of cognate-ness for the forms meaning ‘sacrifice’, calculated using LexStat with the SCA sound class model, clustered using infomap with different thresholds θ . The forms are the same as those in Fig. 8.

Appendix C Posterior tree samples of rejected model choices

C.1 Different data coding

Densitree visualizations of the posterior tree samples from the best model, S , under different ways to automatically cognate-code the data. The order of lects is the same as in Fig. 6, and top to bottom roughly follows the locations of languages from east (Timor) to west (Pantar).

D1: Using ASJP sound classes instead of SCA sound classes

D2: More splitting cognate coding, $\theta = 0.35$

D3: More lumping cognate coding $\theta = 0.75$

D4: Using the Online-PMI cognate coder instead of LexStat

C.2 Lower model likelihood

Densitree visualizations of the posterior tree samples from the convergent models considered in the model selection, which were rejected due to lower model likelihood. The order of lects is the same as in Fig. 6, and top to bottom roughly follows the locations of languages from east (Timor) to west (Pantar).

M: binary covarion, strict clock, birth-death tree prior

T: general binary model, strict clock, uniform tree prior

MT: binary covarion, strict clock, uniform tree prior

Appendix D Innovations supporting the East Alor early split

The following table lists the cognate sets of proto-TAP *mi 'be in, at', proto-Nuclear Alor Pantar *om mi '(be) inside', and reconstructions of intermediate forms. Sources for the reflexes

of proto-TAP *mi ‘be in, at’ etc. are Kaiping et al. (2019a), Klamer (2018: 242–246), and specifically for Adang: Haan (2001: 145), Makalero and Makasai: Huber (2011, 2017), Kaiping et al. (2019a), Klon-Bring: Baird (2008), Kafoa: Baird (2017), Kui: Windschuttel & Shiohara (2017), Reta: Willemsen (2020), Abui-Takalelang: Kratochvíl (2007), Sawila: Kratochvíl (2014), Wersing: Schapper & Hendery (2014), W Pantar: Holton (2014).

D.1 *mi ‘inside’

Language	Forms reflecting *mi	Forms reflecting *om mi
Bunak	mil ‘inside’	
Makasai	mi- ‘Applicative’, (mutuʔu)	
Makalero	mi- ‘Applicative’, (mutuʔ)	
Oirata	(mutʔwainani)	
Fataluku	(mutʔune)	
Wersing	mi- ‘Applicative’, mira ‘inside’, min ‘be at’	
Kula	mɪ ‘be located’, mərə ‘inside’	
Sawila	ming ‘be located’, mirea ‘inside’	
Kamang-Atoitaa	mi- ‘Applicative’, mi ‘inside’	
Kui	mi- ‘Applicative’, mi ‘be in, at’, mare ‘inside’	
Kafoa	mi ‘be at’	-ommi
Abui-Takalelang	mi ‘be in’	-o:mi
Abui-Ulaga	mia ‘be in’	-oni
Klon-Bring	mi ‘be at; to place’, mi ‘Locative’, mi- ‘Applicative’	-omi
Adang	mi ‘be in, at’, mi ‘in, at’	ʔommi
Blagar-Pura	=mi, mi ‘in; to; into; from’	-omi
Reta	mi ‘be in’	-o:mi
Kaera-Abangiwang	ming ‘be in, at’, mi ‘in; at; to; with’	-ommi
Teiwa	meʔ ‘be in’	-ommeʔ
WPantar	me ‘LOC’, migang ‘to set’	-ume

D.2 *hada ‘fire’

Language	Forms reflecting *hada ‘fire’
Bunak	hoto hoto
Makasai	ata
Oirata	aʔa
Fataluku	atɛ
Wersing	ada
Kula	ada
Sawila	ada
Kamang-Atoitaa	atɪ
Kiraman	ar
Kui	ar
Kafoa	ara
Abui	ara
Klon-Bring	əda
Klon-Hopter	ada (wer)
Adang	(awai, aʔfai)
Kabola	(awal)
Hamap	(afail)
Blagar-Pura	ad
Reta	ad
Kaera-Abangiwang	ad (wasing)
Teiwa	har
WPantar	ra

D.3 *hapi ‘fish’

Language	Forms reflecting *habi 'fish'
Bunak	(ikan)
Makasae	afi
Oirata	ahi
Fataluku	api
Wersing	api
Kula	api, apu
Sawila	api
Kamang-Atoitaa	api
Kiraman	ɛb
Kui	eb
Kafoa-Probur	afui
Abui	afu
Klon-Bring	əbi
Klon-Hopter	ʔəbi:
Adang-Lawahing	a:b
Adang-Otvai	hab
Hamap-Moru	ʔab
Kabola-Monbang	hab
Blagar (all dialects)	aba:b
Reta	a:b
Reta	ʔa:b
Sar	haf
Teiwa-Adiabang	haf
Teiwa-Lebang	ħaf
Teiwa-Nule	haf
Deing	haf
WesternPantar-Tubbe	hap (keʔe)