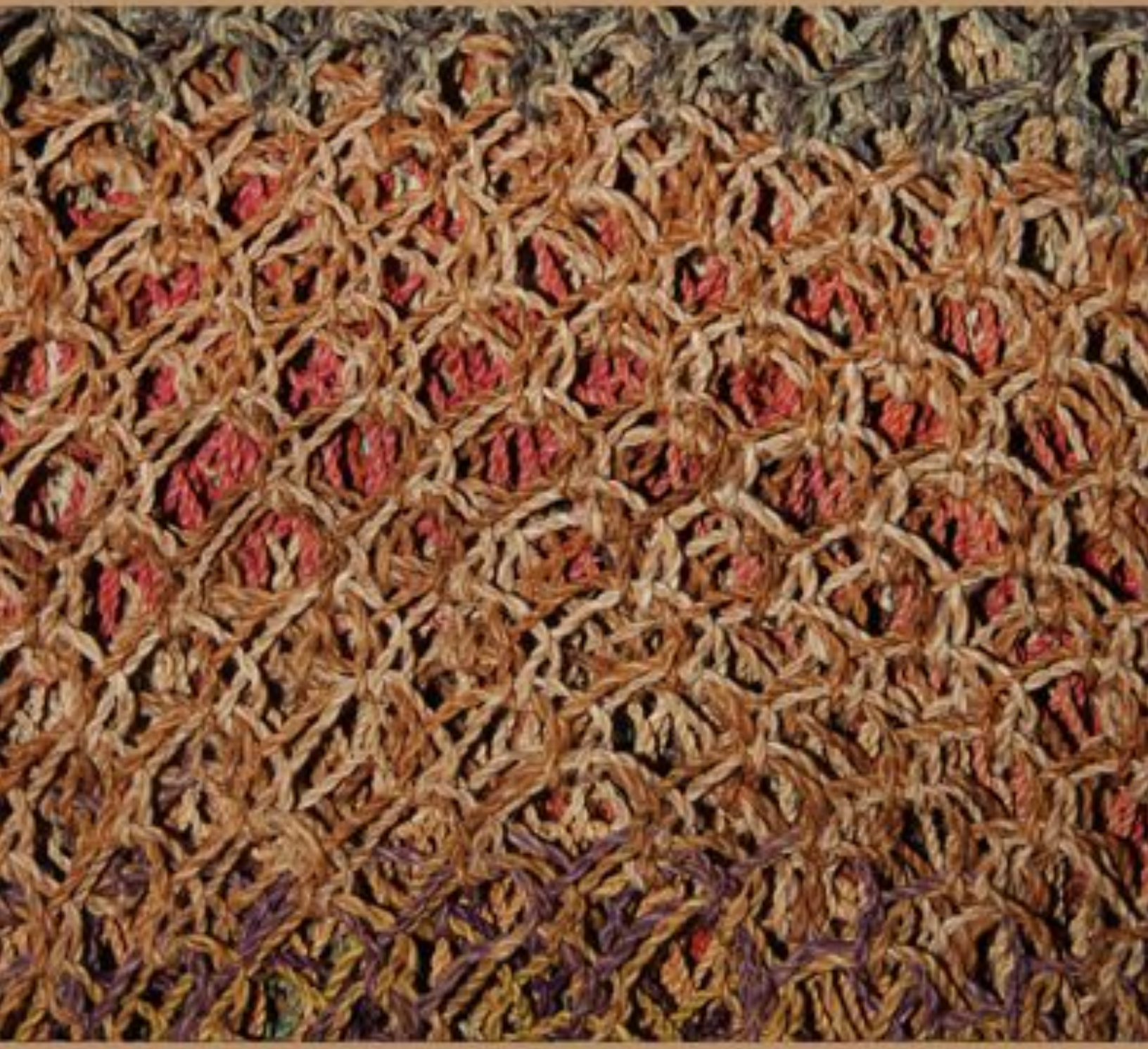
Language Documentation & Conservation Special Publication No. 5

Melanesian Languages on the Edge of Asia: Challenges for the 21st Century



edited by
Nicholas Evans and Marian Klamer

Melanesian Languages on the Edge of Asia: Challenges for the 21st Century

edited by
Nicholas Evans and Marian Klamer

 Language Documentation & Conservation Special Publication No. 5

Published as a Special Publication of Language Documentation & Conservation

Language Documentation & Conservation
Department of Linguistics, UHM
Moore Hall 569
1890 East-West Road
Honolulu, Hawai'i 96822
USA

http://nflrc.hawaii.edu/ldc

University of Hawai'i Press
2840 Kolowalu Street
Honolulu, Hawai'i
96822-1888
USA

Cover design by Susan Ford incorporating a photograph by Darja Hoenigman

Library of Congress Cataloging in Publication data
ISBN 978-0-9856211-2-4

http://hdl.handle.net/10125/4557

Contents

Contributors

I Wayan Arka is affiliated with the Australian National University (as a Fellow in Linguistics at School of Culture, History and Language, College of Asia and the Pacific) and Udayana University Bali (English Department and Graduate Program in Linguistics). His interests are in descriptive, theoretical and typological aspects of Austronesian and Papuan languages of Indonesia. Wayan is currently working on a number of projects: NSF-funded research on voice in the Austronesian languages of eastern Indonesia (2008-2011), ARC-funded projects for the development of computational grammar for Indonesian (2008-2011) and the Languages of Southern New Guinea (2011-2014).

Linda Barwick is Associate Professor in the School of Letters, Art and Media at the University of Sydney, and Director of the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). She trained in Italian language and dialectology with Antonio Comin at Adelaide's Flinders University, and in ethnomusicology with the late Catherine Ellis at the University of Adelaide. Since then she has undertaken ethnomusicological fieldwork in Italy, Australia and the Philippines, and has published widely on Australian Indigenous Music as well as continuing her engagement with Italian traditional music.

Michael Dunn is an evolutionary linguist with a background in linguistic typology and language description. His current research takes a quantitative, phylogenetic approach to language change and linguistic diversity. Recent projects have addressed questions of coevolution of typological parameters, as well as the ecological and social factors influencing language change. He has also worked on the classification and prehistory of Papuan languages, and the phylogeography of Indo-European and Aslian language. He leads the Max Planck Research Group "Evolutionary Processes in Language and Culture" at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands.

Nicholas Evans is Distinguished Professor of Linguistics in the College of Asia/Pacific, Australian National University. He has carried out wide-ranging fieldwork on traditional languages of northern Australia and southern Papua New Guinea. The driving interest of his work is the interplay between documenting endangered languages and the many scientific and humanistic questions they can help us answer. In addition to grammars of two Aboriginal languages, Kayardild and Bininj Gun-wok, dictionaries of Dalabon and Kayardild, edited collections on a number of linguistic topics, and over 120 scientific papers, he recently published the widely-acclaimed crossover book *Dying Words: Endangered Languages and What They Have to Tell Us* which sets out a broad program for engaging with the world's dwindling linguistic diversity. He has also worked as a linguist, interpreter and anthropologist in two Native Title claims in northern Australia, and as a promotor of Aboriginal art by the Bentinck Island women's artists.

Mats Exter studied General Linguistics, Historical-Comparative Linguistics, Phonetics and Finnish Studies at the University of Cologne. He has held post-doctoral research and teaching positions at the Universities of Bonn and Düsseldorf. His work focuses on language description and documentation, experimental phonetics, laboratory phonology, and morphosyntactic typology. He has conducted fieldwork on Wogeo, an Austronesian language of Papua New Guinea, where he has done descriptive research on the phonological and morphosyntactic structure as well as producing a collection of traditional texts. More recently, he has conducted fieldwork on N|uu, a Tuu (formerly 'Southern Khoisan') language of South Africa, focusing on the phonetic and phonological structure of the language.

Michael C. Frank is Assistant Professor of Psychology at Stanford University. He received his PhD from the Department of Brain and Cognitive Sciences at MIT in 2010 and now heads the Language and Cognition Lab, which uses experimental, observational, and computational methods to study language acquisition and language use. He is broadly interested in the reciprocal interactions of language and cognition: both how languages affect the thoughts of their users and how the structure of cognition (especially social cognition) facilitates the acquisition of language in infants and children.

Harald Hammarström studies linguistics and computer science at Uppsala University (Sweden) and went on to do a PhD in computational linguistics at Chalmers University (Sweden). His interests are linguistic typology (especially numeral systems), Papuan languages, language classification and computational techniques for modeling the diversity of human languages. He is currently working as a PostDoc at the MPI EVA (Germany) and Radboud University (the Netherlands) on the documentation of the Papuan language Mor and on areal linguistics in South America.

Darja Hoenigman is a PhD candidate in Anthropology at The Australian National University, working among the Awiakay, a community of 300 people living in Kanjimei village in East Sepik Province of Papua New Guinea. In her current project she is investigating socio-cultural continuity and change in Kanjimei and its relation to linguistic registers. In studying these speech varieties and their relation to the overall social scene, she brings together linguistic anthropology and ethnographic filmmaking.

Marian Klamer teaches at Leiden University and has done primary fieldwork on a dozen Austronesian and Papuan languages in Indonesia over the last two decades. Her research centres on language description and documentation, typology, and historical and contact-induced language change. Her publications include A grammar of Kambera (1998), A grammar of Teiwa (2010), A short grammar of Alorese (2011), and over 50 articles on a variety of topics. Klamer has coordinated numerous research projects on languages of Indonesia, including the NOW-VIDI project 'Linguistic variation in Eastern Indonesia' (2002–2007) and the EuroBABEL project 'Alor Pantar languages: Origins and theoretical impact' (2009–2012), funded by the European Science Foundation

Sebastian Nordhoff is a postdoctoral researcher at the Max Planck Institute for Evolutionary Anthropology in Leipzig. He specializes in language contact and language change and the interface of language description and documentation on the one hand and electronic publication on the other. He is a member of the working group on Open Data in Linguistics of the Open Knowledge Foundation, where he works on integrating typological data into the Linguistic Linked Open Data Cloud.

Ger Reesink studied psychology at the University of Utrecht, after which he spent 15 years in Papua New Guinea under the auspices of SIL. Finishing his affiliation with SIL, he spent more than 15 years at Leiden University, mostly doing research on the languages of the Bird's Head of Papua province, Indonesia. Since 2002 he has been a postdoctoral researcher at the Radboud University and the Max Planck Institute for Psycholinguistics at Nijmegen, involved in typological research of Papuan and Austronesian languages in order to trace the ancient history of genealogical and contact relations.

Nick Thieberger is an ARC QEII Fellow at the University of Melbourne. He recorded Paakantyi (NSW) speakers in the early 1980s and then worked with Warnman speakers (Western Australia) when he was setting up the Wangka Maya language centre in Port Hedland. He built the Aboriginal Studies Electronic Data Archive (ASEDA) at AIATSIS in the early 1990s and then was at the Vanuatu Cultural Centre from 1994–1997. He wrote a grammar of South Efate, a language from central Vanuatu, which was the first grammar to cite a digital corpus of recordings in all example sentences and texts. In 2003 he helped establish the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). He taught in the Department of Linguistics at the University of Hawai'i (2008–2010). He is a co-director of the Resource Network for Linguistic Diversity (RNLD) and the editor of the journal Language Documentation & Conservation. He is developing methods for the creation of reusable data from fieldwork on previously unrecorded languages and training researchers in those methods.

1

Language Documentation & Conservation Special Publication No. 5 (December 2012)
Melanesian Languages on the Edge of Asia: Challenges for the 21st Century,
ed. by Nicholas Evans and Marian Klamer, pp. 1–12
http://nflrc.hawaii.edu/ldc/sp05/
http://hdl.handle.net/10125/4558

Introduction: Linguistic challenges of the Papuan region

Nicholas Evans
The Australian National University

Marian Klamer
Universiteit Leiden

The region where Papuan languages are spoken – centred on the Island of New Guinea, with extensions westward into Timor and the islands of eastern Indonesia, and eastward into the Solomon Islands – is at the same time the most linguistically diverse zone of the planet and the part of the logosphere.[1] It packs around 20% of the world's languages into less than 1% of its surface area and less than 0.1% of its population. The absolute level of linguistic diversity – whether measured in sheer numbers of languages, or in terms of 'maximal clades' of unrelatable units – is comparable to the whole of Eurasia.

Getting the right term to describe the region of interest in this collection is a famously difficult problem. Melanesia is a little too broad – extending out to Fiji, Vanuatu and New Caledonia to the east, a little beyond the scope of the present collection, and on the other

[1] We gratefully acknowledge the support of the various bodies which supported the original conference in Manokwari from 8-10 February, 2010, under the title 'Melanesian Languages on the Edge of Asia': the Australia-Netherlands Research Collaboration, the Australian National University (Department of Linguistics, College of Asia and the Pacific, and the Stephen and Helen Wurm Endowment), the Max Planck Institute for Evolutionary Anthropology, and the Universitas Negeri Papua (Unipa) and the Centre for Endangered Languages Documentation (CELD) for so generously hosting the conference. We would further like to thank the various colleagues who acted as referees for the present collection (in alphabetical order): Sander Adelaar, Wayan Arka, Johan van der Auwera, Matthew Baerman, Rene van den Berg, Sonia Cristofaro, Mary Dalrymple, Philippe Grangé, Simon Greenhill, Harald Hammarström, Andy Pawley, Ger Reesink, Malcolm Ross, Alan Rumsey, Ruth Singer, Lourens de Vries, as well as to Susan Ford and Aung Si for their editorial assistance. Evans would further like to thank the Alexander von Humboldt Stiftung, through an Anneliese-Maier Forschungspreis, for financial support which assisted in some of the final production of this collection.

hand not generally including the Lesser Sunda islands in the Indonesian archipelago.

Nor are definitions in terms of language families easy to make cleanly. The Austronesian languages have wrapped New Guinea and its surrounding islands in a three thousand year embrace that is still being played out in intimate language contact with all the other languages of the region. Some of the papers here concern either Austronesian languages with significant structural resemblances to non-Austronesian languages of the region – see Exter's paper on Wogeo – or various types of historical and typological interaction between Austronesian and non-Austronesian languages – see the papers by Reesink & Dunn and by Klamer.

For the non-Austronesian languages of Melanesia and its surrounds (excluding Australia), the collective name 'Papuan' has been widely used and we continue that practice here. This use, based on definition by exclusion, has hung on for want of a better term long after comparable terms like 'Palaeosiberian' have been abandoned, but includes upwards of forty distinct families and isolates. To get an idea of how distorting a term like this is, consider how unsatisfactory it would be to use a term like 'Eurasian' for the set of languages including Basque, Finnish, Georgian, Ingush, Chinese, Tamil, Cambodian, Japanese, Kurdish, Japanese, Hmong, Ket, Chukchi, Burushaski and all the other non-Indo-European languages of Eurasia (where we partition off Indo-European languages only, in the same way that we partition off Austronesian languages). Yet that is arguably the level of genetic and typological diversity which we face when confronted with the full range of Papuan languages. Despite these problems, we currently have no better term, so the reader is simply cautioned to keep all these caveats in mind each time the word 'Papuan' is used.

Our knowledge of this exuberant linguistic cornucopia lags behind what we know about any other region of the globe. It is likely that the linguist-to-language ratio is lower here than anywhere else, and it is certain that the relative level of language documentation is lower here than anywhere else (see Hammarström and Nordhoff's paper). The inchoate state of Papuan linguistic studies stems from many reasons. These include the recency of linguistic research in the area, the inaccessibility of many sites, the lack of relevant training organisations in the countries concerned, the fragmentation of research across national boundaries and across the academic vs missionary divide, and the general lack of prioritisation that large parts of the linguistic profession have until recently assigned to the documentation of linguistic diversity. We have put together this collection of papers as a sample of just some of the research questions, languages and approaches that currently seem particularly exciting, with the goal of raising interest in this fascinating part of the logosphere, and encouraging linguists from around the world to get involved in research in an area where there is so much just waiting to be discovered,.

The present collection grows out of a conference held at the Centre for Endangered Languages Documentation (CELD) at Universitas Negeri Papua (Unipa) in Manokwari, Indonesia, in February 2010, with the support of the Australian Netherlands Research Council, the Australian National University, and the Max Planck Institute for Evolutionary Anthropology in Leipzig, all of whom we thank for their financial assistance. It is not simply a conference proceedings, however – it represents merely a selection of papers from that conference, supplemented by an additional paper to fill in gaps we thought needed coverage.

We began this introduction by continuing the well-established tradition of stating, in

an approximate and rather unquantified way, that Melanesia, and in particular the island of New Guinea, contains both the greatest concentration of linguistic diversity anywhere on earth, and the lowest level of documentation. The first paper in this volume, *The languages of Melanesia: Quantifying the level of coverage* by Hammarström & Nordhoff, adds precision to this statement by presenting relevant figures from their LangDoc database. This database aims to give comprehensive global listing of all materials existing on all languages, along with an initial, approximate metric of degree of coverage. As the authors point out, there are many shortcomings to their metric. It is relatively unambitious: a language possessing a low-quality grammar of 160 pages and no lexicon or text collection would already be placed at the highest level—well short of the modern gold standard of a Boasian trilogy supplemented by a wide variety of annotated multimedia files—and there is no measure of quality of analysis. Despite these flaws, it has the great virtue of being operationalisable and applicable to all the world's languages in a relatively automatic way, and in their paper they outline their scheme in detail as well as comparing the level of coverage for Melanesia with the rest of the world.

First, regarding the total proportion of the world's languages spoken in Melanesia, their figures count 1347 languages that are 'Melanesian' in their sense (522 Austronesian, 825 non-Austronesian) of the sub-region of Oceania extending from the Arafura Sea in the west to Fiji in the east (see figure 1 in their chapter). This is just over 20% of the world total of 6496 (living) languages on their count (see their table 7), with so-called 'Papuan languages' then making 12.7% of the world's total. In absolute terms, the number of languages in Melanesia (1,347) is almost identical to those in the whole of Eurasia (1,465), these two being surpassed only by Africa (1,986).

Second, for their assessment of level of documentation, they lump together Papuan with all Austronesian, so their figures also include the rest of Indonesia, the Philippines, Malaysia, and so on. Drawing on these figures, they draw some striking conclusions. First, in absolute terms, Papua + Austronesian has the largest number of languages with only a wordlist to their documentation (i.e. the lowest level of documentation which they recognise). The comparison with Australia, another region where professional linguistic research is relatively recent, is salutary: over 42% of Australian languages have a grammar available, compare to half that number (20.48%) for Austronesian + Papuan. Second, in relative terms, Papua + Austronesian has the lowest proportion of languages with the highest rank of description (i.e. a grammar of 150 pages or more), the highest proportion with only a wordlist or less, and the lowest average level of documentation. Third, when Austronesian and non-Austronesian languages are compared within the above categories, the non-Austronesian languages have lower levels of documentation, making their overall documentation status even lower than that for Papua + Austronesian as a whole (see their Table 6).

In the years to come it is to be hoped that LangDoc will be extended to give more accurate metrics in a number of ways – something which will be aided to the extent that more linguists heed the authors' call to put their results in the public domain. But their paper already provides a very clear quantitative basis for our claim above that the Melanesian region – and particularly the Papuan languages within it – is far and away the most linguistically diverse part of the planet, and that conversely it suffers from the lowest level of language documentation found in any quarter of the earth. The combination of

these factors is what makes the study of Melanesian languages an enormous challenge.

Before leaving this paper, we note two important future developments. Firstly it will be crucial to link some form of comprehensive database like LangDoc to actual documents so that it is possible to inspect the actual materials listed there and gradually improve the qualitative ratings through the collective efforts of world scholarship. Secondly, it is desirable that the structure of the LangDoc database allows inspection of data at a number of different geographic levels. While their present article largely treated Melanesia as an undifferentiated whole, their discussion of one geographical variable (distance from coast) shows how more finely articulated geographical characterisations can be made – so that one can compile comparable reports for geographical regions like the Sepik, Bougainville, etc.

The staggering linguistic complexity of Melanesia creates special problems for attempts to classify languages into families and subgroups, especially for efforts that try to reduce the large number of independent maximal clades (over forty on any estimate) by grouping some of them together.

Under these circumstances, the languages of Melanesia have provided a particularly important testing-ground in recent years for new methods which aim to 'break the time barrier' of the classical comparative method, by drawing inferences from the signal in assemblages of typological traits rather than simply in the sound-meaning pairings of the lexicon and grammatical morphology. Though controversial and still subject to fierce critique (see references in Reesink and Dunn paper), it is likely that such methods as applied to Melanesia are here to stay, at the very least as a supplement to the comparative method. Indeed, the situation in Melanesia forces historical linguists to make a virtue of necessity by driving them to develop new methods.

The article by Reesink and Dunn gives an overview of these methods, focussing on the languages of Eastern Indonesia, spoken around the Bird's Head area. As in their other studies, a grave problem with the method is that resemblant signals can signal either shared phylogeny or areal convergence. In the central part of their paper, they consider the case of two Papuan languages of the Bird's Head, Hatam and Meyah, which consistently cluster with the Austronesian language Biak no matter how many 'founding lineages' (K values) are assumed on runs of the 'Structure' algorithm. In this case, then, 'it thus appears that diffusion overrides phylogeny', as they put it.

But they then take a further step, teasing out the fifteen typological features (out of 160 altogether) which align with phylogeny rather than areality, opposing the Papuan languages Hatam and Meyah against the Austronesian language Biak – see their table 6. Does evidence like this hold the key to refining typological-suite based models so that they can filter out areal noise to find the phylogenetic signal? Obviously, if the argumentation proceeds just from a single case, as here, it risks being merely post hoc, but on the other hand it would be possible to iterate this procedure over a number of areas and small groups.

Will iterations of this type, by filtering out the more from the less diffusable over independent cases from around the world, allow us to fine-tune an algorithm like Structure by weighting the evidentiary value of different typological characters as regards to phylogeny vs areality? This will be a crucial question over the next decades of scholarship as more extensive documentation of Melanesia's languages provides us with more information for feeding into comparative enterprises like Reesink and Dunn's. At the same

time, their work reminds us that, to draw maximum benefit from research like theirs, our language documentations need to ensure that matched typological data is obtained – this need not entail 'questionnaire-style' grammars, but feature lists like those in the Appendix to their article do lay down a basic checklist of typological points which all descriptions should make sure to cover.

The next two papers each consider regions of Melanesia in which there have been complex interplays between languages belonging to quite different families, in a social environment where different types of contact appear to have played a role at different points in the past.

Marian Klamer's *Papuan-Austronesian language contact: Alorese from an areal perspective* focuses on Alorese, an Austronesian language abutting the westernmost group of extant Papuan languages on the island of Pantar. She deduces a complex contact history comprising at least two stages played out in different locations.

The first phase, on her model, would have taken place on the island of Flores or nearby, at a time when Papuan languages were still spoken there. It is at this stage that the language ancestral to modern Lamoholot and Alor would have acquired a suite of typological features that are seen as typically 'Papuan' – or, more precisely, as typical of the Papuan languages of the Alor-Pantar region – including post-predicate negation, the marking of possessors, noun-locational order in locative constructions, the presence of a focus particle and the absence of a passive verb form. This 'Papuanisation' of proto-Lamoholot would have taken place under conditions of long-term stable contact involving preadolescents acquiring the complexities of both Papuan and Austronesian languages and melding them into a new system.

In a second phase, following the migration of Alorese speakers to Pantar and the separation this entailed from their Lamoholot cousins, a series of further changes would have taken place. Alorese contrasts drastically with Lamoholot in terms of morphological complexity. Where Lamoholot has two sets of subject affixes to the verb (prefixes for transitives, suffixes for intransitives), Alorese relies on free pronouns with all but a few frequent verbs which retain fossilised agent prefixes. And where Lamoholot has a number of derivational affixes (some productive, some lexicalised), Alorese has no derivational morphology at all – reduplication is its only productive word formation process. These differences suggest a radical process of morphological simplification in the passage from Lamoholot to Alorese. Klamer hypothesises that, in the initial stages of Alorese settlements of Pantar and Alor, Alorese-speaking men would have taken as their wives women speaking a number of different Papuan languages of the inland. Entering the speech community as adults they would have learned a simplified form of Alorese, jettisoning almost all of its morphology. The contact between Alorese and local Papuan languages, however, was neither prolonged nor consistent at this stage. The number of loanwords from local Papuan languages is relatively low (only 14 Alorese terms out of a 270 word-list have a known Papuan source) and is moreover distributed evenly across the different Papuan languages of the locality. This suggests a number of relatively weak contacts and no stable pattern of bilingual contact.

This case study illustrates a type of multi-phase contact scenario likely to have been played out between Austronesian and Papuan speakers in a number of parts of Eastern Indonesia at different phases over the last two to three millennia. The very different

outcomes of the two phases posited in Klamer's model are a salutary reminder of the social and linguistic complexity that must have been involved between two groups who would have been demographically equally poised and interdependent in many ways. At the same time, as Klamer points out, it is only a reconstruction, and we would be on much firmer ground if we were able to draw on contemporary sociolinguistic studies of the types of interaction – social and linguistic – that are occurring between groups along the Papuan-Austronesian interaction zone. As with so many of the questions raised in this issue, the time for this sort of study is running out fast, as the presence of an alternative lingua franca (e.g. Indonesian) radically alters the type of linguistic interaction between such groups.

From Nusa Tenggara we then move east to the Southern New Guinea region, the focus of Nicholas Evans' *Even more diverse than we thought: The multiplicity of Trans-Fly languages*. In contrast to the Austronesian-Papuan interactions in the preceding two articles, here the interactions are between various unrelated Papuan groups. Southern New Guinea is an intriguing zone, of great diversity, about which our level of knowledge dips even lower than the norms for elsewhere in Melanesia.

The Southern New Guinea region is essentially a nucleus of several small language families surrounded by Trans-New Guinea languages which significantly outnumber them demographically, and which at the time of first colonial documentation tended to be far more expansive and militarised than their non-TNG counterparts. It offers an excellent opportunity for historical linguistics to study the mechanisms by which Trans-New Guinea languages have expanded into areas previously characterised by greater levels of deep phylogenetic diversity.

Nonetheless, it is clear that all languages of the region share a number of typological characteristics – to the extent that some languages, which have been classified as TNG, like Marind, pattern typologically with other Southern New Guinea languages (as well as some languages further afield, including Yeli-Dnye and Inanwatan – see Reesink and Dunn, Fig. 1, as well as discussion in footnote 4 of Evans' article.) This suggests that, even if the presence of TNG languages in Southern New Guinea results from expansion at the expense of other groups, there must have been enough stable long-term bi- or multilingualism for significant linguistic convergence to occur. The languages of the Southern New Guinea exhibit high levels of morphological complexity allied with a host of highly unusual typological features, and Evans' paper gives short sketches of two neighbouring but unrelated languages – Nen and Idi – focussing particularly on their complex verbal and case morphology. He shows that, despite the presence of some convergent features and widespread bilingualism and contact between the speakers of these two languaes, there are major differences in how they organise their grammars. (Note in passing that the Reesink et al 2009 sample did not include any language from the Pahoturi River family which Idi belongs to, so it is not clear how far it would fare on their typological profile).

The Southern New Guinea case – rooted as it is in a system of marriage by sister-exchange which favours comparable demographies, interdependence, and intermarriage and multilingualism between neighbouring groups – is a clear case of how prolonged language contact can lead to areal patterns characterised by shared complexification. In illustration of this, Evans considers the way different languages in Southern New Guinea derive a three-valued number contrast. (singular, dual, plural). This is something found in virtually every language in Southern New Guinea except Marind (some languages have an

additional trial or paucal). But the exact route by which such systems are derived varies significantly from language to language. This suggests that whatever series of pathways leads to shared areal features of this type is a long and tortuous one, probably based on the slow patchwork emergence of grammatical solutions to particular semantic targets shared across languages of the region.

Staying in Southern New Guinea and sticking to the topic of grammatical number, Wayan Arka's paper *Projecting morphology and agreement in Marori, an isolate of southern New Guinea* examines issues of how to represent systems of 'constructed' grammatical number featurally, focussing on the TNG-level isolate Marori and other languages with comparable phenomena.

Marori, in line with the South New Guinea pattern described in Evans' article, constructs a three-valued number system by combining two binary values in a system of distributed exponence. However, the base features used to derive this result are different: where Nen crosses a singular vs non-singular with a dual vs non-dual distinction, Marori crosses a singular vs non-singular with a plural vs non-plural distinction. Arka's articleshows how the unification of number values in Marori morphology can be derived within a model in which the features are hierarchically structured, in different ways in different languages. Thus where Marori treats dual as the number value that is neither singular nor plural – and hence relegates the dual to a derived category – the Nen feature structure builds in dual as a primary specified feature, but treats plural as a derived category that is neither singular nor dual. This model is an elegant illustration of how some cross-linguistic variability in feature structure can be built into a robust overall architecture – the presence of an overall feature structure, and of a primary singular vs non-singular cut, remain constant, but the internal makeup of the non-singular subspace differs as between Marori and Nen. The availability of differing feature architectures then makes it possible to model the differences between languages with similar sets of contrasts, but derived in different ways, within a formalism like LFG – we refer the reader to that chapter for the formal details.

The difficulties involved in lining up language-specific descriptive categories with comparative concepts are nicely illustrated, from a different theoretical perspective, in Mats Exter's article *'Realis' and 'Irrealis' in Wogeo: A valid category?* Recall that, in Reesink and Dunn's article, one of their questions (50/87, as listed in their appendix) is 'Is a distinction between realis/irrealis mood available as a morphological choice (1: present, 0: absent)'? But how do we decide what 'realis/irrealis' actually means? Wogeo offers interesting difficulties in answering this question.

Wogeo is a 'mood-dominated' Austronesian language, spoken off the north coast of PNG, with a complex verbal morphology including six prefixal and eight suffixal slots. A basic opposition is between the 'realis' and 'irrealis' forms of the pronominal prefix, illustrated by a pair like *o-lako* 'I go, I went' vs *go-lako* 'I must go, I want to go, I will go (now)'. If this were all there was to the opposition the characterisation would be fairly straightforward, but once we consider more semantically precise combinations problems arise. Wogeo has additional prefixal combinations expressing such meanings as future, tentative ('try doing X'), counterfactual ('would have done X'), proximal imperfective ('am/was doing X, nearby')' and distal imperfective ('am/was doing X (further away)'), which are followed by either the realis or irrealis prefix, e.g. *m-o-lako* [FUT-1sg.realis-go] 'I will go, I can go, I may go'.

For some of these, the choice of prefix makes sense in terms of normally-characterised properties of this opposition, e.g. the realis is used with the two imperfective series. But for others, notably the counterfactual, it is the realis series that is chosen rather than the expected irrealis. Exter then goes on to consider what such cases mean for the overall enterprise of trying to define terms like realis and irrealis in cross-linguistic terms. He ends up arguing against the usefulness of a term with as broad a range as the realis-irrealis contrast – which, if accepted, raises the possibility that typological comparisons may be more successful if they work at much more semantically-specified levels where cross-linguistic comparison can be more precise.

The next two papers examine the embedding of language in its sociocultural and psychological contexts.

Darja Hoenigman's paper *From mountain talk to hidden Talk: Continuity and change in Awiakay registers* examines the diachronic sociolinguistics of special registers in Awiakay, a language of East Sepik province, and in the process throws a fascinating light on how ideologies of the need for linguistic difference intersect with high levels of metalinguistic awareness to drive a dynamic of lexical innovation. Particularly noteworthy is the continuity – in terms of utilising special registers – that holds in the face of significant change – in the form of Christian strictures against the ongoing use of some traditional registers.

Traditionally, Awiakay people used a special register, known as 'mountain talk', to protect themselves from mountain spirits when travelling up into mountain regions; this involved the substitution or avoidance of a number of lexical items. The arrival of Christianity has arrested the use of 'mountain talk', with the recognition it gives to the power of pagan spirits, and knowledge and use of this traditional register is in decline. But at the same time, another special register has come into use, *kay menda* or 'hidden talk'. Travelling outside the village to regional centres such as Wewak, especially when it is for commercial purposes which leaves the travellers vulnerable to theft and predation, is regarded as a risky business and speaking a language impenetrable to outsiders provides good security.

Though Awiakay is traditionally spoken in just one village, and would therefore normally have been incomprehensible to outsiders, the recent arrival of Tok Pisin loanwords creates chink in the armor of linguistic impenetrability. It is precisely these loanwords which get replaced in *kay menda*, through ingenious native coinages some of which have already won full acceptance in the community and others of which still include rival coinages.

Hoenigman's paper includes subtitled video footage of a journey from the village to the regional centre, during which we can witness the camouflaging processes of 'hidden talk' at work, as well as watching the rehearsal and induction of less experienced members of the party while travelling towards the destination. This is of interest not just for the topic of special registers, but more generally for our understanding of how at least some of the processes of linguistic diversification in Melanesia are driven along by very conscious and negotiated processes of change aimed at differentiating one's language from that of other groups.

Her paper concludes by surveying the parallels and differences between the new register of hidden talk and the fading old register of 'mountain talk'. Both are used in unfamiliar,

perilous territory where one goes to obtain valued items, encountering dangerous entities (mountain spirits before, rascals now) and dangers (sickness before, robbery and theft now), preventing these dangers through judicious out-of-the-ordinary language use, and predominantly involving men who are the ones travelling to the dangerous destinations. The most interesting difference, on her comparison, has to do with who is held to have created the special register. In the case of mountain talk this is attributed to 'mountain spirits' deep in the past, whereas in the case of 'hidden talk' the process of creation is still taking place, involves contemporary Awiakay individuals, and is therefore a process that is amenable to direct research on such questions as how rival innovations are selected between, which items are chosen for camouflaging, and how changes are propagated from innovative individuals to the community.

Michael Frank's paper, *Cross-cultural differences in representations and routines for exact number*, leads us from the known diversity of Melanesian languages to the presumed but untested cognitive diversity this subtends. Beller & Bender (2008), whom he quotes in his article, observe that 'there may be no other domain in the field of cognitive sciences where it is so obvious that language (i.e., the verbal numeration system) affects cognition (i.e., mental arithmetic).' Combining this with the likelihood that Melanesian diversity in numeral systems (Lean 1992) is perhaps even greater, in relative and absolutive terms, than in other aspects of the language systems, we have here a fascinating domain for the investigation of how linguistic diversity shapes cognitive diversity – as well as how cultural practices like different counting routines themselves select for the emergence of different types of numeral system.

Frank's paper does not in itself begin the exciting project of investigating how Melanesian diversity in numeral systems produces (or doesn't) significant differences in cognition. Rather, its goal is to clarify the relations between, on the one hand, how exact number is represented linguistically or through other types of representational tool such as the the Mental Abacus, and numerical cognition on the other. Frank adduces experimental evidence that linguistic systems in the form of numerals, but also non-linguistic systems in the form of the Mental Abacus, both provide a 'cognitive technology' enabling the online encoding and manipulation of quantity information. Frank shows that cultural exposure alone does not scaffold the manipulation of exact number, that the lack of exact numeral terms in a language impacts negatively on arithmetical manipulations, and that it is not enough to possess a language with appropriate terms but that one must also be able to access it online in order to successfully carry out arithmetical calculations.

Different numeral systems, such as the use of different bases (2, 5, 6, 10, 20) can be expected to furnish different cognitive strategies in this sense; and so would different numeral sets for different types of counted objects. This suggests that collaborative research on the impact of numeral systems on numerical cognition will yield rich results. Yet the challenges of investigating this interaction are great, and require a type of collaboration between linguists and psychologists that have been all too rare so far:

> The data that lead to this conclusion could not have been gathered by the standard methods of cognitive psychology, nor by the standard methods of field linguistics. Many of the results cited here come from carefully controlled studies performed in the field with populations that possess culturally, linguistically, or cognitively

> interesting numerical representations. This generalization suggests the benefits of psycholinguistic fieldwork that combines experimental design with cross-cultural or cross-linguistic populations (Frank, this volume:234).

In fact, the relevant numerical systems are rather fragile, in some cases significantly more so than other parts of the language system: 'there are many cases where a language is not endangered, or not particularly endangered, but whose numeral systems are endangered' (Comrie 2005). Oksapmin is a salutary Melanesian case, investigated by Saxe (1982) and then Saxe & Esmonde (2005). Though the language is still healthy (Loughnane 2007) its distinctive base-27 body count system is giving way to an English-style decimal system in a modern setting where counting is most commonly applied to money. This great fragility of numeral systems means that there is an exceedingly narrow time window for carrying out the sort of collaborative work on the impact of numeral systems on numerical cognition which is outlined in Frank's article.

The many fascinating questions and research thrown up by the preceding articles – and even more so, the future research which we hope they will stimulate – generate an enormous amount of primary data as fieldworkers of a range of interests and nationalities record materials on numerous language varieties, in increasingly data-rich formats. But endangered data is a problem we need to take almost as seriously as endangered languages themselves – field notes and recordings may end up lost, uncatalogued, unlocatable, or degenerate on old tapes or other materials. Equally problematic are questions about where data should be housed and who should get access to it. Modern digital archives are giving us the power to address these issues in an efficient way. They have the potential to preserve huge amounts of data far into the future while allowing them to be accessible to researchers and community members from all locations. It is with the design and running of one such archive, Paradisec, that the last article, by Nicholas Thieberger and Linda Barwick, is concerned: *The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC): A resource for Melanesian linguistics.*

PARADISEC was established in Australia in 2003, by a team of researchers led by Thieberger and Barwick. It was born as a response to the challenges set out in the preceding paragraph, from an awareness that a vast body of hard-won field data was at risk of vanishing altogether– partly as a result of poor facilities in local archives (e.g. lack of air-conditioning to maintain tapes in good condition), partly through technological changes (e.g. the disappearance of machines able to read old recordings on wax cylinders, wire-recorders etc.), partly through a lack of emphasis in the field of linguistics on the primacy of documentation as opposed to theoretical debate or grammar-writing, and partly through the reluctance of individuals to make their materials available to others until they had analysed them themselves – a moment which sometimes gets overtaken by Alzheimer's or death.

Part of their article is devoted to showing how PARADISEC works, in terms of equipment setup, backup, access, workflow and data ingestion. But Thieberger and Barwick also discuss a number of other important concerns raised in discussions of where archives like PARADISEC fit in a region characterised by such vast discrepancies between countries in terms of living standards, technology, access to digital data, and the potential value of local information. A central issue is the moral tension centred on the foreseeable

and unforeseeable uses of archived information (e.g. in establishing clan ownership of lands, rights to royalties etc.) which require graduated levels of access, on the philosophy that there should be general commitment to permanent archiving, for future safety's sake, but that communities and researchers should be supplied with technologies for regulating access where this is warranted.

Looking in the other direction, the potential for harnessing the collective knowledge of various kinds of expert through cumulative annotation of archived material by different archive users at different locations is a goal that has great potential to galvanise a more collective and interdisciplinary approach to adding commentary and interpretation to primary material through time.

The nine contributions we have outlined can do no more than give a tantalising glimpse of the challenges raised by the languages of Melanesia – for linguists and scholars in allied fields, but also for educators, communication technologists, and development agencies wanting to focus on local knowledge and expertise. Most importantly, this is a challenge of utmost interest to community members wanting to maintain the intellectual wealth held in their linguistic heritage. For them, collaborative work with linguists and others can offer new ways of integrating that heritage with other sorts of language products such as orthographies, dictionaries, grammars, text collections, and digital ethno-encyclopaedias.

For even a fraction of these challenges to be met, many things must happen. We need to attract a new generation of adventurous and capable young scholars to work in this fascinating, diverse and hospitable part of the world. We need to build capacity among local linguists and language workers in the countries where these languages are spoken, so as to reverse the drastic current imbalance between where Melanesian languages are spoken and where future researchers can receive advanced training in how to study them. CELD, the Centre for Endangered Language Documentation in Manokwari, which hosted the conference where most of the papers here were presented, is a promising step in this direction.

There need to be many other developments like this, and international funding agencies need to be convinced that language diversity is a resource, not a handicap. This is particularly relevant at a juncture when key sources of international research funding over the last two decades (the Volkswagenstiftung's DoBeS program, the Hans Rausing Endangered Languages Program, the NWO *Bedreigde Talen* program and the ESF EuroBABEL program) are drawing to a close, or have already. There is vast potential in such new approaches as BOLD or Basic Oral Language Documentation (http://www.boldpng.info/iwlp) and mobile-phone based crowd-sourcing to assist the data-gathering process. But the need for long-term traditional fieldwork drawing on the knowledge of linguists who learn the languages and cultures on-site will remain fundamental. Finally, while there will always be some divergence of interest between missionary organisations and academically-motivated researchers, the vast extent of missionary enterprises through Melanesia means that the potential for fruitful collaborative work is vast, given goodwill on both sides. An important recent initiative is the reestablishment of the journal *Language and Linguistics in Melanesia*, now as an open-access on-line journal (http://www.langlxmelanesia.com/), as a forum for publishing peer-reviewed research and book reviews on the languages of Melanesia.

As can be seen from these considerations, and the fact that almost every paper in this collection is an early step in a new research path, the study of Melanesia's languages offers abundant opportunities to make new discoveries We hope that in the collection of papers gathered here you will find material that invites you into an engaged and diverse international community of scholars dedicated to advancing our understanding of a linguistic territory that is arguably the least charted on earth.

References

Beller, S. & A. Bender. 2008. The limits of counting: Numerical cognition between evolution and culture. *Science* 319. 213-215.

Comrie, Bernard. 2005. Endangered numeral systems. In Jan Wohlgemuth & Tyko Dirksmeyer (eds.), *Bedrohte Vielfalt: Aspekte des Sprach(en)tods* [*Endangered Diversity: Aspects of Language Death*], 203-230. Berlin: Weissensee Verlag.

Frank, Michael C. This volume. Cross-cultural differences in representations and routines for exact number.

Haspelmath, Martin. 2009. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86(3). 663-687.

Lean, G. 1992. Counting systems of Papua New Guinea and Oceania. Lae: Papua New Guinea University of Technology doctoral dissertation.

Loughnane, Robyn. 2007. A grammar of Oksapmin. Melbourne: University of Melbourne PhD thesis.

Saxe, G. B., 1982. Developing forms of arithmetical thought among the Oksapmin of Papua New Guinea. *Developmental Psychology* 18(4). 583-594.

Saxe, G. B. & Esmonde, I. 2005. Studying cognition in flux: A historical treatment of *fu* in the shifting structure of Oksapmin mathematics. *Mind, Culture, and Activity* 12(3/4). 171-225.

Nicholas Evans
nicholas.evans@anu.edu.au

Marian Klamer
M.A.F.Klamer@hum.leidenuniv.nl

2

Language Documentation & Conservation Special Publication No. 5 (December 2012)
Melanesian Languages on the Edge of Asia: Challenges for the 21st Century,
ed. by Nicholas Evans and Marian Klamer, pp. 13–33
http://nflrc.hawaii.edu/ldc/sp05/
http://hdl.handle.net/10125/4559

The languages of Melanesia: Quantifying the level of coverage

Harald Hammarström
Max Plank Institute for Evolutionary Anthopology
Sebastian Nordhoff
Max Plank Institute for Evolutionary Anthopology

The present paper assesses the state of grammatical description of the languages of the Melanesian region based on database of semi-automatically annotated aggregated bibliographical references. 150 years of language description in Melanesia has produced at least some grammatical information for almost half of the languages of Melanesia, almost evenly spread among coastal/non-coastal, Austronesian/non-Austronesian and isolates/large families. Nevertheless, only 15.4% of these languages have a grammar and another 18.7% have a grammar sketch. Compared to Eurasia, Africa and the Americas, the Papua-Austronesian region is the region with the largest number of poorly documented languages and the largest proportion of poorly documented languages. We conclude with some dicussion and remarks on the documentational challenge and its future prospects.

1. Introduction. We will take Melanesia to be the sub-region of Oceania extending from the Arafura Sea and Western Pacific in the west to Fiji in the east – see the map in figure 1.[1] This region is home to no fewer than 1347 (1315 living + 32 recently extinct) attested indigenous languages as per the language/dialect divisions of Lewis (2009), with small adjustments and adding attested extinct languages given in table 1.

[1] The authors wish to thank two anonymous reviewers for helpful comments.

Action	Language	Location	Living/Extinct	Brief Rationale
Added	Bai of Miklucho-Maclay	PNG, Madang	Presumed Extinct	Not the same as Dumun (Z'graggen 1975:13-14)
Added	Nori	PNG, Sandaun	Extinct	Not the same as Warapu (Corris 2005, Donohue & Crowther 2005, Wilkes 1926)
Added	Kaniet of Dempwolff	PNG, Manus	Presumed extinct	Not the same as Kaniet of Thilenius (Blust 1996)
Added	O'oku	PNG, Northern Province	Presumed Extinct	Seemingly a Yareban language (Ray 1938a)
Added	Butam	PNG, New Britain	Extinct	Laufer 1959
Added	Pauwi of Stroeve and Moszkowski	Indonesia, Papua	Presumed Extinct	May have been a mixed village (Moszkowski 1913), but in any case not the same as Robidé van der Aa's Pauwi (Robidé van der Aa 1885) which we count as Yoke [yki]
Added	Batanta	Indonesia, Raja Ampat	Presumed Extinct	Remijsen (2002:42) cites reports of unintelligibility with neighbouring languages and data appears in Cowan (1953)
Added	Mansim	Indonesia, Bird's Head	Rumours of c.50 speakers in the Manokwari area	Reesink 2002
Added	Binahari-Ma	PNG, Northern Province	Alive	Arguably a different language from Binahari-Neme (Dutton 1999)
Added	Nese	Vanuatu	Alive	Crowley 2006a
Added	Womo-Sumararu	PNG, Sandaun	Alive	Donohue and Crowther 2005
Removed	Dororo [drr]	Solomon Islands	Extinct	Not different from Kazukuru (Dunn and Ross 2007)

Action	Language	Location	Living/Extinct	Brief Rationale
Removed	Guliguli [gli]	Solomon Islands	Extinct	Not different from Kazukuru (Dunn & Ross 2007)
Removed	Makolkol [zmh]	PNG, New Britain	Possibly Extinct	Unattested (Stebbins 2010:226)
Removed	Wares [wai]	Indonesia, Papua	-	Unattested or same as Mawes [mgk] (Wambaliau forthcoming)
Removed	Yarsun [yrs]	Indonesia, Papua	-	Unattested or same as Anus [auq] or Podena [pdn] (van der Leeden 1954)

TABLE 1. Adjustments concerning the languages of Melanesia to the language catalogue of Lewis (2009). We have not added totally unattested, very poorly attested languages (e.g., Ambermo, attested in two numerals, Fabritius 1855), or once attested languages whose attestation has disappeared (e.g., Rutan, only 3 words now remaining, Crowley 2006b:3).

FIGURE 1. Map of Melanesia adapted from http://en.wikipedia.org/wiki/Melanesia accessed 10 July 2011. The countries present in Melanesia are Papua New Guinea, Indonesia, Fiji, France (New Caledonia), Solomon Islands and Vanuatu.

The present paper seeks to describe the current state of description of the languages of Melanesia in detail (in the online appendix at http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4559/melanesia_appendix.pdf) and in general (in the body of the paper) based on a database of annotated bibliographical references. This database of references, called LangDoc (Hammarström & Nordhoff 2011), spans the entire world but we restrict it to the Melanesian subset in the present survey.

2. Assessing Status of Description. To assess status of description we first a) collect all relevant bibliographical references, b) annotate them as to (target-)language and type (grammar, wordlist etc), and c) for each language, mark its status of description according to the most extensive or sum description it has.

2.1. Collecting References. Language documentation and description is, and has been, a decentralized activity carried out by missionaries, anthropologists, travellers, naturalists, amateurs, colonial officials, and not least linguists. In order to comprehensively collect all relevant such items, we have, in essence, gone through all handbooks and overviews concerning the Melanesian region, in the hope that specialists on families and (sub-)regions have the best knowledge on what descriptive materials actually exist. This is supplemented by a) intensive searching as to (sub-)regions for which there is no recent expert-written handbook/overview paper and b) whole-sale inclusion of relevant existing bibliographical resources such as the WALS, the SIL Bibliography, SIL Papua Guinea Bibliographies, the library catalogue of MPI EVA in Leipzig and so on – see Hammarström and Nordhoff (2011) for a little more detail regarding this procedure and alternatives.

Everything published by a locatable publisher has been included as well as MAs and PhDs since they should, in principle, be findable via the national library or the degree-giving institution. However, field notes, manuscripts, self-published items and items published by a local bible society have not been included since they cannot be located systematically. In our experience, locating manuscripts too often turns out to be a wild goose chase and including them in the current survey would do more harm than good, in particular, it would give a false picture of the state of (accessible) description. However, we have included a small number of manuscripts and/or fieldnotes where the item in question has been posted on the internet and/or is verified to be located in a publicly accessible archive (e.g., the KITLV in Leiden), and thus meets the accessibility criterion.

It should be stressed, however, that the amount of original and valuable data sitting in unpublished form is highly significant. To give just a few examples, Capell (1962) cites a large number of missionary manuscripts from the islands east of the Papuan mainland, the archives of the SIL in Jayapura and Ukarumpa (cf. Silzer & Heikkinen-Clouse 1991) hold a huge number of unpublished survey wordlists and/or grammar sketches spanning (in our impression) at least 50% of the languages of Melanesia, and linguists Mark Donohue and William Foley have unpublished field data from Indonesian Papua and the Sepik-Ramu region respectively which is enough for several full grammars and dozens of grammar sketches (p.c. Mark Donohue 2008 and William Foley 2010). If unpublished material is included, the descriptive picture of the languages of Melanesia changes significantly, especially on the breadth side, with far more data on the lesser-known languages (cf. Carrington 1996).

In total, the bibliographical database contains 11 290 references pertaining to Melanesia.

2.2. Annotation. Bibliographical references are annotated as to identity, i.e., the iso-639-3 code of the language(s) treated, and type of description, i.e., grammar, wordlist etc. As to type, the following hierarchy has been used:

- grammar: an extensive description of most elements of the grammar: 150 pages and beyond
- grammar sketch: a less extensive description of many elements of the grammar 20–150 pages (typically 50 pages)
- dictionary: 75 pages and beyond
- specific feature: description of some element of grammar (i.e., noun class system, verb morphology etc)
- phonology: phonological description with minimal pairs
- text: text (collection)
- wordlist: a couple of hundred words
- minimal: a small number of cited morphemes or remarks on grammar
- sociolinguistic: document with detailed sociolinguistic information
- comparative: inclusion in a comparative study with or without cited morphemes, e.g., lexicostatistical survey
- handbook/overview: document with meta-information about the language (i.e., where spoken, non-intelligibility to other languages etc.)
- ethnographic: ethnographic information on the group speaking a language

The hierarchy is an ad-hoc amalgam of existing annotation, automatizability properties and bias towards typologist usage (with grammar at the top, trumping text and dictionary, and form-function pairs rated higher than sociolinguistic information). It is in many ways imperfect, but it is more informative than nothing. Other existing schemas could not be felicitously adopted, e.g., Moore (2007:33) is similar to the present scheme but credits the existence of various types (scientific articles, dissertations, etc.) rather than their actual content, and AIATSIS (2011:285-297) is also similar to the present scheme but so much more detailed (several hundred categories including vocabulary/animals, vocabulary/body parts, etc.) that it could not be automatized or done by hand within the scope of the present project. Bibliographical references in the present project have been annotated both automatically and by hand. Some examples are shown in Table 2.

Reference	Language	Type	Comment
Lindström, Eva. (2002) Topics in the Grammar of Kuot. Stockholm University doctoral dissertation, 265pp.	Kuot [kto]	grammar	although it contains some text and a Swadesh word-list at the end, it counts as grammar
Franklin, Karl J. & C. L. Voorhoeve. (1973) Languages near the intersection of the Gulf, Southern Highlands and Western Districts. In Karl J. Franklin (ed.), *The linguistic situation in the Gulf District and adjacent areas, Papua New Guinea* (Pacific Linguistics: Series C 26), 149-186. Canberra: Research School of Pacific and Asian Studies, Australian National University	Fasu [faa], Foe [foi], Fiwaga [fiw], Kewa [kew]	overview; comparative; minimal	There is a discussion of comparative matters and a number of morphemes are given (for each language).
Wirz, Paul. (1924) *Anthropologische und ethnologische Ergebnisse der Central Neu-Guinea Expedition* 1921-1922. *Nova Guinea* XVI. 1-148.	Zwart Valley = Dani-Western [dnw]	ethnographic; grammar sketch	It contains a grammar sketch in addition to ethnographic data.
Hughes, Jock. (1987) The languages of Kei, Tanimbar and Aru: Lexicostatistic classification. In Soenjono Dardjowidjojo (ed.), *Miscellaneous studies of Indonesian and other languages in Indonesia, part 9* (NUSA: Linguistic Studies of Indonesian and Other Languages in Indonesia 27), 71-111. Jakarta: Universitas Katolik Indonesia Atma Jaya.	Mariri [mqi], East Tarangan [tre], Lorang [lrn], Lola [lcd], Koba [kpd], Kompane [kvp], Batuley [bay], Barakai [baj], Karey [kyd]	overview; comparative	No actual words or wordlists are included, just results of comparing wordlists.

TABLE 2. Examples of the annotation scheme used in the present survey.

Automatic annotation is possible when the title words contain the language name and/or word(s) revealing the type of the document, e.g., "A grammar of Tauya" can be automatically recognized as [tya] and grammar. Exactly how this is done and what percentages of correctness are to be expected is described in Hammarström (2008, 2011).

For most references, number of pages is recorded, and is used to rank within categories.

2.3. STATUS OF DESCRIPTION PER LANGUAGE. For each language, the references concerning it are aggregrated and its status of description is straightforwardly assessed as

per the annotation hierarchy. In addition, for the purposes of the current presentation, it has been simplified into a more distilled scheme as per Table 3.

type	distilled type	numerical value
grammar	Grammar	4
grammar sketch	grammar sketch	3
dictionary	phonology/dictionary/specific/text	2
text	phonology/dictionary/specific/text	2
specific feature	phonology/dictionary/specific/text	2
wordlist	wordlist or less	1
minimal	wordlist or less	1
sociolinguistic	wordlist or less	1
comparative	wordlist or less	1
handbook/overview	wordlist or less	1
ethnographic	wordlist or less	1
<type annotation lacking>	wordlist or less	1

Table 3. The full- and distilled description level hierarchy used in the present survey.

There may be missing extant references and manual as well as automatic annotation has gaps and errors. The claim we are able to make is that at least the status of description for every language should be correct. That is, the outcome has been screened at the language level by an informed human, and inasmuch as errors of omission and annotation remain, they do not alter the (correct) status of description of any language. Thus, for a language which only has a published wordlist to its documentation it may be that there are several wordlists published, but only one of them is accurately reflected in the database (accurately reflecting the others would not change the status of description away from wordlist), and, if a language is given a certain status of description, the claim is that there is, in reality, no other descriptive publication that would give it a higher mark. Of the publications that are the witness to the status of description of a language (the most significant items of description) 95% have been personally inspected by the authors, but, since this was done over a long period of time it is no guarantee of consistency and we are not in a position to assess the quality of a description.

It should be noted again that the above hierarchy reflects descriptive status and has a bias towards typologist usage. For example, a language that has a grammar, dictionary and text collection will be ranked the same (grammar) as a language with only a grammar, even though the former is better documented overall. An index of overall documentation (e.g., with points separately for grammatical-, lexical- and textual documentation) could be computed from the same database. We do not do this for the present survey since we cannot venture the same claim of completeness as with the grammar-oriented scheme above. In other words, the database screening is likely to have missed cases of missing texts and dictionaries for languages which already have a grammar (sketch). The database is released to the public so that others who are more interested in overall documentation can complete the database and compute figures of their own.

The fact that "grammar" is the highest weighted category of description should not be taken to mean that a language with a grammar is completely described – it merely means that it is the highest category of grammatical description that is commonly distinguished by linguists, i.e., there are as yet no descriptions that are called "super-grammars" or the like. However, grammars can be more or less comprehensive and a correlate of this (with validity only on average) may be the number of pages, which is recorded in the present database. Nor is length more than a rough proxy for quality and comprehensiveness – it would rank a rambling and obtuse document above a concise and elegant one – but it has the virtue of being operationalisable and applicable to the data we have.

3. Status of Description of Melanesian Languages. Results of the full survey are given in the online appendix (http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4559/melanesia_appendix.pdf), sorted by family, author and language. We review the generalities here.

	Living	Extinct	Total	Total as percentage
grammar	207	0	207	15.4%
grammar sketch	245	7	252	18.7%
phonology or sim.	107	2	109	8.1%
wordlist or less	756	23	779	57.8%
			1347	

Table 4. Raw number of languages in Melanesia and their level of description.

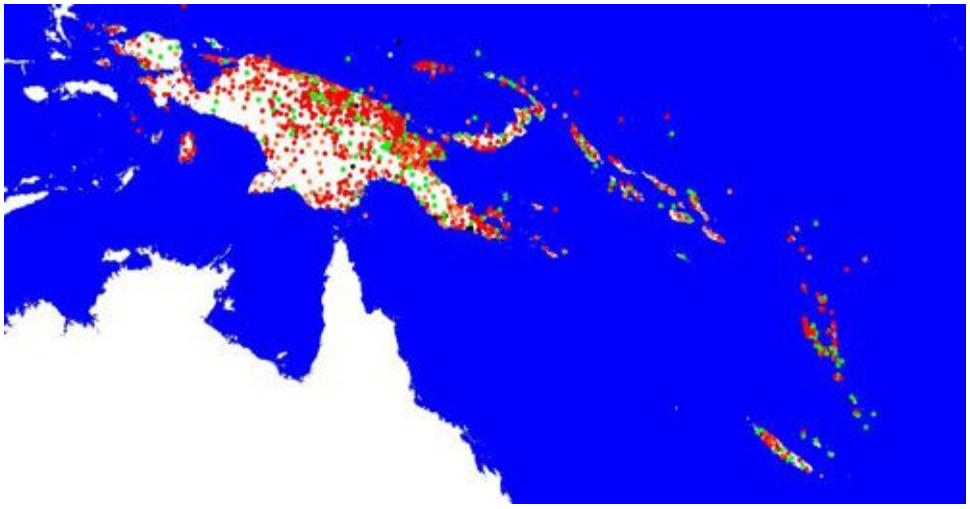


Figure 2. The location and description level of Melanesian languages. The colour coding is grammar = green, grammar sketch = orange or light gray (if extinct), phonology or sim. = orange red or slate gray (if extinct), wordlist or less = red or black (if extinct).

Raw numbers of languages described to various degrees are shown in table 4 and a map is shown in figure 2. The numbers speak for themselves, yet the most conspicuous fact is that more than half of the languages of Melanesia have only a wordlist or less of published descriptive material. Any non-trivial generalizing statement concerning the grammar of languages of Melanesia can only be at most half-fully grounded empirically. For example, Wurm (1954), drawing on data and experience from Capell, was acquainted with all Melanesian languages described at the time, and lists some 20 tone languages, whereas surveys of tone on New Guinea half a century later (Cahill 2011, Donohue 1997) turn up far more and far different tonal languages in Melanesia.

Historically speaking, early wordlists were catalogued superbly by Ray (1893, 1912, 1914, 1919, 1920, 1923, 1926, 1929, 1938a, 1938b) for the entire Melanesian area, and the history of research has been adequately surveyed qualitatively by area experts (Beaumont 1976, Chowning 1976, Dutton 1976, Grace 1976, Haudricourt 1971, Healey 1976, Hooley 1976, Laycock 1975, 1976, Laycock and Voorhoeve 1971, Lincoln 1976, Lithgow 1976, Lynch and Crowley 2001, Schütz 1972, Taylor 1976, Tryon and Hackman 1983, Voorhoeve 1975b, Z'graggen 1976). We supplement these with some quantitative results in Figure 3. As can be seen, language description in Melanesia takes off in the second half of the 19th century with travellers, colonial officers, and missionaries producing wordlists. From there description increases at a steady pace, due mostly to missionaries and German scholars. A sharp rise in the number of items produced every year, and a corresponding (but less sharp) increase in the overall descriptive status, happens after 1950, presumably due to the establishment of the SIL in Papua New Guina (Hooley 1968, Foley 1986:13). The pace has since been kept up mainly by SIL missionaries and academic linguists in Australia and other western countries. Very little has so far been produced by Melanesians themselves; notable exceptions include Flassy (2002), Nekitel (1985), Sumbuk (1999). There are more than a dozen languages whose corresponding ethnic groups have a monograph-length ethnographic description, yet the languages are not described beyond a wordlist, e.g., Gnau [gnu] (Lewis 1975) or Banaro [byz] (Juillerat 1993).

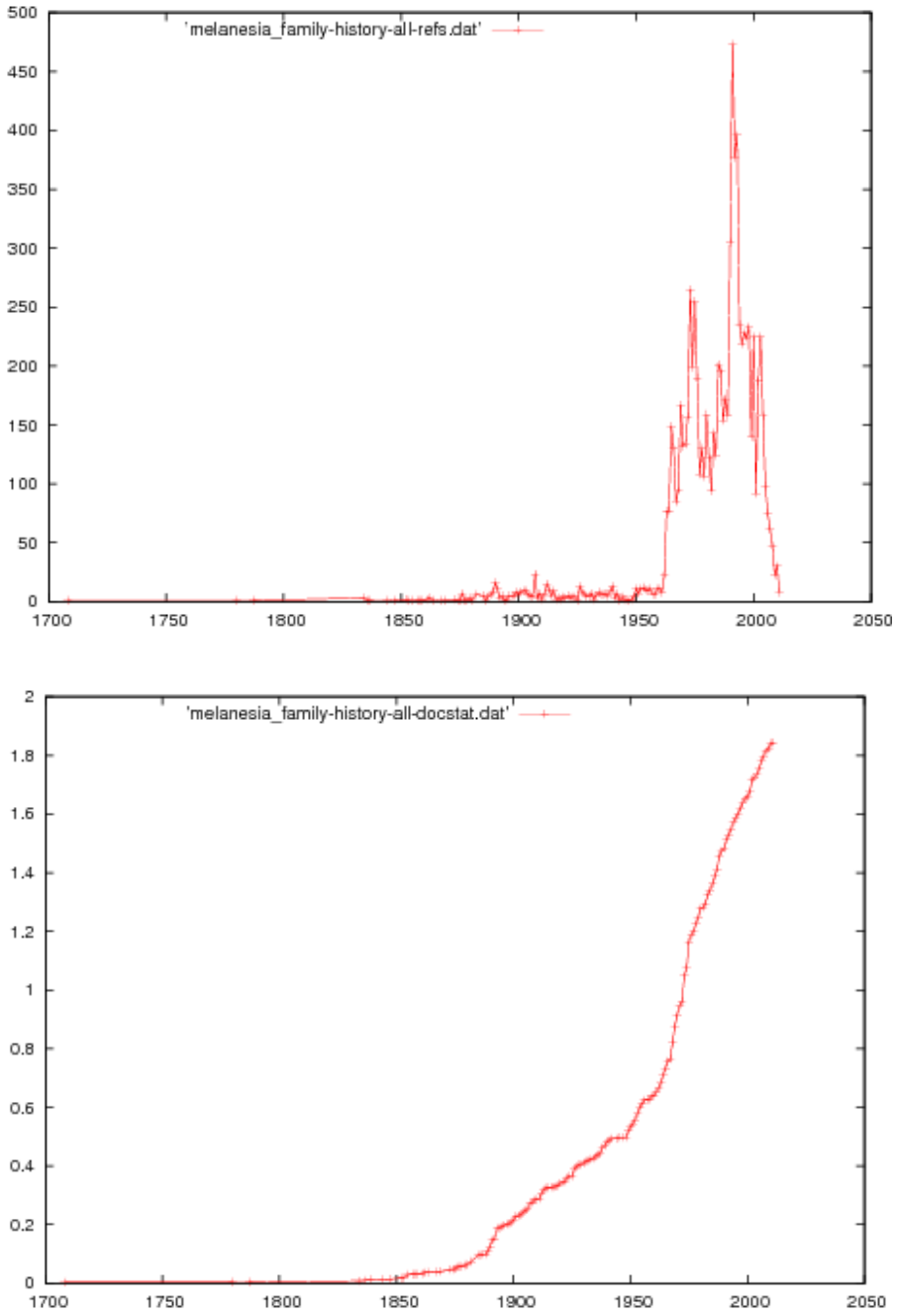


FIGURE 3. The upper diagram shows the raw number of publications per year concerning languages of Melanesia. The lower diagram shows the average description level as it increases through time.

In the early times, languages near the coast were much better known than inland languages. At the present time, this correlation is much diluted. Table 5 shows the median and average distances (as the crow flies) to the coast for the various levels of description, which shows little difference. The slight tendency for grammars to be written of languages nearer to the coast is not statistically significant for average distances, but it is so for median distances. This means that half of the languages with grammars are within 14.97 kms to the coast whereas half of the languages of other categories are 10-15% further away, and that languages with grammars that are not near the coast (the exceptions) are so far away that they blur the tendency on average. This overall lack of a stronger trend must be taken to mean that flight and river access inland, balances the amount of neglected languages on the coast and immediate coastal hinterlands.

	Average distance to coast (kms)	p≈	Median distance to coast (kms)	p≈
grammar	44.91	0.340	14.97	0.026
grammar sketch	46.84	0.463	17.90	0.462
phonology or sim.	46.71	0.466	16.75	0.346
wordlist or less	46.85	0.373	20.09	0.133
overall	46.51		17.95	

TABLE 5. Average and median distance (as the crow flies) for languages of various levels of description. Significance testing is by selecting 1000 random subsets of the corresponding size from the total pool of 1347 languages and checking how many of those have an average/median distance lower viz. higher than the distance to be tested.

As is well-known, the languages of Melanesia divide into two classes, the Austronesian languages (522 languages) and the non-Austronesian languages (825 languages). The Austronesian languages are more coastal (average 12.79 kms and median 9.92 kms from the coast) than the Papuan ones (average 67.92 kms and median 44.66 kms), but since there is only a weak or no trend that favours the description of coastal languages, we can check fairly easily if there is a bias towards the description of Austronesian or non-Austronesian languages. Figure 4 shows that, historically, there was a long time during which AN languages were better described on average (presumably due to being coastal) and in recent times the slightly higher level has been regained. The current average level of description for AN languages in Melanesia is 2.04 against 1.84 for non-AN languages. The difference is slight but highly significant $p \approx 0.002$. The difference is hardly due to the tendency for full grammars to be coastal, as the AN languages have higher representation at all levels (beyond wordlist) as per Table 6. We do not know what the reason for this bias is.

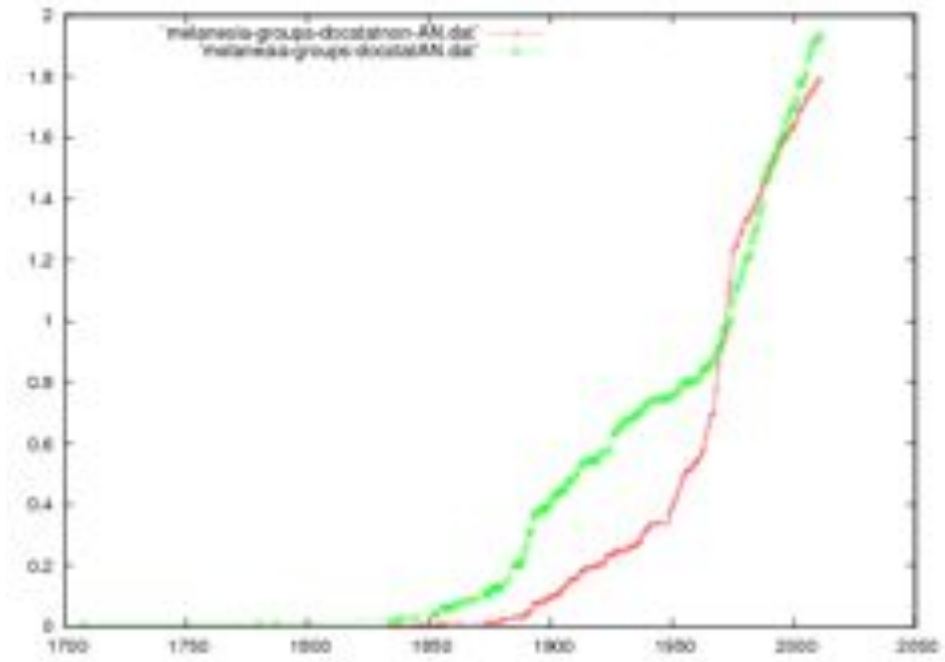


FIGURE 4. The average description level for Austronesian (AN, green) and non-Austronesian (non-AN, red) languages through time.

	Austronesian		non-Austronesian		total	
	number	%	number	%	number	%
grammar	93	17.82	114	13.82	207	15.37
grammar sketch	104	19.92	148	17.94	252	18.71
phonology or similar	55	10.54	54	6.55	109	8.09
wordlist or less	270	51.72	509	61.70	779	57.83

TABLE 6. Numbers and proportions of Austronesian (AN) and non-Austronesian (nAN) languages at different levels of description.

It is difficult to say which is the best described language of Melanesia as that would require a quality judgment that we are not in a position to make. However, the description with the largest number of pages is Lichtenberk (2008)'s 1409-page grammar of To'aba'ita

(an Oceanic Austronesian language of the Solomon Islands). In fact, it is also the longest grammar of any lesser-known language in the world, in terms of number of pages devoted to grammatical description. The second longest grammar of a language of Melanesia is Aikhenvald (2008)'s 727-page grammar of Manambu (a Ndu language). As far as can be told with documents accessible to us, the least described languages whose existence seems certain enough, are Kehu [khh] and Kembra [xkw], two seemingly isolated languages in Indonesian Papua. Kehu is known from from two unpublished minuscule wordlists (Moxness 1998, Whitehouse n.d.) at least one of which is from a non-native speaker, and Kembra is known from a minuscule wordlist taken up from a transient speaker by Doriot (1991) attributed to a village named Kembra near the confluence of the Sobger and Nawa (Kiambra appears at the right place on a colonial map, Hoogland 1940).

Arguably the most prolific author of descriptive work on Melanesian languages has been the Dutch Catholic priest Petrus Drabbe (Voorhoeve 2000) who can count to his name no less than 4 languages with grammars, another 19 with grammar sketches and wordlists for 6 more spanning a range of different families. Linguist Terry Crowley wrote 6 grammars and 9 grammar sketches of Austronesian languages before his premature death in 2005. Linguists such as Arthur Capell, Stephen Wurm, Sidney Ray, Malcolm Ross, J. C. Anceaux, J. A. Z'Graggen, Darrell Tryon and C. L. Voorhoeve have between them published wordlists (or similar bits of information) of several hundred languages, either collected themselves or by others.

A current discussion among linguists as to priorities for documentation – the context being that time is running out – is whether to describe an undescribed isolated language or whether to describe an undescribed language from a family with other described languages. At present, we count 45 language isolates for the Melanesian region (see Hammarström 2010a,b:appendix for a justification of this figure). The 45 isolates have an average description level of 2.20 and the 1 298 non-isolates have 1.91. The difference, however, is not statistically significant at conventional levels of significance ($p \approx 0.070$). That is, there is no overall principle at work that has favoured the description of isolates rather than non-isolates. Nevertheless, there is a conspicuously large *absolute* number of underdescribed isolates and small families in the Melanesian region, especially lowland New Guinea – see Hammarström (2010b) for details.

4. Melanesian Languages in Relation to the Rest of the World. The bibliographical database LangDoc spans the entire world in a fairly uniform way, allowing us to compare Melanesia to other conventional macro-areas of the world. The total database contains over 160 000 references collected and annotated in much the same way as the Melanesian subpart (Hammarström and Nordhoff 2011). Although the Eurasian, Australian and Meso-American sections have not been screened as thoroughly as the other areas yet, the general trends of the comparisons with Melanesia should still be trustworthy. For this section, we will consider all Papuan-Austronesian languages together, not just the Melanesian ones, in order to appropriately cover all of the world's languages. This entails that the Eurasia figures do not include the Austronesian languages of South East Asia, the Philippines and Indonesia. Figures are shown in Table 7.

	Africa	Australia	Eurasia	North America	Papua+AN	South America
grammar	780 [20]	94 [28]	537 [40]	264 [25]	415 [1]	260 [25]
grammar sketch	483 [35]	30 [22]	135 [18]	67 [32]	428 [10]	82 [28]
phonology or sim.	120 [5]	15 [2]	109 [2]	44 [9]	157 [2]	31 [17]
wordlist or less	603 [77]	45 [53]	684 [112]	105 [39]	978 [40]	30 [125]
Total	1986 [137]	184 [105]	1465 [172]	480 [105]	1978 [53]	403 [195]
Average desc.	2.68	2.69	2.31	2.91	2.12	2.88
grammar (%)	37.68	42.21	35.25	49.40	20.48	47.66
living undoc (%)	28.40	15.57	41.78	17.95	48.15	5.02

TABLE 7. The number of languages at various levels of description broken up by macro-areas. The numbers outside brackets refer to strictly living languages and those within brackets refer to extinct. The last row gives the proportion of living languages with only a wordlist of less.

In absolute terms, Papua+Austronesian has the largest number of languages with only a wordlist to their documentation. In relative terms, Papua+Austronesian has the lowest proportion of grammars, the highest proportion of languages with only a wordlist or less, and the lowest average level of documentation. The Melanesia subpart scores slightly lower on all relative accounts. Therefore, Papua+Austronesian, and the languages of Melanesia in particular, can rightly be called the linguistically least known area of the world.

5. 21ST CENTURY CHALLENGES IN DOCUMENTATION. As is clear from the figures above, a formidable challenge for linguistic science is to provide descriptions of the vast number of un(der)described languages in the Melanesian region before it is too late.

On the optimistic side, a) the trend from the past century predicts a continued large production of grammatical descriptions and, b) it seems, impressionistically, that people from a wider array of countries of the world are taking interest in the Melanesian languages, and c) infrastructure in Melanesia is making it easier to reach and live in otherwise remote areas.

On the pessimistic side, a) at the same pace as infrastructure is developing the languages become endangered, b) violence, tropical diseases, visa/permit-matters and lack of funding continue to deter Westerners from in situ fieldwork, c) harnessing of local talent and interest, and the training of linguists from the region, remains extremely undeveloped, and d) large amounts of descriptive work never reach the scientific community, as if such materials had no scientific merit.

A few comments are in order.

The failure of local interest to develop into active descriptive work is not endemic to Melanesia per se, but is widespread in all of the language-rich countries of the world. However, exceptions such as Brazil and Ethiopia show that it is possible for local universities and communities to take a productive interest in local languages.

In addition to unpublished materials alluded to above, many valuable descriptive works are difficult to access, in particular, a large number of unpublished PhD and MA-theses. PhD and MA theses are in many instances the most extensive description there is of a language. Many universities (for instance, the Australian National University) that regularly keep MA-theses do not allow interlibrary loans of them precisely when theirs is the only copy. Other universities, including the convenors of the 3L Language Documentation school, i.e., Leiden University, Université Lumière Lyon II and SOAS, either do not regularly keep awarded MA theses at all, or do not keep them in a manner that allows systematic access (such as the Department library or the main University library). Perhaps the most blatant example of a university in antipathy of its scientific production actually being used is Université Libre de Bruxelles, as the first author experienced personally after making the trip to Bruxelles to read the presumably only library copy of Levy (2002)'s PhD grammar of Nubia-Awar - by far the most extensive description of that language. According to regulations, nobody – be it registered library card holders or visitors – is allowed to *read* this thesis (let alone borrow or photocopy from!) without the written consent of the author.

Similarly, finished documents and reports from SIL Papua New Guinea and SIL Indonesia cannot be systematically accessed, although many items have been made accessible in publication series and other outlets. Dissemination is a scientific principle, and scholarly institutions – be they missionary organizations or universities – that actively or passively restrict access to, or effectively let scientifically valuable documents be thrown away, do not fully merit the label 'scientific institution'. If descriptive work continues to be disvalued in the above exemplified ways, there is less incentive for more descriptive work to be produced.

Apart from first-hand descriptive fieldwork, there are less obvious ways in which one can contribute to the description of Melanesian languages. A non-trivial number of languages of Melanesia have scripture translations, i.e., bodies of text with translation, but no published grammatical descriptions. The languages for which scripture translations are said to exist are given in Lewis (2009). Partial but substantial analyses of grammar can be done on the basis of text data from scripture translations, without fieldwork in situ. Comparative and typological work on languages of Melanesia can help generate interest in producing more detailed descriptions. The digital era allows for tools on management, annotation and interoperability of language resources which can free up time for strictly human-needed analysis for language description. And, if nothing else, publishing or making available legacy resources is a valuable contribution. Prime examples are the publication of Anceaux's gigantic wordlist collection from Indonesian Papua by Smits and Voorhoeve (1992a, 1992b, 1994, 1998), and the digitization of Arthur Capell and Donald Laycock's fieldnotes from Papua New Guinea by PARADISEC (see Thieberger & Barwick, this volume).

6. Conclusion. 150 years of language description in Melanesia has produced at least some grammatical information for almost half of the languages of Melanesia, almost evenly spread among coastal/non-coastal, Austronesian/non-Austronesian and isolates/

large families. Nevertheless, only 15.4% of these languages have a grammar and another 18.7% have a grammar sketch. Compared to Eurasia, Africa and the Americas, the Papua-Austronesian region is the region with the largest number of poorly documented languages and the largest proportion of poorly documented languages.

References

AIATSIS. 2011. *AIATSIS Subject Thesaurus*. http://www1.aiatsis.gov.au/thesaurus/data/SubjectThesaurus.pdf (10 June, 2011.)

Aikhenvald, Alexandra Y. 2008. *The Manambu language of East Sepik, Papua New Guinea*. Oxford & New York: Oxford University Press.

Bakker, Peter & Mikael Parkvall. 2010. Catalogue of Pidgin languages. Paper presented at the second APiCS conference. Max Planck Institute for Evolutionary Anthropology, Leipzig

Beaumont, Clive H. 1976. History of research in Austronesian languages: New Ireland. In Wurm, *Austronesian Languages*, 171-177.

Blust, Robert. 1996. The linguistic position of the Western Islands, Papua New Guinea. In John Lynch & Pat Fa'afo (eds.), *Oceanic Studies: Proceedings of the first international conference on Oceanic linguistics* (Pacific Linguistics C 133), 1-46. Canberra: Pacific Linguistics, Australian National University.

Cahill, Michael. 2011. Tonal diversity in languages of Papua New Guinea. *SIL Electronic Working Papers* 2011-008. http://www.sil.org/silewp/2011/silewp2011-008.pdf.

Capell, Arthur. 1962. *Linguistic survey of the south-western Pacific.* New and rev. edn. (South Pacific Commission Technical Paper 136). Noumea: South Pacific Commission.

Carrington, Lois. 1996. *A linguistic bibliography of the New Guinea area* (Pacific Linguistics D 90). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Chowning, Ann. 1976. History of Research in Austronesian Languages: New Britain. In Wurm, *Austronesian languages*, 179-195.

Corris, Miriam. 2005. A grammar of Barupu, a language of Papua New Guinea. Sydney: University of Sydney PhD thesis.

Cowan, H. K. J. 1953. *Voorlopige Resultaten van een Ambtelijk Taalonderzoek in Nieuw-Guinea*. 'S-Gravenhage: Martinus Nijhoff.

Crowley, Terry. 2006a. *Nese: A diminishing speech variety of northwest Malakula (Vanuatu)* (Pacific Linguistics 577). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Crowley, Terry. 2006b. *Tape: A declining language of Malakula (Vanuatu)* (Pacific Linguistics 575). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Donohue, Mark. 1997. Tone systems in New Guinea. *Linguistic Typology* 1. 347–386.

Donohue, Mark & Melissa Crowther. 2005. Meeting in the middle: Interaction in north-central New Guinea. In Pawley et al. *Papuan Pasts,* 167–184.

Doriot, Roger E. 1991. 6-2-3-4 Trek, April-May, 1991. Unpublished ms.

Drabbe, Peter. 1926. *Spraakkunst der Fordaatsche taal* (Verhandelingen van het Koninklijk Bataviaasch Genootschap van Kunsten en Wetenschappen LXVII:Eerste Stuk). Batavia: Albrecht.

Drabbe, Peter. 1949. Bijzonderheden uit de Talen van Frederik-Hendrik-Eiland: Kimaghama, Ndom en Riantana. *Bijdragen tot de Taal-, Land- en Volkenkunde* 105. 1–24.

Dunn, Michael & Malcolm Ross. 2007. Are Kazukuru languages really Austronesian? *Oceanic Linguistics* 46(1). 210–231.

Dutton, Tom. 1999. From pots to people: Fine-tuning the prehistory of Mailu Island and neighbouring coast, south-east Papua New Guinea. In Roger M. Blench & Matthew Spriggs (eds.), *Archaeology and Language*, vol. 3 (One World Archaeology 34), 90-108. London & New York: Routledge.

Dutton, Tom E. 1976. History of Research in Austronesian Languages: Eastern part of south-eastern mainland Papua. In Wurm, *Austronesian languages*, 129-140.

Etherington, Paul Anthony. 2002. *Nggem morphology and syntax.* Darwin: Northern Territory University (CDU) MA thesis.

Fabritius, G. J. 1855. Anteekeningen omtrent Nieuw-Guinea. *Tijdschrift voor Indische Taal-, Land- en Volkenkunde* IV. 209–215.

Flassy, Don Augusthinus Lamaech. 2002. *Toror: A name beyond language and culture fusion*. Jakarta: Balai Pustaka.

Foley, William A. 1986. *The Papuan languages of New Guinea* (Cambridge Language Surveys). Cambridge: Cambridge University Press.

Foley, William A. 2005. Linguistic prehistory in the Sepik-Ramu Basin. In Pawley et al., *Papuan pasts*, 109-144.

Galis, Klaas Wilhelm. 1955. Talen en dialecten van Nederlands Nieuw-Guinea. *Tijdschrift Nieuw-Guinea* 16. 109–118, 134–145, 161–178.

Grace, George W. 1976. History of research in Austronesian languages of the New Guinea area: General. In Wurm, *Austronesian languages*, 55-72.

Hammarström, Harald. 2008. Automatic annotation of bibliographical references with target language. In *Proceedings of the workshop on multi-source, multilingual information extraction and summarization (MMIES '08)*, 57-64. Manchester: Association for Computational Linguistics.

Hammarström, Harald. 2010a. A full-scale test of the language farming dispersal hypothesis. *Diachronica* XXVII(2). 197–213. With Appendix at http://www.benjamins.com/jbp/series/DIA/27-2/art/02ham.app.pdf.

Hammarström, Harald. 2010b. The status of the least documented language families in the world. *Language Documentation & Conservation* 4. 177–212.

Hammarström, Harald. 2011. Automatic annotation of bibliographical references for descriptive language materials. In Pamela Forner, Julio Gonzalo, Jaana Kekäläinen, Mounia Lalmas & Maarten de Rijke (eds.), *Proceedings of the CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation* (LNCS 6941), 62-73. Berlin: Springer.

Hammarström, Harald & David Kamholz. 2010. A note on Duvle-Wano pidgin. Paper presented at the second APiCS conference. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Hammarström, Harald & Sebastian Nordhoff. 2011. LangDoc: Bibliographic infrastructure for linguistic typology. *Oslo Studies in Language* 3(2). 31–43.

Haudricourt, André G. 1971. New Caledonia and the Loyalty Islands. In Sebeok, *Linguistics in Oceania*, 359-396.

Healey, Alan. 1964. The Ok language family in New Guinea. Canberra: ANU PhD thesis.

Healey, Alan. 1976. History of research in Austronesian languages: Admiralty Islands area. In Wurm, *Austronesian Languages,* 223-231.

Hoogland, J. 1940. Memorie van Overgave van de Onderafdeling Hollandia. Nationaal Archief, Den Haag, Ministerie van Koloniën: Kantoor Bevolkingszaken Nieuw-Guinea te Hollandia: Rapportenarchief, 1950-1962, nummer toegang 2.10.25, inventarisnummer 24.

Hooley, Bruce A. 1968. SIL research in New Guinea. *Kivung* 1(2). 63–70.

Hooley, Bruce A. 1976. History of research in Austronesian languages: Morobe province. In Wurm, *Austronesian languages*, 115-128.

Juillerat, Bernard. 1993. *La révocation des Tambaran: les Banaro et Richard Thurnwald revisités* (CNRS Ethnologie). Paris: CNRS.

Laufer, Carl. 1959. P. Futschers Aufzeichnungen über die Butam-Sprache (Neubritannien). *Anthropos* 54. 183–212.

Laycock, Donald C. 1975. A hundred years of Papuan linguistic research: Eastern New Guinea area. In Wurm, *Papuan Languages*, 43-116.

Laycock, Donald C. 1976. History of research in Austronesian languages: Sepik provinces. In Wurm, *Austronesian Languages*, 73-93.

Laycock, Donald C. & C. L. Voorhoeve. 1971. History of research in Papuan languages. In Sebeok, *Linguistics in Oceania*, 509-540.

Laycock, Donald C. & John A. Z'Graggen. 1975. The Sepik-Ramu phylum. In Wurm, *Papuan Languages*, 731-764.

Leeden, Alexander Cornelis van der. 1954. *Verslag over taalgebieden in het Sarmische van de Ambtenaar van het Kantoor voor Bevolkingszaken* volume 35. Hollandia: Gouvernement van Nederlands-Nieuw-Guinea, Dienst van Binnenlandse Zaken, Kantoor voor Bevolkingszaken.

Levy, Catherine. 2002. A tentative description of Awar phonology and morphology: Lower Ramu family, Papua-New Guinea. Brussels: Université Libre de Bruxelles doctoral dissertation.

Lewis, Gilbert. 1975. *Knowledge of illness in a sepik society: a study of the Gnau, New Guinea*. London: Athlone Press.

Lewis, Paul M. (ed.). 2009. *Ethnologue: Languages of the world*. 16th edn. Dallas: SIL International. .

Lichtenberk, Frantisek. 2008. *A grammar of Toqabaqita* (Mouton Grammar Library 42). 2 vols. Berlin: De Gruyter.

Lincoln, Peter C. 1976. History of research in Austronesian languages: Bougainville Province. In Wurm, *Austronesian Languages*, 197-222.

Lithgow, David. 1976. History of research in Austronesian languages: Milne Bay Province. In Wurm, *Austronesian Languages*, 157-170.

Lynch, John & Terry Crowley. 2001. *Languages of Vanuatu: A new survey and bibliography* (Pacific Linguistics 517). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Marsden, William. 1834. On the Polynesian, or East-Insular languages. In *Miscellaneous Works*, 65-65. London: Parbury, Allen and Co.

Moore, Denny. 2007. Endangered languages of lowland tropical South America. In Matthias Brenzinger (ed.), *Language Diversity Endangered* (Trends in Linguistics: Studies and Monographs), 29-58. Berlin: De Gruyter.

[Max Moszkowski]. 1913. Wörterverzeichnisse von Papua-Sprachen aus holländisch-Neuguinea. *Anthropos* VIII. 254–259.

Moxness, Mike. 1998. A brief, second-hand report on the Kehu (Keu?). 24 words from the memory of a non-native speaker. Unpublished ms.

Nekitel, Otto. 1985. Sociolinguistic Aspects of Abu', a Papuan Language of the Sepik Area, Papua New Guinea. Canberra: ANU PhD thesis.

Pawley, Andrew, Robert Attenborough, Jack Golson & Robin Hide (eds.). 2005. *Papuan pasts: Studies in the cultural, linguistic and biological history of the Papuan-speaking peoples* (Pacific Linguistics 572). Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University.

Ray, Sidney H. 1893. The languages of the New Hebrides. *Journal and Proceedings of the Royal Society of New South Wales* XXVII. 101–167, 469–470.

Ray, Sidney H. 1912. Notes on languages in the east of Netherlands New Guinea. In A. F. R. Wollaston (ed.), *Pygmies and Papuans: The stone age to-day in Dutch New Guinea*, 322-345. London: Smith, Elder & Co.

Ray, Sidney H. 1913-1914. The languages of the Papuan gulf district, Papua. *Zeitschrift für Kolonialsprachen* IV. 20–67.

Ray, Sidney H. 1919. The languages of northern Papua. *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 49. 317–341.

Ray, Sidney H. 1919/1920. The Polynesian languages in Melanesia. *Anthropos* 14/15. 46–96.

Ray, Sidney H. 1923. The Languages of the Western Division of Papua. *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 53. 332–360.

Ray, Sidney H. 1926. *A comparative study of the Melanesian island languages*. Cambridge: Cambridge University Press.

Ray, Sidney H. 1929. The languages of the Central Division of Papua. *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 59. 65–96.

Ray, Sidney H. 1937/1938b. The languages of the Eastern Louisiade Archipelago. *Bulletin of the School of Oriental and African Studies* 9(2). 363–384.

Ray, Sidney H. 1938a. The languages of the Eastern and South-Eastern Division of Papua. *JRAI* 68. 153–208.

Reesink, Ger P. 2002. Mansim, a lost language of the Bird's Head. In Ger P. Reesink (ed.), *Languages of the eastern Bird's Head* (Pacific Linguistics 524), 277-340. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Remijsen, A. C. L. 2002. Word-prosodic systems of Raja Ampat Languages. Leiden: Rijksuniversiteit te Leiden doctoral dissertation.

Robidé van der Aa, Pieter Jan Baptist Carel. 1885. Reizen van D. F. van Braam Morris naar de noordkust van Nederlandsch Nieuw-Guinea eerste vaart op de Amberno- of Rochussen-Rivier. *Bijdragen tot de Taal-, Land- en Volkenkunde van Nederlandsch-Indië, 4e volg., Deel X* 34. 73–114.

Schütz, Albert J. 1972. *The languages of Fiji*. Oxford: Clarendon Press.

Sebeok, Thomas A. (ed.). 1971. *Linguistics in Oceania* (Current Trends in Linguistics 8). Berlin: De Gruyter.

Silzer, Peter J. & Heljä Heikkinen-Clouse. 1991. *Index of Irian Jaya languages* (Special Issue of Irian: Bulletin of Irian Jaya). 2nd edn. Jayapura: Program Kerjasama Universitas Cenderawasih and SIL.

Smits, Leo & C. L. Voorhoeve. 1992a. *The J. C. Anceaux collection of wordlists of Irian Jaya languages A: Austronesian languages* (Part I) (Irian Jaya Source Material No. 4 Series B 1). Leiden-Jakarta: DSALCUL/IRIS.

Smits, Leo & C. L. Voorhoeve. 1992b. *The J. C. Anceaux collection of wordlists of Irian Jaya languages A: Austronesian languages* (Part II) (Irian Jaya Source Material No. 5 Series B 2). Leiden-Jakarta: DSALCUL/IRIS.

Smits, Leo & C. L. Voorhoeve. 1994. *The J. C. Anceaux collection of wordlists of Irian Jaya languages B: Non-Austronesian (Papuan) languages* (Part I) (Irian Jaya Source Material No. 9 Series B 3). Leiden-Jakarta: DSALCUL/IRIS.

Smits, Leo & C. L. Voorhoeve. 1998. *The J. C. Anceaux collection of wordlists of Irian Jaya languages B: Non-Austronesian (Papuan) languages* (Part II) (Irian Jaya Source Material No. 10 Series B 4). Leiden-Jakarta: DSALCUL/IRIS.

Stebbins, Tonya N. 2010. The Papuan languages of the Eastern Bismarcks: migration, origins and connections. In Bethwyn Evans (ed.), *Discovering history through language: Papers in honour of Malcolm Ross* (Pacific Linguistics 605), 223-243. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Sumbuk, Kenneth Memson. 1999. Morphosyntax of Sare. Waikato: University of Waikato PhD thesis.

Taylor, A. J. 1976. History of research in Austronesian Languages: Western part of south-eastern mainland Papua. In Wurm, *Austronesian Languages*, 141-155.

Tryon, Darrell T. & B. D. Hackman. 1983. *Solomon Islands languages: An internal classification* (Pacific Linguistics C 72). Canberra: Research School of Pacific and Asian Studies, Australian National University.

van der Leeden, Alexander Cornelis (see Leeden).

Voorhoeve, Bert. 2000. Een moeilijk begin: Zending, missie en taalwetenschap in Nieuw-Guinea. In Willem van der Molen & Bernard Arps (eds.), *Woord en schrift in de Oost: De betekenis van zending en missie voor de studie van taal en literatuur in Zuidoost-Azië* (Semaian 19), 184-199. Leiden: Universiteit Leiden.

Voorhoeve, C. L. 1975a. Central and Western Trans-New Guinea Phylum Languages. In Wurm, *Papuan Languages*, 345-460.

Voorhoeve, C. L. 1975b. A Hundred Years of Papuan Linguistic Research: Western New Guinea Area. In Wurm, *Papuan Languages,* 117-142.

Wambaliau, Theresia. Forthcoming. Draft Laporan Survei pada Bahasa Mawes di Papua, Indonesia. *SIL Electronic Survey Reports*.

Whitehouse, Paul. n.d. Type-up of anonymous Kehu wordlist from SIL Indonesia (the wordlist presumably comes from Ron Baird in the 1980s). Unpublished ms.

Wilkes, J. R. Adams. 1926. Appendix B: Vocabulary of native languages. *Australian Report on the Administration of New Guinea* 1924-1925. 75–78.

Wurm, Stefan. 1954. Tonal languages in New Guinea and the adjacent islands. *Anthropos* 49(3/4). 697–702.

Wurm, Stephen A. (ed.). 1975. *Papuan languages and the New Guinea linguistic scene* (Pacific Linguistics C 38). Vol. 1 of *New Guinea area languages and language study.* Canberra: Research School of Pacific and Asian Studies, Australian National University.

Wurm, Stephen A. (ed.). 1976. *Austronesian languages* (Pacific Linguistics C 39). Vol. 2 of *New Guinea area languages and language study*. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Z'graggen, John A. 1975. *The Languages of the Madang District, Papua New Guinea* (Pacific Linguistics B 41). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Z'graggen, John A. 1976. History of Research in Austronesian Languages: Madang Province. In Wurm, *Austronesian Languages*, 95-114.

Harald Hammarström
h.hammarstrom@let.ru.nl

Sebastian Nordhoff
sebastian_nordhoff@eva.mpg.de

3

Language Documentation & Conservation Special Publication No. 5 (December 2012)
Melanesian Languages on the Edge of Asia: Challenges for the 21st Century,
ed. by Nicholas Evans and Marian Klamer, pp. 34–71
http://nflrc.hawaii.edu/ldc/sp05/
http://hdl.handle.net/10125/4560

Systematic typological comparison as a tool for investigating language history

Ger Reesink

Max Planck Institute, Nijmegen

Michael Dunn

Max Planck Institute, Nijmegen

Similarities between languages can be due to 1) homoplasies because of a limited design space, 2) common ancestry, and 3) contact-induced convergence. Typological or structural features cannot prove genealogy, but they can provide historical signals that are due to common ancestry or contact (or both). Following a brief summary of results obtained from the comparison of 160 structural features from 121 languages (Reesink, Singer & Dunn 2009), we discuss some issues related to the relative dependencies of such features: logical entailment, chance resemblance, typological dependency, phylogeny and contact. This discussion focusses on the clustering of languages found in a small sample of 11 Austronesian and 8 Papuan languages of eastern Indonesia, an area known for its high degree of admixture.

1. Introduction. The practice of proposing families on the basis of typological comparison is one of the guilty secrets of historical linguistics. It is a basic principle of the historical linguistic tradition that genealogical relationships between languages can only be established by the comparative method, which detects sets of cognates on the basis of regular sound changes and shared irregularities, and thus allows the positing and reconstruction of a proto-language[1]. In spite of this, some early classifications of the more than 800 Papuan languages are based on just a handful of lexical correspondences, supplemented by observations of structural and typological similarities (Greenberg 1971;

[1] The original research conducted for this study was supported by funds from NWO (Netherlands Organization for Scientific Research) for the Program "Breaking the time barrier: Structural traces of the Sahul past" of Professor Pieter Muysken (360-70-210), Radboud University Nijmegen and Professor Stephen C. Levinson, Max Planck Institute for Psycholinguistics, Nijmegen. We thank two anonymous reviewers for comments on an earlier version and Angela Terrill for editorial assistance.

Wurm 1975, 1982). These proposals have been severely criticized (see Pawley 1998, 2005 for a summary), but the influence of typological data at the stage of genealogical hypothesis generation remains.

Typological features of languages are subject to the same evolutionary processes which create genealogical history in other aspects of samples of related languages. There is a tendency for more closely related languages to be more similar on the level of linguistic structure, just as they are more similar in terms of e.g. shared vocabulary. The evolutionary and statistical properties of lexical and sound change have been extensively examined: a great deal is known about what kinds of sound changes are likely, as there is too about what kinds of words tend to be lost, replaced, semantically or phonologically mutated, and so forth. Less is known about the evolutionary and statistical properties of typological/structural features, Even where lexical cognates cannot be identified because of phonological and semantic drift, there remains the possibility that other aspects of language retain traces of the historical relations between languages, whether due to genealogical descent or contact. Area specialists may be able to make generalizations about languages of one or another family on the basis of typological features even where comparative method reconstruction has not been carried out. Hypothesis generation on the basis of structural features of language relies intrinsically on statistical arguments.

As in biology, there are a number of different historical factors that lead languages to be similar: common ancestry, contact (hybridization), and chance convergence (homoplasy). The smaller the design space the higher the probability that convergence is the result of chance rather than geneaological or geographical factors. In biological evolution therefore, the more degrees of freedom in a given domain, the more powerful is the mutation and selection process, resulting in greater disparity and diversity of species. This suggests for linguistic evolution that the greater degree of freedom of lexical elements allows for a more exact measure of phylogenetic relationship on the basis of cognacy sets. Structural features have a much more limited design space, thus convergent evolution will cause homoplasies that need to be distinguished from historical signals, be they phylogenetic or due to hybridization. Large scale chance convergence is less likely, however, when a great number of features are compared, provided these have a measure of independence. See for a more extensive argumentation Dunn et al. (2008:715) where we answer the skepticism expressed by Harrison (2003). We come back to this point in the conclusion.

In this paper we examine the statistical properties of structural features of languages with an eye to their potential in illuminating historical relations. We use the languages of eastern Indonesia, previously identified as an interesting area including both diffusion and inheritance, as a case study. We identify various traits of these languages as present either through diffusion or genetic inheritance.

We adopt a systematic, probabilistic approach using computational models. There are a number of reasons for this, both practical and theoretical. Practically, computational models are able to process a multitude of traits for a great number of languages, while minimizing the apophenic effects of observer preconceptions , where 'apophenic' refers to the human tendency to see meaningful patterns or connections in random or meaningless data. Theoretically, computational models provide us with consistent and testable results, comparable over different hypotheses, and having useful statistical properties such as explicit likelihood scores. A further advantage of computational methods over the

Comparative Method is that the former approach allows hypothesis generation and testing in a way not possible with the Comparative Method. We show that while the Comparative Method illuminates genealogy, structural features can illuminate a long-term history of contact.

The use of structural data in phylogenetic inference has been applied in a few earlier studies which are summarized in section 2. In section 3 we discuss the number and nature of structural features that have been used in those studies. In particular, we pay attention to the issue of trait independency. Section 4 presents the results of a small-scale study, illustrating how structural features provide some clusterings in a set of genealogically diverse Austronesian and Papuan languages of eastern Indonesia. Here we attempt to distill which set of features contributes most strongly to the clusterings. The conclusion in section 5 summarizes discoveries and remaining issues of a standardized approach to typological comparison.

2. Previous studies employing structural features. The use of structural data in phylogenetic inference has been applied in an investigation into the relationships between twenty-two languages of the Oceanic subgroup of the Austronesian family and fifteen Papuan languages of Island Melanesia, reported in two publications (Dunn et al. 2005, Dunn et al. 2008). Although the Papuan languages of this sample had been claimed to form a genealogical group (the East-Papuan phylum, see Wurm 1975), this genealogical unity had been challenged by Ross (2001) and Dunn et al. (2002).

Dunn et al. (2005) used a maximum parsimony analysis of the distribution of 125 abstract structural features and found a reasonable congruence between the consensus tree and the traditional classification of the Oceanic languages in their sample, while the Papuan tree showed some geographic clustering, possibly reflecting ancient relationships (due to inheritance or diffusion through contact). For a critical debate on the merits of that study see Donohue and Musgrave (2007) and Dunn et al. (2007). Croft (2008:230) remarks, "although the result from Dunn et al. (2005) is surprising to a historical linguist, it may be that a cluster of typological traits will provide more precision in classification than will individual traits; also some typological traits are quite stable and therefore may be useful indicators of phylogeny."

Dunn et al. (2008) explained various computational methods in more detail, showed how they can be extended and refined and explored how a phylogenetic signal can be distinguished from possible contact. That study used a Bayesian algorithm to carry out a phylogenetic analysis on a set of 115 abstract phonological and grammatical features. While a certain degree of possible admixture of structural features was detectable between some Oceanic and some Papuan languages, the overal clustering of the languages distinguished the Papuan languages from the Oceanic languages, and the Papuan languages could be clustered into three (geographically, archaeologically) plausible subgroups. The clustering of the Papuan languages into three groups was shown not to be the result of degrees of contact with Oceanic languages, leaving as the most plausible hypothesis that the historical signal found on the basis of structural features is most likely due to a common ancestry, ancient contact between Papuan lineages, or both.

One of the questions raised by these studies (Dunn et al. 2008:737) was how the eastern Papuan languages of Island Melanesia would cluster if a much greater sample

of Papuan languages were investigated. In their critique on Dunn et al. (2005) Donohue and Musgrave (2007:11) "proposed that comparison with Austronesian languages should include representative Austronesian languages from beyond Island Melanesia, in order to obtain an idea of the degree of diversity of these features that can be expected in a family over a 10,000 year (in the Austronesian case, 6,000 year) time frame."

For a follow-up study designed to apply the structural method to a much larger sample of languages, the set of structural features was critically reviewed, revised and expanded. See below for a comparison of some revised questions and the Appendix for both questionnaires.

In the second study (Reesink et al. 2009) we compared a large sample of 121 languages from the Sahul region (i.e. New Guinea and Australia), made up of 55 Papuan, 17 Australian and 48 Austronesian languages, and one Andamanese language, using the revised and expanded set of 160 structural features. Since the linguistic situation of Sahul is complex, combining great time depth with long-term and intensive contact situations, we used a Bayesian algorithm originally developed to discover population structure on the basis of recombining genetic markers, i.e. a model of inheritance and admixture. The Structure algorithm (Pritchard et al. 2000) models evolutionary change and admixture and simultaneously determines both the most likely number of ancestral groups and the most likely contribution of each of these ancestral populations to each of the observed individuals (in this case, languages). The results of Reesink et al. (2009) study suggest 10 ancestral linguistic populations, some of which largely correspond to clearly defined or proposed phylogenetic groups (see figure 1), while others exhibit a high degree of hybridization. Where there are very different degrees of hierarchical relatedness the inferred populations may be nested within known genealogical groupings. The 10 ancestral populations inferred by the structure algorithm can be characterized as follows:

The **Austronesian family** is captured by three groups:

dark green	The Austronesian languages of Borneo and the Phillipines
pale blue	Oceanic languages of mainland New Guinea, New Britain, and Vanuatu
dark purple	All other Oceanic languages of the sample

The Tsou language of Taiwan is equally related to the *dark green* and *dark purple* groups

Other major families

dark blue	Trans-New-Guinea (note that this does not include some of the languages hypothesised to belong to the TNG periphery, such as the Alor-Pantar languages)
light green	Pama-Nyungan languages

Areal groupings

light orange	Non-Pama-Nyungan languages
dark orange	North coast Papuan
light purple	South coast Papuan

pink	East Papuan (plus Bukiyip and Yimas in the north of New Guinea)
red	West Papuan (the Alor-Pantar languages, plus some difficult to classify languages of Halmahera)

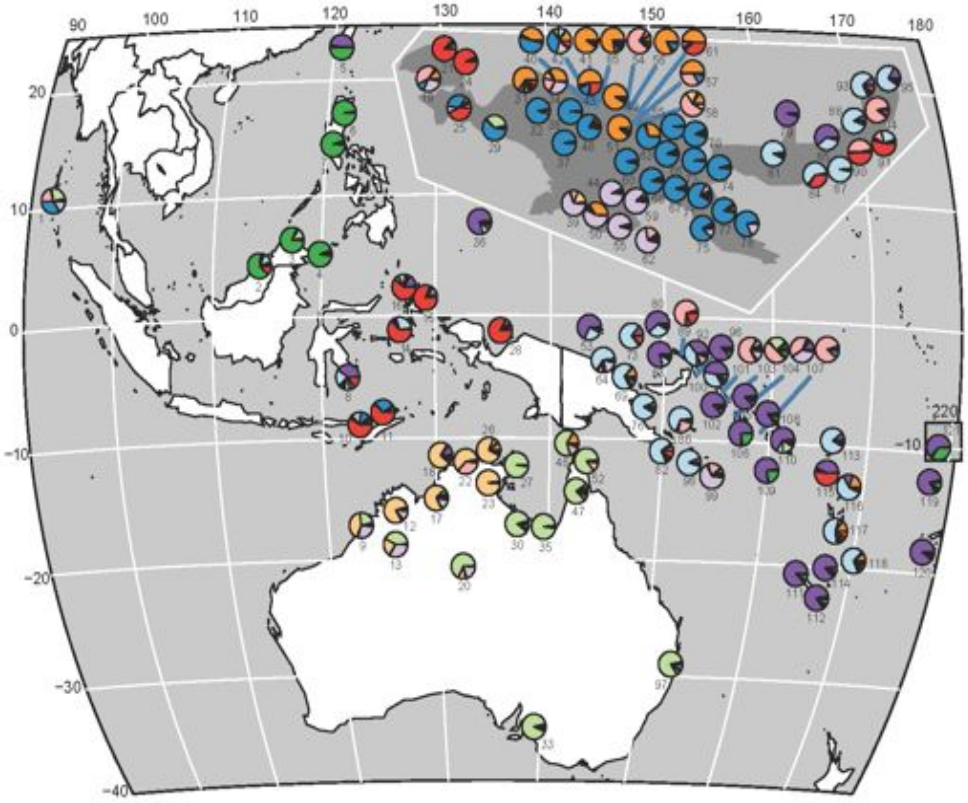


FIGURE 1. The geographic patterning of Structure results for 10 founding populations (Reesink et al. 2009). The pie charts indicate the proportional contribution of each of the founding populations to each language. Languages are identified by number:

Legend

(Fuller details of the interpretation of each population are given in Reesink et al. 2009: 4-7.)

1. Onge [oon]	42. Imonda [imn]	83. Tungag [lcm]
2. Belait [beg]	43. Isaka [ksi]	84. Mangseng [mbh]
3. Kimaragang [kqr]	44. Arammba [stk]	85. Nakanai [nak]
4. Sama [ssb]	45. Namia [nnm]	86. Kilivila [kij]
5. Tsou [tsu]	46. Telefol [tlf]	87. Mengen [mee]
6. Ilocano [ilo]	47. Kuuk Thayorre [thd]	88. Meramera [mxm]

7. Tagalog [tgl]
8. Muna [mnb]
9. Bardi [bcj]
10. Klon [kyo]
11. Abui [abz]
12. Ngarinyin [ung]
13. Gooniyandi [gni]
14. Taba [mky]
15. Tidore [tvo]
16. Tobelo [tlb]
17. Murrinhpatha [mwf]
18. Tiwi [tiw]
19. Inanwatan [szp]
20. Warlpiri [wbp]
21. Meyah [mej]
22. Mawng [mph]
23. Bininj Gun-wok [gup]
24. Hatam [had]
25. Mairasi [zrs]
26. Burarra [bvr]
27. Djambarrpuyngu [djr]
28. Biak [bhw]
29. Kamoro [kgq]
30. Garrwa [gbc]
31. Bauzi [bvz]
32. Nggem [nbq]
33. Ngarrinyeri [nay]
34. Orya [ury]
35. Kayardild [gyd]
36. Ulithian [uli]
37. Korowai [khe]
38. Una [mtg]
39. Marind [mrz]
40. Menggwa Dla [kbv]
41. Abau [aau]
48. Kala Lagaw Ya [mwp]
49. Mende [sim]
50. Gizrra [tof]
51. Yessan-Mayo [yss]
52. Uradhi [urf]
53. Wuvulu-Aua [wuv]
54. Bukiyip [ape]
55. Bine [bon]
56. Ambulas [abt]
57. Alamblak [amp]
58. Yimas [yee]
59. Kiwai Southern [kjd]
60. Kewa [kew]
61. Kamasau [kms]
62. Meriam Mir [ulk]
63. Kobon [kpw]
64. Manam [mva]
65. Usan [wnu]
66. Tauya [tya]
67. Yagaria [qgr]
68. Hua [ygr]
69. Takia [tbc]
70. Waskia [wsk]
71. Menya [mcr]
72. Nabak [naf]
73. Kele [sbc]
74. Selepet [spl]
75. Koiari [kbk]
76. Yabem [jae]
77. Korafe [kpr]
78. Umanakaina [gdn]
79. Bali [bbn]
80. Mussau [emi]
81. Kove-Kaliai [kvc]
82. Gapapaiwa [pwg]
89. Kuot [kto]
90. Kol [kol]
91. Sulka [sua]
92. Madak [mmx]
93. Tolai [ksd]
94. Mali [gcc]
95. Duke of York [rai]
96. Siar [sjr]
97. Bandjalang [bdy]
98. Sudest [tgo]
99. Yélî Dnye [yle]
100. Halia [hla]
101. Rotokas [roo]
102. Banoni [bcm]
103. Motuna [siw]
104. Bilua [blb]
105. Sisiqa [qss]
106. Roviana [rug]
107. Lavukaleve [lvk]
108. Kokota [kkk]
109. Rennellese [mnv]
110. Longgu [lgu]
111. Cèmuhî [cam]
112. Xârâcùù [ane]
113. Aiwoo [nfl]
114. Iaai [iai]
115. Buma [tkw]
116. Mwotlap [mlv]
117. South Efate [erk]
118. Sye [erg]
119. Rotuman [rtm]
120. Fijian [fij]
121. Marquesan [mrq]

Among the conclusions to be drawn from the Reesink et al 2009 study are:

- Structural features of language can be used to help clarify historical relationships.
- In the study, large known groups of languages are recapitulated:

  - The Austronesian family with Oceanic as subgroup
  - The putative Trans New Guinea family, as proposed by Ross (2005), appeared as a solid block with the exception of the Alor-Pantar languages Klon and Abui and the Marind family (Marind and Inanwatan), separated from various non-TNG clusters
  - Australian languages are separated in Pama-Nyungan versus a non-PN cluster.
- However, some clusters represent hybridization rather than phylogeny, especially the cluster containing both Papuan and Austronesian languages of eastern Indonesia.

Some important questions remain: which features are responsible for the clustering? To what extent are structural features independent? Is it possible to distinguish phylogeny from lateral transfer? The issue of relative (in)dependence of structural features will be addressed in section 3 and in section 4 we will take a closer look at the hybrid cluster of eastern Indonesia identified above, applying the Structure algorithm to a new sample of Austronesian and Papuan languages of that area.

3. Relatively (in)dependent traits. After chance resemblance of features (due to the limited design space of language structure at the level of granularity that we have data for; see Dunn et al. 2005, Dunn et al. 2008 and Reesink et al. 2009), the main factors leading to resemblances between languages can be divided into two groups. Firstly, there are factors indicative of historical signal. These include shared inheritance from a common ancestral language, and diffusion through contact between speakers of different linguistic communities. Secondly there are factors which, while in some cases historically determined, do not allow us to infer individual language histories. These include logical entailment, typological dependency (implicational universals), and functionally motivated similarities due to system constraints (Croft 2008:230). For the purposes of making historical inferences about languages, this second set of factors acts as noise at best (obscuring a signal where present), and is misleading at worst (creating the appearance of a signal where one is absent). This is not to say that these factors are intrinsically bad for linguistic analysis: for making historical inferences about typological features this is exactly reversed. In an investigation of implicational universals shared history is the confound (see Dunn et al. 2011).

3.1. Establishing a set of structural features. For the original questionnaire used by Dunn et al. (2005), features were selected on the basis of what in the literature (Dunn et al. 2002; Foley 1998, 2000; Lynch et al. 2002) was known as typical or common characteristics of various Austronesian and Papuan lineages. Some improvements on that set was done for Dunn et al (2008:731), in part in response to commentary in Donohue and Musgrave (2007); see also Dunn et al. (2007). But at the start of the study reported in Reesink et al. (2009) we carried out a major overhaul of the questionnaire in consultation with colleagues (acknowledged in Reesink et al. 2009). Many questions were better defined, a number of questions were removed and others were added. In table 1 and table 2 we give some examples of original questions which could not easily be answered for many languages and which were replaced by questions whose terms were better defined

and more easily identified in a given description.

The questions whether there are adjectives and how they function attributively and predicatively caused some difficulties in the first version. This was solved by the new formulations, which specifically are meant to capture whether adjectival notions are nouny or verby in a particular language.[2]

	ADJECTIVES 2005/2008 [LANGUAGE]		ADJECTIVES 2006/2009 [PLOS BIOLOGY]
40	Is there lexical overlap between a significant proportion of adjectives and verbs (including zero-derivation)?	69	Do core adjectives (defined semantically as property concepts; value, shape, age, dimension) act like verbs in predicative position?
41	Does the same lexical set of adjectives function both attributively and predicatively?	70	Do core adjectives (defined semantically as property concepts; value, shape, age, dimension) used attributively require the same morphological treatment as verbs?

Table 1. Questions relating to Adjectives in two versions

The original questionnaire used for Dunn et al. 2005 and 2008 contained a number of questions attempting to collect data on Tense-Aspect-Mood categories. Those questions were phrased in terms of "how many pure tenses are distinguished?" and "how many fused tense/mood categories are distinguished?" It was stipulated to " include affixes, clitics and satellite particles associated with verbs forming a constituent with the verb on some level, but exclude optional adverbials". Since the terms 'pure' versus 'fused' are not easily interpreted and because the answers were not binary as they are for all other traits, these questions weren't even used for those studies.

Thus, only the few questions in column 2 in table 2 were part of the analyses in the two studies, which meant that potentially important information regarding Tense marking could not be used. The revised questions in table 2 yield more clearly interpretable codes, and they restrict the traits to clearly morphological categories marked on the verb.

[2] As acknowledged in Reesink et al (2009:9), for comments and additions resulting in the latest version we thank Sjef Barbiers, Milly Crevels, Nick Evans, Rob Goedemans, Eva Lindström, Pieter Muysken, Gunter Senft, Leon Stassen, and Hein van der Voort (Workshop 15 May 2006, Radboud University and Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands). In particular the reformulation of the questions regarding adjectives is due to Leon Stassen.

	TAM 2005/2008 [LANGUAGE]		TAM 2006/2009 [PLOS BIOLOGY]
46	Do the same morphemes systematically encode both TAM and person?	79	Do verbs have prefixes/proclitics, other than those that ONLY mark A, S or O (do include portmanteau: A & S + TAM)?
47	Do verbs have prefixes/proclitics?	80	Do verbs have suffixes/enclitics, other than those that ONLY mark A, S or O (do include portmanteau: A & S + TAM)?
48	Do verbs have suffixes/enclitics?	81	Can infixation be used on verbs for derivational, aspectual, or voice-changing purposes?
		82	is there present tense regularly morphologically marked on the verb?
		83	is there past tense regularly morphologically marked on the verb?
		84	is there future tense regularly morphologically marked on the verb?
		85	are there multiple past or future tenses, distinguishing distance from Time of Reference, marked on the verb?
49	is a distinction between punctual/ continuous aspect available as a morphological choice?	86	is a distinction between punctual/ continuous aspect available as a morphological choice?
50	is a distinction between realis/ irrealis mood available as a morphological choice?	87	is a distinction between realis/irrealis mood available as a morphological choice?

TABLE 2. Features relating to Tense-Aspect-Mood affixation in two versions

For a full comparison of the differences between the two versions we refer to the Appendix. We continue with a discussion of the relative (in)dependence of traits in the most recent questionnaire.

3.2. LOGICAL ENTAILMENT. In spite of our attempt to minimize logical entailment between features in our database, there are some cases where we judge it innocuous to allow features with some degree of logical dependency between them to remain. For example, consider the possible values for two questions relating to the phonotactics of a language in (1):

(1) (a) Are there word-final consonants?

(b) Are there consonant clusters (not counting prenasalized consonants) in syllable coda?

The two questions are clearly not totally independent from each other, as particular values of certain features logically entail particular values of others:

if (a) = 1, then (b) = 1 or 0; if (a) = 0, then (b) = 0.
if (b) = 1, then (a) = 1; if (b) is 0, then (a) = 1 or 0.

However, the bias added by this dependency is small as this entailment is only partial, outweighed by the added statistical power we get from including data with the logically independent values. Given the large number of features in our analysis, it is not likely that this one case of partial dependency has seriously affected the results in our earlier analyses.

3.3. Chance resemblance due to limited design space. Most or all of the structural features of language have a far more restricted degree of freedom than lexical items. They are a fundamentally different kind of data with different statistical properties. For example, the two questions about the behavior of adjectival elements in predicative and attributive position (see table 1) were formulated to capture whether a language has verby (Y to both questions) or nouny (N to both) adjectives, or in between (Y to predicative; N to attributive verb-like behavior). A language which would have N to verb-like behavior in predicative position, but Y to verb-like behavior in attributive position was considered as unlikely. However, in our sample we do find this anomalous situation in the non-TNG language Imonda. Thus, the maximum number of four possibilities is available. This holds also for the two questions whether a language has prepositions or postpositions. There are languages with Y or N to both questions in addition to those that have only one or the other.

With regard to the order of Possessor and Possessum, the design space allows for three possibilities: the Possessor may (1) precede or (2) follow or (3) may do both. A negative value of both questions is of course not possible.

While such limited degrees of freedom may create homoplasies that do not reflect shared history, large-scale chance convergence is rendered unlikely through the use of a large number of features.

3.4. Typological dependency – implicational universals. Typological dependencies have been widely discussed since the sixties when Joseph Greenberg launched his language universals project. Most generalizations deal with word order properties in the clause and the nominal constituent. For example, it is well-known that OV order and postpositions are commonly found together, as are VO order and prepositions. Dunn et al (2011) has argued that there is a strong lineage-specific element to these apparent universals. Dryer (2005) presents data showing that the correlation is not perfect. Of a total of 1033 languages, 427 have OV and postpositions and 417 have VO and prepositions, while 10 languages combine OV with prepositions and 38 have VO together with postpositions. In addition, 141 languages do not fall into one of these four categories. For example, Dutch has prepositions but has both OV and VO order. On the other hand, Jabêm has SVO order with both prepositions and postpositions. Thus, while there is a strong typological tendency for the values of these features to be correlated, by removing some of these questions

important information is lost.

There are indeed rather high correlations between the questions on tense marking in table 2. However, combining past and future tense as reported by Dahl and Velupillai (2005; chapters 66 and 67 in WALS), there is no clear typological dependency cross-linguistically: of the 110 languages that mark future tense, there are 48 that mark a simple past tense, 26 with 2-3 degrees of remoteness, 1 with 4 or more degrees, and 35 with no past tense marking. In other words, if some of these questions were removed a considerable amount of information would be lost.

3.5. Functionally motivated – system constraints. Somewhat related to typological dependency is convergence due to system constraints. Some of our features may at first blush be mutually exclusive or inclusive. For instance, languages tend to have prepositions or postpositions, but relatively infrequently have both or neither. The raw counts for these features in our complete database (ignoring for the moment that these observations are phylogenetically dependent) are shown in table 3. The conditional probability of having prepositions given postpositions is 15%, and the conditional probability of having postpositions given prepositions is only 11%. A diachronic account for the development of adpositions predicts that the order of adposition and noun phrase will typically be fixed.

		Postpositions	
		Present	Absent
Prepositions	Present	11	87
	Absent	61	13

Table 3. Postpositions and prepositions.

Heine and Kuteva (2007) describe typical grammaticalization pathways such as *relational noun>adposition, adverb>adposition,* or *verb+complement>adposition+noun phrase,* which each have as their starting point a construction which most commonly already has fixed ordering. Even if two orders of adpositions and noun phrases are possible, the order will most likely be fixed with respect to the particular adposition selected. Given the constraint on adposition systems that there will usually be only one kind, there is a negative correlation between having prepositions and having postpositions.

Similar kinds of system constraints exist in other parts of the grammar. For example, there is a tendency for agreement affixes for transitive and intransitive subjects to be marked the same way. Thus, there is a positive correlation between having prefixes for marking transitive subjects (A) and intransitive subjects (S), and likewise there is a positive correlation between having suffixes for subjects of transitive (A) and subjects of intransitive clauses (S), as shown in table 4.

		S suffix	
		Present	Absent
A suffix	Present	11	4
	Absent	7	66
		S prefix	
		Present	Absent
A prefix	Present	35	3
	Absent	3	47
		O suffix	
		Present	Absent
O prefix	Present	1	0
	Absent	51	34

TABLE 4: A, S and O as prefixes and suffixes.

Some of these tendencies are nevertheless not strong. While there is a negative correlation between having object suffixes (O) and having object prefixes (see table 4), the amount of the variance this correlation explains of the (phylogenetically uncorrected) data is barely significant. This is despite a strong phylogenetic bias, in that no Austronesian languages have an object prefix, and most Trans New Guinea languages do. This would be expected to have the effect of exaggerating the apparent negative correlation between object prefixes and suffixes.

3.6 SHARED INHERITANCE. Correlations between features in a linguistic data set cannot be interpreted as causal with any validity without taking into account the confound introduced by possible genealogical relationships between the languages. This issue is known as Galton's problem: variables in languages related by common descent or diffusion are not statistically independent. Any apparent causal correlations between features of languages linked by shared history might be no more than 'duplicate copies of the same original' (Galton in Tylor 1889:270). This was alluded to above (section 3.4), with the example of object prefixes. Object prefixes are absent in Austronesian languages and highly frequent in Trans New Guinea languages. So, any other feature which is rare in Austronesian and common in TNG will correlate with the presence or absence of object prefixes. In a data set limited to Austronesian and TNG languages we could expect absurd correlations, such as positive correlations between object prefixes and altitude, negative correlations between object prefixes and navigational technology. These correlations are driven by accidents of history rather than any causal link.

In the full set of features we find substantial correlations, either positive or negative, which are clearly the product of shared history. For instance, there are positive correlations between Decimal counting systems, Prepositions, and the Inclusive/Exclusive distinction for non-singular first person. Likewise there is a negative correlation between these features

and verbal past tense.

Figure 2 illustrates the accidental nature of the negative correlation between decimal counting systems and verbal past tense marking. A decimal system of counting predominates in Austronesian, although there are quite a number of AN languages which exhibit a quinary system. And there is internal evidence in many of the Papuan languages with decimal systems that this occurred through contact with Austronesian speaking communities. The left panel of figure 2 shows a possible reconstruction of the history of decimal counting systems in a sample of Austronesian languages. The case for reconstructing decimal systems for proto-Austronesian seems strong. There are two sub-branches of languages lacking decimal counting systems and decimal counting systems occur on every level of the phylogeny. The right panel shows a somewhat different story. Verbal past tense marking occurs sporadically throughout the tree, but there are no cases where it makes sense to reconstruct past tense marking to an earlier node of the tree. The negative correlation between these features is apparently because of the relative stability of the two features, and their states in the ancestral languages. In other words, the clustering of such statistically dependent features is due to a genealogical signal.

The final possible cause of typological similarity between languages, diffusion through contact, will be discussed in section 4.

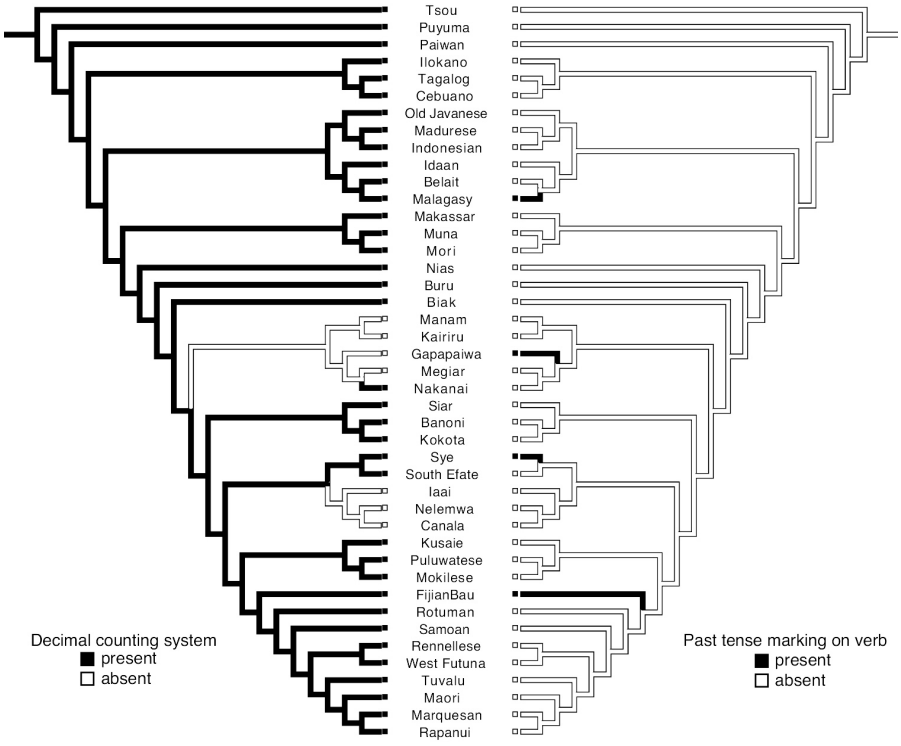


FIGURE 2. Decimal counting systems and past tense marking on the verb in Austronesian tree

4. Linguistic Populations in east Indonesia. The study examining linguistic traces of the Sahul Past (Reesink et al. 2009) employed the STRUCTURE algorithm (Pritchard et al. 2000), as shown in section 2. The method assumes a model in which there are a number (K) of unspecified or unknown populations, each of which is characterized by a set of allele frequencies at each locus. Individuals in any sample are assigned (probabilistically) to populations, or jointly to two or more populations if their genotypes indicate that they are admixed. The different values of the linguistic characters are the analogical equivalent of the genetic alleles, while a language is the equivalent of an individual in the biological studies. In other words, just as an individual's autosomal DNA is inherited from a number of different ancestors belonging to one or more biological populations, so a language may have inherited structural features from one or more different populations. The structure algorithm computes the most likely contribution of a given number (K) of ancestral populations to each of the individuals.

As stated at the end of section 2, Reesink et al. (2009) did find some striking correspondence between earlier defined linguistic families. However, as already mentioned, structural features cannot be used to claim or refute genealogical relationships between languages, see also Croft (2004). This is illustrated in the fact that the striking correspondence does not amount to full agreement among the groupings found by the different methods. A rather robust linguistic population identified by the Structure algorithm (Reesink et al. 2009) as the 'red' or 'West Papuan' cluster (see figure 1) contains all the Papuan languages of eastern Indonesia and the Bird's Head in the sample: Klon and Abui from the Alor-Pantar family, Tobelo and Tidore from North Halmahera, and Meyah and Hatam from the Bird's Head, as well as the two AN languages Taba and Biak. This cluster has also contributions to Papuan and Austronesian languages along the north coast of New Guinea and in the Bismarck archipelago. We concluded in that study: "This finding suggests an area of millennia of contact between AN and Papuan non-TNG speaking groups" (Reesink et al. 2009:8).

Given that earlier studies had shown a great degree of heterogeneity among the Papuan groups in east Indonesia (see for example Reesink 2005), it was rather surprising to see them clustered together with a few AN languages thrown in. Thus, new research questions are raised: 1) which features are responsible for a certain clustering; and 2) is it possible to differentiate phylogeny and diffusion?

In order to answer these questions a new study was conducted with two more AN languages from the same region, Tetun spoken in East Timor and Buru of the Moluccas, both classified as members of the Central Malayo-Polynesian subgroup. The validity of this subgroup, proposed by Blust as a linkage (1993), has been challenged by Donohue and Grimes (2008) and reaffirmed as most likely descending from a dialect chain by Blust (2009). We now report the results of this new study.

The Structure algorithm was applied this time to just a small sample of Austronesian languages of (eastern) Indonesia and Papuan languages of the same area. Since the algorithm simultaneously determines both the most likely number of ancestral groups and the most likely contribution of each of these populations to each of the observed individuals, we wanted to focus on the similarities and differences between just these languages, avoiding clustering that might ignore intragroup differences when compared to Papuan and Australian languages with different profiles, as was done in the major studies.

In figure 3 the clustering of these languages is shown for two to five ancestral populations (K2-5). For each specified number of clusters (K), the algorithm assigns a certain weight to each allele (in our case, the state of a particular feature). This clustering is independent for each K value, so that individuals may be assigned to different clusters (arbitrarily given a particular colour) on the basis of the amalgamated weights of the feature-states within each K.

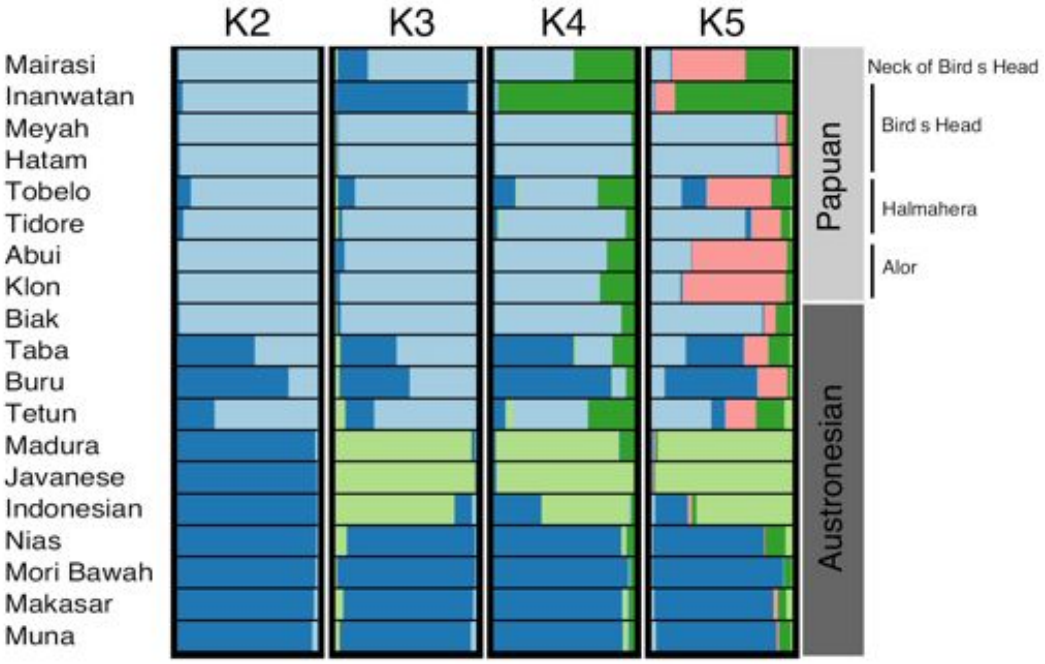


Figure 3. Clustering of AN and Papuan languages of eastern Indonesia

The K values 3, 4, and 5 hardly differ in their likelihood score. At K3 and K4 the light blue cluster contains AN and Papuan languages, while at K5 we find some differentiation. A new cluster (pink) is detected contributing mainly to Klon and Abui of the Alor-Pantar group, Tobelo of North Halmahera and Mairasi, spoken in the 'neck' connecting the Bird's Head to the rest of New Guinea. Thus at K5 we find a separation of a number of the Papuan languages, but still not all.

Clustering by the Structure algorithm is based on differential weighting to each of the 160 features per cluster. The same feature may have a higher or lower weighting for different K values. Space does not allow us to give a full list of different weights of each feature for each value of K, but in table 5 a sample pertaining to word order is given. These values show that presence of V final, Postpositions, and Object Prefix have a lower weight for the light blue cluster at K4 and thus, together with the values of other features, cannot differentiate Klon, Abui, Tobelo and Mairasi from the other Papuan and AN languages. At K5 these features have a stronger weighting, and thus a new cluster is identified.

Light blue	**K4**	**K5**	**Pink K5**
Verb final	0.38	0.24	0.61
Postposition	0.31	0.21	0.49
Object Prefix	0.38	0.22	0.56
Verb medial	0.64	0.79	0.45
Preposition	0.72	0.85	0.56

TABLE 5. Allele weights for features in contributing populations

Some values of the features in table 5 may look like a system constraint, or a typological correlation, but the overall correlation between Object prefix and Verb-final word order in the sample of 121 languages is rather weak (r = 0.40). The contribution of these features is therefore relatively independent.

In figure 3 it is clear that in all independent runs at all K values, the two unrelated Papuan languages of the Bird's Head, Hatam and Meyah, consistently cluster with the AN language Biak. It thus appears that in this case diffusion overrides phylogeny.

Is it possible to differentiate the two historical processes by extant structural features? We know from the comparison of their lexicons that Biak belongs to the South Halmahera-West New Guinea subgroup of the AN family and that Hatam and Meyah belong to two different Papuan families, albeit with perhaps a very remote common ancestor (Reesink 2002). Are there any traces in their structural features that still betray their genealogical affiliation? In other words, to what extent are these languages different in the set of structural features employed?

In order to find such traces we have to go into the nitty-gritty of the data. Table 6 lists all fifteen features (out of 160) on which the two unrelated Papuan languages Hatam and Meyah both agree with each other, presumably due to shared diffusion of Papuan traits, and are different in value from Austronesian Biak.

	Hatam	Meyah	Biak
Weight sensitive stress	-	-	+
Syllable position stress	-	-	+
Definite/specific articles	-	-	+
Indefinite article required	-	-	+
Difference comitative vs coordination	-	-	+
Gender in third person	-	-	+ (3pl.animate)
Numeral classifiers	+	+	-
Possession by suffix	-	-	+
Quinary counting system	+	+	-

	Hatam	Meyah	Biak
Attributive adjectives require same morphology as verbs	-	-	+
Copula for predicative N(P)	-	-	+
Aspectual auxiliaries	-	-	+
Causative by Serial Verb Construction	+	+	-
Nouns can be reduplicated	-	-	+
Other elements than N or V can be reduplicated	+	+	-

TABLE 6. Hatam and Meyah values agree and differ from Biak

These facts show very faint traces of structural features that may betray phylogenetic affiliation. For example, possession by suffix seems tightly linked to the AN family. In many AN languages to the north-west of this geographic region the Possessor normally follows the Possessum, and when that is expressed by a pronoun it can easily become encliticized or suffixed. This order is still present in Biak. It should be noted that the feature Possession by prefix (a separate question in our database) is not part of the list separating Hatam and Meyah from Biak, because for this trait all three languages have a positive value. This order is typical of the Papuan languages of the Bird's Head (and other regions of east Indonesia), and has diffused to a few AN languages in the Cenderawasih Bay area, in Biak and Ambai for plural possessors, in Waropen for both singular and plural (see Klamer et al. 2008:129). While all other Papuan languages of North Halmahera and the Bird's Head have a gender distinction for third person singular, the two east BH families that Hatam and Meyah belong to, do not. Yet Biak has adopted this Papuan trait in the form of a gender distinction between animate and inanimate for third person plural pronouns.

Of course, as mentioned above, single structural features can never be diagnostic for genealogical relatedness, and this is illustrated for these heterogeneous languages which have converged to such a degree that even their full structural profile obscures their descent. While the Comparative Method illuminates their genealogy, structural features illuminate their long term history of contact.

5. CONCLUSION. The results of large-scale comparison of structural features in a great number of languages from different lineages can be summarized as follows.

In population genetics the distribution and frequency of mutations in unrelated individuals are used to trace ancestral populations. In the studies reviewed in section 2 we practice *population linguistics*, that is, we attempt to find clusters between individual languages that are NOT immediate family. Where cognate-based methods cannot be applied, profiles of abstract structural features can discover plausible groupings in hitherto unrelated clusters of languages. These groupings may be the result of remote common ancestry, diffusion or both. In the case of a putative family like the Papuan TNG family, the result obtained by structural features may strengthen the tentative conclusions based on

pronominal forms. We do not claim that we now have conclusive evidence for TNG as a bona fide family, but simply that the proposed unity has some firmer footing.

Chance resemblances due to the limited degrees of freedom structural features have (Harrison 2003; section 3.3 above) can to some extent be overcome by considering a large number of features. Typological dependencies such as implicational universals and functionally motivated convergences are an empirical matter: how strong are they? They apparently differ in different lineages (Dunn et al. 2011).

The results reported in section 2 show that a large set of structural features does reveal a phylogenetic signal in that higher level linguistic groupings are identified. Due to their limited design space and relative ease of diffusion they cannot unequivocally identify lower level language families. As shown in section 4, the Structure algorithm cannot separate different lineages in eastern Indonesia, at least not with a strong likelihood. A matter for further research would be to investigate whether a different set of features could do better. It may be that a small set of diagnostic traits is masked by a much larger number of features that are shared by languages of different families by a Bayesian inference algorithm such as Structure, as illustrated for Hatam, Meyah and Biak in section 4.

While structural features can be diffused, complete substitution is quite rare. The basic morpho-syntactic profile, linked to the semantic-pragmatic way of representing the natural and social world of any particular speech community is quite robust through many descending generations. Therefore, the linguistic clusters found on the basis of full profiles provide information about their historical provenance. If it is possible to reconstruct/determine the ancestral state of a particular feature in a (putative) family, as for example shown by the presence of a decimal counting system and absence of past tense marking in the Austronesian family, then aberrant values in daughter languages can be accounted for by hybridization.

References

Blust, Robert. 1993. Central and Central-Eastern Malayo-Polynesian. *Oceanic Linguistics* 32(2). 241-293.

Blust, Robert. 2009. The position of the languages of eastern Indonesia: A reply to Donohue and Grimes. *Oceanic Linguistics* 48(1). 6-77.

Croft, William. 2004. Typological traits and genetic linguistics. http://www.unm.edu/~wcroft/Papers/Typ-Gen.pdf. (5 November, 2010.)

Croft, William. 2008. Evolutionary linguistics. *Annual Review of Anthropology* 37. 219–234.

Dahl, Östen & Velupillai, Viveka. 2011. The Past Tense. In Dryer & Haspelmath, http://wals.info/chapter/66. (8 September, 2012.)

Dahl, Östen & Velupillai, Viveka. 2011. The Future Tense. In Dryer & Haspelmath, http://wals.info/chapter/67. (8 September, 2012.)

Donohue, Mark, & Simon Musgrave. 2007. Typology and the linguistic macrohistory of Island Melanesia. *Oceanic Linguistics* 46(2). 325–364.

Donohue, Mark, & Charles E. Grimes. 2008. Yet more on the position of the languages of Eastern Indonesia and East Timor. *Oceanic Linguistics* 47(1). 114-158.

Dryer, Matthew S. 2011. Relationship between the order of object and verb and the order of adposition and noun phrase. In Dryer & Haspelmath, http://wals.info/chapter/95. (10 September, 2012.)

Dryer, Matthew S. & Martin Haspelmath (eds.). 2011. *The world atlas of language structures online*. http://wals.info/.

Dunn, Michael, Ger Reesink & Angela Terrill. 2002. The East Papuan languages: A preliminary typological appraisal. *Oceanic Linguistics* 41(1). 28–62.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley & Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309. 2072–2075.

Dunn, Michael, Robert A. Foley, Stephen C. Levinson, Ger Reesink & Angela Terrill. 2007. Statistical reasoning in the evaluation of typological diversity in Island Melanesia. *Oceanic Linguistics* 46(2). 388-403.

Dunn, Michael, Stephen C. Levinson, Eva Lindström, Ger Reesink, & Angela Terrill. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language* 84(4). 710-759.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson, & Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473. 79–82.

Foley, William A. 1998. Toward understanding Papuan languages. In Miedema et al., 503–18.

Foley, William A. 2000. The languages of New Guinea. *Annual Review of Anthropology* 29. 357-404.

Greenberg, Joseph H 1971. The Indo-Pacific hypothesis. In Thomas A. Sebeok (ed.), *Linguistics in Oceania* (Current Trends in Linguistics 8), 807–71. The Hague: Mouton.

Harrison, S. P. 2003. On the limits of the comparative method. In Brian D. Joseph & Richard D. Janda (eds.), *The handbook of historical linguistics*, 213–43. London: Blackwell.

Heine, Bernd & Tania Kuteva. 2007. *The genesis of grammar*. Oxford: Oxford University Press.

Klamer, Marian, Ger Reesink & Miriam van Staden. 2008. Eastern Indonesia as a linguistic area. In Pieter Muysken (ed.) *From linguistic areas to areal linguistics*, 95-149. Amsterdam: Benjamins.

Lynch, John, Malcolm Ross & Terry Crowley. 2002. *The Oceanic Languages*. London: Curzon.

Miedema, Jelle, Cecilia Odé & Rien A.C. Dam (eds.). 1998. *Perspectives on the Bird's Head of Irian Jaya.* Amsterdam: Rodopi.

Pawley, Andrew K. 1998. The Trans New Guinea phylum hypothesis: A reassessment. In Miedema et al., 655-690.

Pawley, Andrew. 2005. The chequered career of the trans New Guinea hypothesis: Recent research and its implications. In Pawley et al., 67-107.

Pawley, Andrew, Robert Attenborough, Jack Golson & Robin Hide (eds.). 2005. *Papuan Pasts, Studies in the cultural, linguistic and biological history of the Papuan-speaking peoples*. Canberra: Pacific Linguistics.

Pritchard, Jonathan, Matthew Stephens, & Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155. 945–959.

Reesink, Ger. 2002. *Languages of the eastern Bird's Head*. Canberra: Pacific Linguistics.
Reesink, Ger. 2005. West Papuan languages: Roots and development. In Pawley et al., 185-218.
Reesink, Ger, Ruth Singer & Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population methods. *PloS Biology* 7(11). e1000241.
Ross, Malcolm. 2001. Is there an East Papuan phylum? Evidence from pronouns. In Andrew Pawley, Malcolm Ross and Darrell Tryon (eds.), *The boy from Bundaberg: Studies in Melanesian linguistics in honour of Tom Dutton,* 301-321. Canberra: Pacific Linguistics.
Ross, Malcolm. 2005. Pronouns as a preliminary diagnostic for grouping Papuan languages. In Pawley et al., 15-65.
Wurm, Stephen A. 1975. The East Papuan phylum in general. In Stephen A. Wurm (ed.), *Papuan languages and the New Guinea linguistic scene*, 783-803. Canberra: Pacific Linguistics.
Wurm, Stephen A. 1982. *Papuan languages of Oceania* (Ars Linguistica 7). Tübingen: Gunter Narr.

Ger Reesink
ger.reesink@hccnet.nl

Michael Dunn
michael.dunn@mpi.nl

Appendix

In this table the lists of characters used for Dunn et al. (2008) in *Language* and for Reesink et al. (2009) in *PloS Biology* are compared.

	characters 'Language 2008'		**characters 'PloS 2009'**
		1	Are there as many points of articulation for nasals as there are for stops? *Only consider points of articulation where a nasal is phonetically possible* (1: present, 0: absent)
		2	Is there contrast between heterorganic and homorganic sequence of nasal and velar stop? *For example, does the language permit a phonetic contrast between -nk- and -ŋk- clusters* (1: present, 0: absent)
1	Are there fricative phonemes?	3	Are there fricative phonemes? (1: present, 0: absent)
2	Are there phonemic prenasalised stops?	4	Are there phonemic prenasalised stops? (1: present, 0: absent)
3	Is there a phonemic distinction between l/r?	5	Is there a phonemic distinction between l/r? (1: present, 0: absent)
4	Is there a phonemic velar fricative or glide?	6	Is there a phonemic velar fricative or glide? (1: present, 0: absent)
5	Is there a voicing contrast between oral (i.e. non-prenasal) stops?	7	Is there a voicing contrast between oral (i.e. non-prenasal) stops? (1: present, 0: absent)
		8	Is there a laminal/apical contrast? (1: present, 0: absent)
		9	Are there retroflexed consonants? (1: present, 0: absent)
6	Is there phonemic consonant length?	10	Is there phonemic consonant length? (1: present, 0: absent)
7	Is there phonemic vowel length?	11	Is there phonemic vowel length? (1: present, 0: absent)
8	Are there contrastive phonation types for vowels? (e.g. nasal, creaky, etc)		

	characters 'Language 2008'		**characters 'PloS 2009'**
		12	Are there two or more contrastive central vowels *Do not include length contrasts* (1: present, 0: absent)
9	Is there lexically determined suprasegmental prominence? *suprasegmental prominence can be loudness, duration, pitch, i.e. stress or tone phenomena (don't include phonemic vowel length)*	13	Is there lexically determined suprasegmental prominence? *suprasegmental prominence can be loudness, duration, pitch, i.e. stress or tone phenomena (don't include phonemic vowel length)* (1: present, 0: absent)
		14	Is there weight-sensitive suprasegmental prominence *suprasegmental prominence can be loudness, duration, pitch, i.e. stress or tone phenomena* (1: present, 0: absent)
		15	Is there syllable position sensitive suprasegmental prominence? *suprasegmental prominence can be loudness, duration, pitch, i.e. stress or tone phenomena* (1: present, 0: absent)
		16	Is there a tonal system? *I.e. two or more contrastive tones* (1: present, 0: absent)
10	Are there word-final consonants?	17	Are there word-final consonants? (1: present, 0: absent)
11	Are there consonant clusters?	18	Are there consonant clusters (not counting prenasalized consonants) in syllable onset? (1: present, 0: absent)
		19	Are there consonant clusters (not counting prenasalized consonants) in syllable coda? (1: present, 0: absent)
12	Are there definite or specific articles?	20	Are there definite or specific articles? (1: present, 0: absent)
13	Are there indefinite or non-specific articles?	21	Is an indefinite NP obligatorily accompanied by an indefinite (or non-specific) article? *Disregard if only on personal names* (1: present, 0: absent)
14	Is the order of NP elements Art N?	22	Are there prenominal articles? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
15	Are NPs N-initial (except for articles)?	23	Are there postnominal articles? (1: present, 0: absent)
		24	What is the relative position of numeral and noun in the NP? (*multistate* 1; Num-N; 2: N-Num; 3: both.)
		25	What is the relative position of demonstrative and noun in the NP? (*multistate* 1: Dem-N; 2: N-Dem; 3: both.)
		26	Are there 'discontinuous noun phrases'? *Can an argument be expressed by multiple N/NP throughout the clause > i.e. the Australian type.* (1: present, 0: absent)
		27	Is there a difference between the marking of NP coordination ('John and Mary went to market') and the marking of comitative phrases ('John went to market with Mary')? (1: present, 0: absent)
16	Is there an inclusive/exclusive distinction?	28	Is there an inclusive/exclusive distinction? (1: present, 0: absent)
		29	Is there a minimal-augmented system? *i.e. four basic pronominal forms for 1sg, 2sg, 3sg and 1+2, which each can be affixed for plural (or dual etc.)* (1: present, 0: absent)
		30	Is there a gender distinction in 3rd person pronouns (or demonstratives, if no 3rd person pronouns)? *either two- or threefold* (1: present, 0: absent)
		31	Is there a dual (or unit augmented) in addition to a plural (or augmented) number category in pronouns? (1: present, 0: absent)
17	Are 1st and 2nd persons conflated in any context?	32	Are 1st and 2nd persons conflated in any context? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
18	Are 2nd and 3rd persons conflated in non-singular numbers? (*Morphologically in any paradigm. Disregard pragmatics/politeness*)	33	Are 2nd and 3rd persons conflated in non-singular numbers? *morphologically in any paradigm. Disregard pragmatics/politeness* (1: present, 0: absent)
		34	Are person categories neutralized under some conditions? *e.g. in non-singular, under NEG, in certain TAM* (1: present, 0: absent)
19	Are more than 2 degrees of distance morphologically marked in demonstratives?	35	Is there an opposition between three or more distance terms in the demonstrative system? (1: present, 0: absent)
20	Are any of the spatial demonstratives not speaker-based? *Speaker-based spatial demonstratives are demonstratives that take as their decitic centre the speaker. By contrast, some demonstratives take not the speaker but the addressee as the deictic centre, for example a demonstrative might mean 'close to the speaker'; and some take both speaker and addressee as the deictic centre e.g. 'far from speaker and addressee'.*		
21	Is elevation morphologically marked in demonstratives?	36	Is elevation morphologically marked in demonstratives? (1: present, 0: absent)
		37	Is the opposition visible-non-visible marked on demonstratives? (1: present, 0: absent)
22	Are demonstratives classified?	38	Are demonstratives classified? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
23	Are there declensions (partly) determined by number of the noun? *By noun declensions is meant e.g nouns divided into groups which have formally different sets of morphological marking. Do not include place names which can act as bare adjuncts*	39	Are there declensions (partly) determined by number of the noun? *By noun declensions is meant e.g nouns divided into groups which have formally different sets of morphological marking. Do not include place names which can act as bare adjuncts* (1: present, 0: absent)
24	Are there declensions (partly) determined by gender of the noun? *By noun declensions is meant e.g nouns divided into groups which have formally different sets of morphological marking. Do not include place names which can act as bare adjuncts*	40	Are there declensions (partly) determined by gender of the noun? *By noun declensions is meant e.g nouns divided into groups which have formally different sets of morphological marking. Do not include place names which can act as bare adjuncts* (1: present, 0: absent)
25	Are there nouns which are suppletive for number? (*Only yes if present for more than **2** (basic) kin terms*)	41	Are there nouns which are suppletive for number? *Only answer yes if present for more than 2 (basic) kin terms*
		42	Can singular number be marked on the noun itself? *Number marking on noun does not count phrase level clitic or reduplication; absence of plural marking does not count as singular marking; exclude derivational forms (e.g. deverbal, deadjectival)* (1: present, 0: absent)
26	Can dual number be marked on the noun itself? *Number-marking on N does not count phrase-level clitic or reduplication*	43	Can dual number be marked on the noun itself? *number-marking on N does not count phrase-level clitic or reduplication* (1: present, 0: absent)
		44	Can plural number be marked on the noun itself? *number-marking on N does not count phrase-level clitic or reduplication* (1: present, 0: absent)
27	Is number marking prohibited on certain (types of) nouns? *(do not include proper nouns, e.g. place names or personal names)*	45	Is number marking prohibited on certain (types of) nouns? *(do not include proper nouns, e.g. place names or personal names)* (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
		46	Are there associative plurals? *e.g. Mary-PL = Mary and her family* (1: present, 0: absent)
		47	Is there a productive morphologically marked Action/state nominalization (arrive-arrival)? *if a language is precategorial, include the morphological mechanisms to produce such 'nominalizations'* (1: present, 0: absent)
		48	Is there a productive morphologically marked Agentive nominalization (sing-er)? (1: present, 0: absent)
		49	Is there a productive morphologically marked Object nominalization (sing; song)? (1: present, 0: absent)
28	Are there noun classes/genders? *By noun classes/genders is meant a system of dividing all or almost all of the nouns of a language into morphological classes which determine agreement phenomena beyond the noun itself.*	50	Are there noun classes/genders? *By noun classes/genders is meant a system of dividing all or almost all of the nouns of a language into morphological classes which determine agreement phenomena beyond the noun itself.* (1: present, 0: absent)
		51	Is sex a relevant category in the noun class/gender system? (1: present, 0: absent)
		52	Is shape a relevant category in the noun class/gender system? (1: present, 0: absent)
		53	Is animacy (without reference to sex) a relevant category in the noun class/gender system? (1: present, 0: absent)
		54	Is plant status a relevant category in the noun class/gender system? (1: present, 0: absent)
		55	Does the language only have a gender distinction in 3rd person pronouns? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
		56	Is there concord within the NP, i.e. agreement of elements within the NP with the noun class of a noun? *related to class/gender* (1: present, 0: absent)
29	Are there numeral classifiers? *i.e. free or bound morphemes which are non-agreeing, noun categorization devices, the choice of which is determined by lexical selection*	57	Are there numeral classifiers? *i.e. free or bound morphemes which are non-agreeing, noun categorisation devices, the choice of which is determined by lexical selection* (1: present, 0: absent)
30	Are there possessive classifiers? *i.e. free or bound morphemes which are non-agreeing, noun categorisation devices, the choice of which is determined by lexical selection*	58	Are there possessive classifiers? *i.e. free or bound morphemes which are non-agreeing, noun categorisation devices, the choice of which is determined by lexical selection* (1: present, 0: absent)
31	Are there possessive classes? *i.e. different nouns treated differently in possession according to semantically-based groupings. Include alienable/inalienable*.		
32	Is alienable/inalienable a relevant distinction?	59	Is alienable/inalienable a relevant distinction? (1: present, 0: absent)
33	Are there different possessive constructions?	60	Are there different possessive constructions? (1: present, 0: absent)
		61	Can possession be marked by a prefix? *even if only on a restricted numer of kin terms. Emphasis is on *can** (1: present, 0: absent)
		62	Can possession be marked by a suffix? *even if only on a restricted numer of kin terms. Emphasis is on *can** (1: present, 0: absent)
34	Can possession be marked on the nominal possessor?	63	Can possession be marked on the nominal possessor? (1: present, 0: absent)
35	Can possession be marked on the nominal possessee?	64	Can possession be marked on the nominal possessee? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
36	If the order of elements in a possessive construction is fixed, is it possessor-possessed?	65	What is the relative position of possessor and possessed in the attributive possessive construction? (*multistate* 1:Possessor-Possessed; 2:Possessed-Possessor; 3: both)
		66	Are there different orders of elements in a possessive phrase for different classes of possession? *emphasis on *for different types of possession** (1: present, 0: absent)
37	Is there a decimal counting system? (*i.e. elements of decimal; even lexical **10**, **10**+**5** qualify*.)	67	What is the counting system? (*multistate* 1:Decimal; 2:Quinary; 3: Body-part tallying; 4: minimal) [Other systems, like senary, are not scored]
38	Is there evidence for any element of a quinary counting system? (*e.g. expressions for 5+1, 10+5+1*.)		
39	Are there words for particular amounts of a thing? *(e.g. ten possums)*		
40	Is there lexical overlap between a significant proportion of adjectives and verbs (including zero-derivation)?	68	Do core adjectives (defined semantically as property concepts; value, shape, age, dimension) act like verbs in predicative position? (1: present, 0: absent)
41	Does the same lexical set of adjectives function both attributively and predicatively?	69	Do core adjectives (defined semantically as property concepts; value, shape, age, dimension) used attributively require the same morphological treatment as verbs? (1: present, 0: absent)
42	Is there case marking for core nominal NPs (i.e., S, A or O function)? *For case marking, include any affixal marking which appears in the NP and shows the function of the NP in the clause; adpositions are not counted.*	70	Is there case marking for core nominal NPs (i.e., S, A or O function)? *for case marking, include any affixal marking which appears in the NP and shows the function of the NP in the clause; do not count adpositions* (1: present, 0: absent)
		71	Is there case marking for core pronouns? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
43	Is there case marking for oblique nominal NPs ? *e.g. locationals, instrumentals, etc.; adpositions are not counted.*	72	Is there case marking for oblique nominal NPs ? *e.g. locationals, instrumentals, etc. do not count adpositions* (1: present, 0: absent)
		73	Is there case marking for oblique pronouns? (1: present, 0: absent)
44	Are there prepositions?	74	Are there prepositions? (1: present, 0: absent)
45	Are there postpositions?	75	Are there postpositions? (1: present, 0: absent)
		76	Are there adpositions to mark core NPs? (1: present, 0: absent)
		77	Are there adpositions to mark oblique NPs? (1: present, 0: absent)
		78	Is there a distinction between locational and directional adpositions? (1: present, 0: absent)
46	Do the same morphemes systematically encode both TAM and person?	79	Do verbs have prefixes/proclitics, other than those that ONLY mark A, S or O (do include portmanteau: A & S + TAM)? *A, S, and O affixes are dealt with in 3.3* (1: present, 0: absent)
47	Do verbs have prefixes/proclitics?		
48	Do verbs have suffixes/enclitics?		
		80	Do verbs have suffixes/enclitics, other than those that ONLY mark A, S or O (do include portmanteau: A & S + TAM)? (1: present, 0: absent)
		81	Can infixation be used on verbs for derivational, aspectual, or voice-changing purposes? (1: present, 0: absent)
		82	Is there present tense regularly morphologically marked on the verb? (1: present, 0: absent)
		83	Is there past tense regularly morphologically marked on the verb? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
		84	Is there future tense regularly morphologically marked on the verb? (1: present, 0: absent)
		85	Are there multiple past or future tenses, distinguishing distance from Time of Reference, marked on the verb? (1: present, 0: absent)
49	Is a distinction between punctual/ continuous aspect available as a morphological choice?	86	Is a distinction between punctual/ continuous aspect available as a morphological choice? (1: present, 0: absent)
50	Is a distinction between realis/irrealis mood available as a morphological choice?	87	Is a distinction between realis/irrealis mood available as a morphological choice? (1: present, 0: absent)
		88	Is there an apprehensive modal category marked on the verb *also known as 'evitative', 'lest', etc* (1: present, 0: absent)
51	Is the S participant (at least sometimes) marked by a suffix/ enclitic? *pertains to verb morphology*	89	Is the S participant (at least sometimes) marked by a suffix/enclitic? *pertains to verb morphology* (1: present, 0: absent)
52	Is the S participant (at least sometimes) marked by a prefix/ proclitic? *pertains to verb morphology*	90	Is the S participant (at least sometimes) marked by a prefix/proclitic? *pertains to verb morphology* (1: present, 0: absent)
53	Is the A participant (at least sometimes) marked by a suffix/ enclitic? *pertains to verb morphology*	91	Is the A participant (at least sometimes) marked by a suffix/enclitic? *pertains to verb morphology* (1: present, 0: absent)
54	Is the A participant (at least sometimes) marked by a prefix/ proclitic? *pertains to verb morphology*	92	Is the A participant (at least sometimes) marked by a prefix/proclitic? *pertains to verb morphology* (1: present, 0: absent)
55	Is the O participant (at least sometimes) marked by a suffix/ enclitic? *pertains to verb morphology*	93	Is the O participant (at least sometimes) marked by a suffix/ enclitic? *pertains to verb morphology* (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
56	Is the O participant (at least sometimes) marked by a prefix/ proclitic? *pertains to verb morphology*	94	Is the O participant (at least sometimes) marked by a prefix/proclitic? *pertains to verb morphology* (1: present, 0: absent)
57	Are variations in marking strategies of core participants based on TAM distinctions?	95	Are variations in marking strategies of core participants based on TAM distinctions? *this question refers to variations (if they occur) in 89-94* (1: present, 0: absent)
58	Are variations in marking strategies based on verb classes?	96	Are variations in marking strategies based on verb classes? *this question refers to variations (if they occur) in 89-94* (1: present, 0: absent)
59	Are variations in marking strategies based on clause type, e.g. main vs subordinate?	97	Are variations in marking strategies based on clause type, e.g. main vs subordinate? *this question refers to variations (if they occur) in 89-94* (1: present, 0: absent)
60	Are variations in marking strategies based on person distinctions?	98	Are variations in marking strategies based on person distinctions? *this question refers to variations (if they occur) in 89-94* (1: present, 0: absent)
61	Do verb stems alter according to the **number** of a core participant?		
62	Do verb stems alter according to the **person** of a core participant?	99	Do verb stems alter according to the person of a core participant? (1: present, 0: absent)
63	Is number ever marked separately from person on the verb?	100	Is number ever marked separately from person on the verb? (1: present, 0: absent)
64	Are person, number and any TAM category (i.e. **3** or more categories in all) marked by portmanteau morphemes on verbs?	101	Are person, number and any TAM category (i.e. 3 or more categories in all) marked by portmanteau morphemes on verbs? (1: present, 0: absent)
65	Are categories such as person, number, gender related to a single participant discontinuously marked on a verb?	102	Are categories such as person, number, gender related to a single participant discontinuously marked on a verb? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
66	Is a non-core participant marked on the verb? *Include affixes, clitics and satellite particles associated with verbs forming a constituent with the verb on some level, but exclude optional adverbials*.	103	Are benefactive nominals marked on the verb? (1: present, 0: absent)
		104	Can instruments be marked on the verb? (1: present, 0: absent)
67	Can recipients be treated as a transitive object, i.e. as Direct object?	105	Can recipients be treated as a transitive object, i.e. as Direct Object? (1: present, 0: absent)
68	Are there syntactically ditransitive verbs?	106	Are there syntactically ditransitive verbs? (1: present, 0: absent)
69	Is negation marked morphologically on the verbs? *i.e. affixation, stem alternation, neutralization of some inflection*	107	Is negation marked morphologically on the verbs? *i.e. affixation, stem alternation, neutralization of some inflection* (1: present, 0: absent)
70	Is direction marked on verbs *Includes affixes, clitics and satellite particles associated with verbs forming a constituent with the verb on some level, but excludes optional adverbials.*	108	Can locative or direction be morphologically marked on the verb? *Locative as Direct Object ('she sleeps mat') does not qualify* (1: present, 0: absent)
71	Are there suppletive verbs for number of participants	109	Are there suppletive verbs for number of participants? *(list them all if feasible, otherwise give an estimate of the number and/or proportion of nouns)* (1: present, 0: absent)
		110	Are there suppletive verbs for tense or aspect? (1: present, 0: absent)
72	Are there conjugation classes?	111	Are there conjugation classes? (1: present, 0: absent)
73	Are there (several) verbs which can be used either transitively or intransitively with no morphological marking? *say no if it's only one or two stems; Intended here is the 'break' and 'open' type; not John eats/ eats the bread*	112	Are there (several) verbs which can be used either transitively or intransitively with no morphological marking? *say no if it's only one or two stems; Intended here is the 'break' and 'open' type; not John eats/ eats the bread* (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
74	Is there transitivising morphology (include clitics)?	113	Is there transitivising morphology (include clitics)? (1: present, 0: absent)
75	Is there morphology (include clitics) to mark a reflexive action? *free word/ particle does not count; neither a default P/N co-reference*	114	Is there morphology (include clitics) to mark a reflexive action? *free word/ particle does not count; neither a default P/N co-reference* (1: present, 0: absent)
76	Is there morphology (include clitics) to mark a reciprocal action? *free word/particle does not count; neither a default P/N co-reference*	115	Is there morphology (include clitics) to mark a reciprocal action? *free word/particle does not count; neither a default P/N co-reference* (1: present, 0: absent)
77	Do verbs classify the shape, size, consistency or position of absolutive arguments by means of incorporated nouns, verbal affixes or suppletive verb stems? *not included here are positional verbs that classify a referent in such terms*	116	Do verbs classify the shape, size, consistency or position of absolutive arguments by means of incorporated nouns, verbal affixes or suppletive verb stems? *not included here are positional verbs that classify a referent in such terms - covered by 127* (1: present, 0: absent)
78	Is there a copula for predicate nouns? *e.g. John is a teacher*	117	Is there a copula for predicate nouns? *e.g. John is a teacher* (1: present, 0: absent)
79	Are there serial verb constructions? *(i.e. two or more verbs in juxtaposition, functioning as a single predicate, with no morphology to mark their relationship with each other. Each of the verbs is a separate phonological word but the construction as a whole is expressed in one intonational unit. Morphology is shared to a greater or lesser extent.)*	118	Are there serial verb constructions? *(i.e. two or more verbs in juxtaposition, functioning as a single predicate, with no morphology to mark their relationship with each other. Each of the verbs is a separate phonological word but the construction as a whole is expressed in one intonational unit. Morphology is shared to a greater or lesser extent.)* (1: present, 0: absent)
80	Is there one or more auxiliary?		
		119	Are there modal auxiliaries? (1: present, 0: absent)
		120	Are there aspectual auxiliaries? (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
		121	Are there tense auxiliaries? (1: present, 0: absent)
81	Is verb compounding a regular process? *(i.e. two or more verb stems acting as one phonological and grammatical word)*	122	Is verb compounding a regular process? *(i.e. two or more verb stems acting as one phonological and grammatical word)* (1: present, 0: absent)
82	Are there verb-adjunct (aka light-verb) constructions? *(i.e. constructions involving a non-predicating element expressing the lexical meaning of the construction, in conjunction with a semantically fairly empty verb, which enables the element to function as a predicate by providing the necessary morphology, e.g. eye do for 'see'; or sneeze hit for 'sneeze')*	123	Are there verb-adjunct (aka light-verb) constructions? *(i.e. constructions involving a non-predicating element expressing the lexical meaning of the construction, in conjunction with a semantically fairly empty verb, which enables the element to function as a predicate by providing the necessary morphology, e.g. eye do for 'see'; or sneeze hit for 'sneeze')* (1: present, 0: absent)
83	Is there incorporation of any element into verbs?		
		124	Is there incorporation of nouns into verbs a productive intransitivizing process? (1: present, 0: absent)
		125	Is there productive incorporation of other elements (adjectives, locatives, etc.) into verbs? (1: present, 0: absent)
84	Is there one or more existential verb? *exclude e.g. positional verbs*	126	Is there one or more existential verb? *exclude e.g. positional verbs (3.8.02)* (1: present, 0: absent)
		127	Are there positional (classificatory) verbs? *(i.e. in answer to a question 'Where is the X', does the verb used in the answer depend on the type of referent (e.g. do you have to say 'The X sits/stands/lies/etc on the table'). List them all.* (1: present, 0: absent)
85	Is the verb 'give' morphologically peculiar (different from most other verbs)? *e.g. stem suppletion, different affixation*	128	Is the verb 'give' morphologically peculiar (different from most other verbs)? *e.g. stem suppletion, different affixation* (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
86	Is there a notably small number, i.e. about 100 or less, of verbs in the language?	129	Is there a notably small number, i.e. about 100 or less, of verbs in the language? (1: present, 0: absent)
87	Is a pragmatically unmarked constituent order SV for intransitive clauses?	130	What is the pragmatically unmarked order of S and V in instransitive clauses? (*multistate* 1: SV; 2: VS; 3: both)
88	Is a pragmatically unmarked constituent order VS for intransitive clauses?		
89	Is a pragmatically unmarked constituent order verb-initial for transitive clauses?	131	Is a pragmatically unmarked constituent order verb-initial for transitive clauses? (1: present, 0: absent)
90	Is a pragmatically unmarked constituent order verb-medial for transitive clauses?	132	Is a pragmatically unmarked constituent order verb-medial for transitive clauses? (1: present, 0: absent)
91	Is a pragmatically unmarked constituent order verb-final for transitive clauses?	133	Is a pragmatically unmarked constituent order verb-final for transitive clauses? (1: present, 0: absent)
		134	Is the order of constituents the same in main and subordinate clauses? (1: present, 0: absent)
		135	Do clausal objects occur in the same position as nominal objects? (1: present, 0: absent)
92	Is constituent order fixed? *Do not consider 'left or right-dislocation', accompanied by intonational signals*	136	Is constituent order fixed? *Do not consider 'left or right-dislocation', accompanied by intonational signals* (1: present, 0: absent)
93	Can negation be marked clause-finally? *This includes suffixes on verb-final clauses; prefixes on clause-final verbs do not count; Don't include elliptical 'Pete didn't'*	137	Can negation be marked clause-finally? *This includes suffixes on verb-final clauses; prefixes on clause-final verbs do not count; Don't include elliptical 'Pete didn't'* (1: present, 0: absent)
94	Can negation be marked clause-initially? *Don't include elliptical 'Not Mary'*	138	Can negation be marked clause-initially? *Don't include elliptical 'Not Mary'* (1: present, 0: absent)

	characters 'Language 2008'	**characters 'PloS 2009'**
95	Is there a difference between imperative and declarative negation?	139 Is there a difference between imperative and declarative negation? (1: present, 0: absent)
96	Are verbal and non-verbal predicates marked by the same negator?	140 Are verbal and non-verbal predicates marked by the same negator? (1: present, 0: absent)
97	Are S and O conflated morphologically in at least some basic constructions, i.e. simple main clauses?	141 Are S and O conflated morphologically in at least some basic constructions, i.e. simple main clauses? (1: present, 0: absent)
98	Are S and A conflated morphologically in at least some basic constructions, i.e. simple main clauses?	142 Are S and A conflated morphologically in at least some basic constructions, i.e. simple main clauses? (1: present, 0: absent)
99	Are S and O conflated morphologically across clause boundaries, i.e. acting as syntactic pivot?	143 Are S and O conflated morphologically across clause boundaries, i.e. acting as syntactic pivot? (1: present, 0: absent)
100	Are S and A conflated morphologically across clause boundaries, i.e. acting as syntactic pivot?	144 Are S and A conflated morphologically across clause boundaries, i.e. acting as syntactic pivot? (1: present, 0: absent)
101	Do S and O operate in the same way, and differently from A, for the purpose of any syntactic construction?	145 Do S and O operate in the same way, and differently from A, for the purpose of any syntactic construction? (1: present, 0: absent)
102	Is there a morpho-syntactic distinction between predicates expressing controlled versus uncontrolled events or states?	146 Is there a morpho-syntactic distinction between predicates expressing controlled versus uncontrolled events or states? (1: present, 0: absent)
		147 Is there a morphologically marked passive construction? *morphological marking includes some verbal affixation or some periphrastic element in the VP or clause* (1: present, 0: absent)
		148 Is there a morphologically marked antipassive? *morphological marking includes some verbal affixation or some periphrastic element in the VP or clause* (1: present, 0: absent)

	characters 'Language 2008'		**characters 'PloS 2009'**
		149	Is there a morphologically marked inverse? *i.e. different marking by verbal affixation or pronominal clitics referring to A and O, depending on person, animacy or definiteness* (1: present, 0: absent)
103	Is there clause chaining? *i.e. chains of morphologically stripped-down medial clauses which are dependent on a single clause (usually, but not necessarily, final) for their TAM or participant marking specification*	150	Is there clause chaining? *i.e. chains of morphologically stripped-down medial clauses which are dependent on a single clause (usually, but not necessarily, final) for their TAM or participant marking specification* (1: present, 0: absent)
		151	Is there a morphologically-marked switch reference system? (1: present, 0: absent)
104	Is there a morphologically marked distinction between simultaneous and sequential clauses?	152	Is there a morphologically marked distinction between simultaneous and sequential clauses? (1: present, 0: absent)
105	Is the verb 'say' or a quotative construction used in desiderative constructions? *(e.g. 'I said for him to go' for 'I wanted him to go')*	153	Is the verb 'say' or a quotative construction used in desiderative constructions? *(e.g. 'I said for him to go' for 'I wanted him to go')* (1: present, 0: absent)
106	Are there purposive non-finite subordinate clauses?		
107	Are there temporal non-finite subordinate clauses?		
108	Are there complement clauses?		
109	Are causatives formed by serial verb constructions?	154	Are causatives formed by serial verb constructions? (1: present, 0: absent)
110	Are causatives formed by bound affixes/clitics?	155	Are causatives formed by bound affixes/clitics? (1: present, 0: absent)
111	Are causatives formed by constructions involving 'say'?	156	Are causatives formed by constructions involving 'say'? (1: present, 0: absent)
112	Is topic or focus marked morphologically? *i.e. by affixes or clitics*.		

	characters 'Language 2008'		**characters 'PloS 2009'**
113	Is there tail-head linkage? *(i.e. a discourse strategy in which the final verb of one sentence is repeated as the first verb of the next sentence)*	157	Is there tail-head linkage? *(i.e. a discourse strategy in which the final verb of one sentence is repeated as the first verb of the next sentence)* (1: present, 0: absent)
114	Are verbs reduplicated?	158	Are verbs reduplicated? (1: present, 0: absent)
115	Are nouns reduplicated?	159	Are nouns reduplicated? (1: present, 0: absent)
		160	Are elements apart from verbs or nouns reduplicated? (1: present, 0: absent)

4

Language Documentation & Conservation Special Publication No. 5 (December 2012)
Melanesian Languages on the Edge of Asia: Challenges for the 21st Century,
ed. by Nicholas Evans and Marian Klamer, pp. 72-108
http://nflrc.hawaii.edu/ldc/sp05/
http://hdl.handle.net/10125/4561

Papuan-Austronesian language contact: Alorese from an areal perspective

Marian Klamer
Leiden University

This paper compares the grammar and lexicon of Alorese, an Austronesian language spoken in eastern Indonesia, with its closest genealogical relative, Lamaholot, spoken on east Flores, as well as with its geographical neighbours, the Papuan languages of Pantar. It focusses on the question how Alorese came to have the grammar and lexicon it has today. It is shown that Alorese and Lamaholot share a number of syntactic features which signal Papuan influences that must have been part of Proto-Lamaholot, suggesting (prehistoric) Papuan presence in the Lamaholot homeland in east Flores/Solor/ Adonara/ Lembata. The data indicate that Proto-Lamaholot had a rich morphology, which was completely shed by Alorese after it split from Lamaholot. At the same time, lexical congruence between Alorese and its current Papuan neighbours is limited, and syntactic congruence virtually absent. Combining the comparative linguistic data with what little is known about the history of the Alorese, I propose a scenario whereby Lamaholot was acquired as non-native language by spouses from different Papuan clans who were brought into the Lamaholot communities that settled on the coast of Pantar at least 600 years ago. Their morphologically simplified language was transferred to their children. The history of Alorese as reconstructed here suggests that at different time depths, different language contact situations had different outcomes: prehistoric contact between Papuan and Proto-Lamaholot in the Flores area resulted in a complexification of Proto-Lamaholot, while post-migration contact resulted in simplification. In both cases, the contact was intense, but the prehistoric contact with Papuan in the Flores area must have been long-term and involve pre-adolescents, while the post-migration contact was probably of shorter duration and involved post-adolescent learners.

1. Introduction. This article is about Alorese (Alor), an Austronesian language in eastern Indonesia.[1] It focusses on the question how Alorese came to have the grammar and lexicon it has today. By comparing Alorese with its closest relative, Lamaholot, as well as with its non-Austronesian neighbouring languages, we reconstruct some of its history and structural features.

Alorese (also refered to as *Bahasa Alor, Alor, Coastal Alorese,* Barnes 2001: 275) is spoken by 25,000 speakers in pockets along the coasts of western Pantar and the Bird's Head of Alor island, as well as on the islands Ternate and Buaya (Stokhof 1975:8-9, Grimes et al. 1997, Lewis 2009), see figure 1. Klamer (2011) is a sketch grammar of the language. Alorese is the only indigenous Austronesian language spoken in the Alor Pantar archipelago. It shows significant dialectal variation; for example, lexical differences exists between the dialect of Baranusa (Pantar island) and the dialect spoken on Alor. The data discussed in this paper is mainly from the Baranusa dialect. All data are primary data collected during fieldwork in 2003.

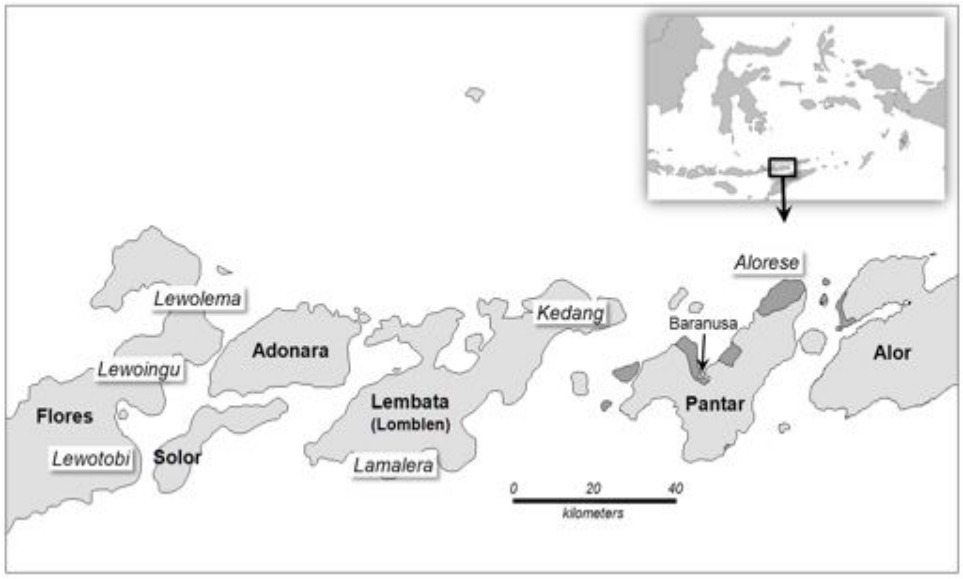


Figure 1. Alorese as spoken Alor, Pantar, Buaya and Ternate (dark grey areas); Lamaholot varieties as spoken on Flores, Solor, Lembata.

In earlier sources, it has been suggested that Alorese is a dialect of Lamaholot (Stokhof 1975:9, Keraf 1978:9, Steinhauer 1993:645), and likewise, the map in Blust (2009a:82)

[1] I would like to thank the two anonymous reviewers of this volume and Nick Evans as co-editor for insightful comments and detailed suggestions for improvement. Many thanks also to Sander Adelaar, Antoinette Schapper, Ger Reesink, and Hein Steinhauer who commented on earlier versions of this paper.

indicates that Lamaholot is spoken on Alor and Pantar. A recent historical comparison by Doyle (2010) suggests that genealogically Alorese is indeed closely related to Lamaholot.

Lamaholot (abbreviated as LMH) is spoken on the eastern part of Flores, and on the islands of Solor and Lembata. Lamaholot has 150,000-200,000 speakers. Although it is usually referred to as a single language, it is better thought of as a dialect chain. Known varieties include the following (see figure 1):[2]

(i) LMH-Lewotobi, spoken in Wulunggitang and Ile Bura, in the western-most part of the Lamaholot speaking region on Flores (Nagaya 2009a,b).
(ii) LMH-Lewolema, spoken in the village Belogili-Balukhering, north of the town Larantuka on east Flores. Pampus (1999, 2001) are word lists of this variety.
(iii) LMH-Lewoingu (Lewolaga), spoken in the village Leworook, south of Larantuka and described in Nishiyama and Kelen (2007).
(iv) LMH-Solor, spoken on Solor island and described by Arndt (1937) and Bouman (1943), lexical survey data collected by Klamer (2002).
(v) LMH-Lamalera, spoken on south Lembata. Keraf (1978) is a description of the morphology of this variety.

To the west, the Lamaholot speaking area is bordered by the language Sika (Lewis & Grimes 1995). A neighboring language in the east is Kedang (Samely 1991), spoken on north Lembata. While Kedang is geographically close to both Lamalera (south Lembata) and Alorese (north-west Pantar), it is genealogically only remotely related to either variety (Doyle 2010).

A comparison of 200+ basic word lists of LMH-Lewoingu, LMH-Solor and LMH-Lamalera with Alorese renders lexical similarity percentages of Alorese versus these three other varieties that range between 52.6 % and 58.8 % (Klamer 2011:18-19). This suggests that Alorese is lexically distinct enough to be qualified as a language of its own. In addition, significant morphological differences exist between Lamaholot and Alorese (see section 3.3 below, and Klamer 2011). For these reasons, the current paper considers Alorese a language on its own, and different from Lamaholot in any of its varieties listed above.

In this paper, I first compare the syntax, morphology and basic vocabulary of Alorese with Lamaholot in sections 2 and 3, followed by a comparison with its non-Austronesian neighbours in section 4. For the syntactic comparison, Alorese will be contrasted mostly with the LMH-Lewoingu variety, as Nishiyama & Kelen (2007) (henceforth N&K 2007) is to date the only published source on a Lamaholot variety that contains syntactic details. (When possessive structures are compared I also refer to LMH- Lamalera, as Keraf 1978

[2] Abbreviations: AL=alienable, DIST=distal, FIN=final, FOC=focus, INAL=inalienable, IND=Indonesian, LOC=location, NEG=negation, OBL=oblique, PL=plural, POSS=possessor, PRF=perfective, RDP=reduplication, REAL=realis, SEQ=sequential, SG=singular.

contains information on this topic.) For the morphological comparison, Alorese will be contrasted with both LMH-Lewoingu (N&K 2007) and LMH-Lamalera (Keraf 1978). For the comparison of the Alorese lexicon and syntax with its non-Austronesian Alor-Pantar neighbours, I refer to the Alor Pantar Lexical Database (listed as such in the references) and for the grammatical constructions I present published and unpublished field data collected by colleagues and myself as indicated in the text.

The structure of the paper is as follows. In section 2, I identify a number of 'Papuan' features found in both Alorese and Lamaholot, and investigate what these suggest about the shared history of the two languages. In section 3, I investigate to what extent Alorese and Lamaholot are syntactically or morphologically different, and what these differences suggest about the history of Alorese, after it split from Lamaholot. In section 4, I investigate some lexical and syntactic changes that occurred after its speakers settled on Pantar and Alor, by comparing the Alorese lexicon and grammar with the lexicon and grammar of its non-Austronesian neighbours. In section 5, I present some notes on the history and ethnography of the Alorese speakers and in section 6, I summarize the reconstruction of the history of the Alorese language and its speakers, and suggest a scenario how it developed into the language it is today.

2. Papuan features in Alorese and Lamaholot

2.1. Introduction. The term 'Papuan' is often used to refer to the perhaps 800 languages spoken in New Guinea and its vicinity that do not belong to the Austronesian language family.[3] In this paper I use 'Papuan' to refer to languages that are not Austronesian and are spoken in eastern Indonesia. The Papuan languages spoken in the Alor archipelago just north of Timor are geographically closest to the Lamaholot speaking region, and will therefore be focussed on in the discussion of 'Papuan' features in this section. The Alor-Pantar languages form a closeknit family (Holton et al. 2012), and are in turn related to the non-Austronesian languages of Timor and Kisar, with whom they form the Timor-Alor-Pantar family (Schapper et. al. 2012). A higher order affiliation of the Timor-Alor-Pantar family to another Papuan group cannot be established (Holton et al. 2012, Robinson & Holton 2012), although a long-standing assumption, beginning with Wurm et al. (1975), has it that the Timor Alor Pantar languages belong to the Trans-New Guinea family.

The non-Austronesian populations in eastern Indonesia must have predated the arrival of the Austronesian speakers (cf. Pawley 2005:102, Ross 2005:18), but there is no reason to assume that Papuan languages spoken in eastern Indonesia today descend from a single prehistoric group. It is far more plausible that they derive from a complex mix of prehistoric populations and various waves of immigrants.

Over the past decade a body of literature has appeared which argued for the relevance of certain particular structural features in the typological characterization of the languages of eastern Indonesia (see also Reesink & Dunn, this volume). In the Austronesian languages of this area, certain features are considered to represent a 'Papuan' influence (e.g. the existence of a post-predicate negator, Reesink 2002; see Florey 2010 for a modification),

[3] As in Tryon (1995:3): "The term 'Papuan' is a convenient term for the non-Austronesian languages of Papua New Guinea and eastern Indonesia, not all of which are demonstrably related."

while other features found in Papuan languages are suggestive of Austronesian influence (e.g. verb-object order correlating with the typical head-initial phrase structure found in Austronesian languages (Clark 1990, Tryon 1995). Works discussing Austronesian-Papuan contact in eastern Indonesia proposing features that diffused as the result of this contact include Grimes (1991), Reesink (2002), Klamer (2002), Donohue (2004), Himmelmann (2005), Klamer, Reesink & Van Staden (2008) and Klamer & Ewing (2010).

The current section identifies a number of features that are part of the Austronesian languages Alorese and Lamaholot, but at the same time are generally recognized as features that are typical for a 'Papuan' language, not an Austronesian one (in the general sense of 'Papuan' in the sources just mentioned). I investigate what the presence of these features suggest about the history of these languages. Highlighted features are: post-predicate negation (section 2.2); the marking of possessors (section 2.3); the noun-locational order in locative constructions (section 2.4); the presence of a focus particle (section 2.5); and the absence of a passive verb form and construction (section 2.6). The results are summarised and discussed in section 2.7. The Papuan languages closest to Lamaholot are the Alor-Pantar languages spoken on west Pantar, see figure 1. The Papuan features discussed in the following sections will therefore be illustrated with examples from languages spoken on Pantar: Teiwa, Blagar, Adang, Sar and Kaera. It is however important to bear in mind that in the Lamaholot-speaking region itself no Papuan language is currently spoken.

2.2. Post-predicate negation. The canonical Austronesian position for negations is to precede the predicate, but in the Papuan languages in the Alor Pantar region it follows the predicate, as illustrated for Teiwa in (1).

(1)	*Na*	*iman*	*ga-pak-an*	*iman*	*suk-an*	*maan.*
TEI	1SG	they	3SG -call- REAL	3PL	exit.come.down- REAL	NEG

'I called them [but] they didn't come out' (Klamer 2010:25)

Both Alorese (Alor) and Lamaholot (LMH) also have a final, 'post-predicate', negation, as shown in (2) and (3).

(2)	*Akhirnya,*	*kujo*	*ha*	*no*	*nele*	*n-ei*	*tobo*	*kaha*	*lang*
Alor	finally(IND)	crab	this	3SG	crawl	3SG-go	sit	coconut.shell	under

'Finally, this crawled to sit underneath a coconut shell

mu	*no*	*pana*	*ha*	*n-ei*	*tahi*	*lahe.*
SEQ	3SG	walk	this	3SG-go	sea	NEG

then he did not go to the sea [again].'

(3)	*Go*	*bərin*	*na*	*hala'.*
LMH	I	hit	him	NEG

'I don't hit him.' (N&K 2007: 69)

The Alorese negator *lahe* is a metathetised form of the Lamaholot negator *hala'* found in

LMH-Lewoingu, LMH-Lewolema, and LMH-Solor.[4,5]

2.3. Possessive marking. In the nominal domain, three Papuan features relating to possessive structures are relevant: (i) the replacement of possessive suffixes by possessor pronouns that precede the possessed noun, (ii) the marking of distinct classes of alienable and inalienable nouns and (iii) the relative order of possessor and possessee.

2.3.1. Replacing possessive suffixes by prenominal possessor pronouns. In Papuan languages, possessors typically precede the possessed, and the person and number features of a possessor are encoded as a prefix on the noun, as illustrated for Teiwa in (4):

(4) *Rai* *ga-yaf*
TEI king 3SG-house
'The king's house'

The possessor pronouns of Alorese and LMH-Lewoingu and LMH-Lamalera are given in (5). In LMH-Lewoingu and LMH-Lamalera possessors are encoded as suffixes or as free pronouns following the possessee, as illustrated in (6). Alorese has no possessive suffixes and uses a free possessor pronoun, which precedes the possessee, as illustrated in (7). (See also section 2.3.3.)

(5) Pronouns and affixes to encode possessors

	Alorese	LMH-Lewoingu (N&K 2007:13, 23-30)		LMH-Lamalera (Keraf 1978:85-95)	
1SG	go	go'en	-kən	goe	-k, -ka
2SG	mo	mo'en	-ko	moe	-m, ma
3SG.AL	ni[6]	na'en	-nən	nae	non-segmental[7]
3SG.INAL	no	na'en	-nən	nae	no suffix (C-final stem)/ non-segmental (V-final stem)
1PL.EXCL	kame	kame'en	-kən	kame	-kem
1PL.INCL	ite	tete'en	-te	tite	-te
2PL	mi	mion	-ke	mio	-kre, re
3PL	fe / fereng	ra'en	-ka	rae	-ri

[4] Identical metathesis patterns occur in other words; compare Alorese *mareng* 'night' with LMH-Lewolema *remã,* LMH-Lewoingu *rəman*; and Alorese *kamore* 'rat' with LMH-Lewolema *kərome,* LMH-Lewoingu *kərome*.

[5] The LMH-Lamalera negation listed in Keraf (1978) is *take.* This word functions in LMH-Lewoingu as negative answer'no'.

[6] Alternative pronunciation *ne*.

[7] 3rd person possessor suffixes differ for stems ending in a consonant or in a vowel. Inalienable nouns ending in a consonant have no suffix. For all the other stems, 3rd person singular possessor features are expressed as lengthening of the stem vowel and/or consonant, and/or vowel

(6) a. *Lango-kən*
LMH house-1SG
'My house' (N&K 2007:23)

b. *Lango* *go'en*
house 1SG
'My house' (N&K 2007:23)

(7)	*Mato*	*kete*	*ni*	*ning*	*anang*	*labi.*
Alor	frog	that	3SG	POSS	child	many

'That frog has many children' or 'That frog's children are many'

2.3.2. Marked distinction between alienable and inalienable nouns. Both Lamaholot and Alorese have a marked distinction between alienable and inalienable nouns. This distinction is not a typical feature of the Austronesian family as a whole, although it is found in some Austronesian languages of eastern Indonesia (see Klamer 2002 for examples). The Papuan languages of Alor Pantar all mark the distinction. In Blagar, for instance, inalienables have an (obligatory) possessor prefix (a), while alienables have a free possessor pronoun (b):

(8) a. *N-amal* b. *Ne* *quu*
Blagar 1SG.INAL-voice 1SG.AL tuber
'My voice' 'My tuber' (Steinhauer 1993:150-151)

In LMH-Lamalera, the distinction is also marked, this time by the obligatory vs. optional use of a possessor morpheme: inalienable nouns must always have a possessor suffix, while alienable nouns can occur without a possessor. Both inalienable and alienable possession are expressed by the same morphemes, except for the 3rd person singular possessor, as shown in the rightmost column of (5) above.

In Alorese and in LMH-Lewoingu, inalienable possession is expressed by a dedicated suffix that attaches to body part nouns. In Alorese, the fossilized suffix is a root-final consonant *–ng* [ŋ]. In LMH-Lewoingu, it is *–('V)n* [ʔVn].[8] Examples are given in (9); most of the forms in Alor and Lamaholot are cognates. Reconstructed Proto-Central Malayo Polynesian (PCMP) and Proto-Malayo Polynesian forms are included for comparison.[9]

nasalization, and/or stress shift (see Keraf 1978: 84-93 for details).

[8] The V stands for any vowel: depending on the open/closeness of the final root syllable, the final vowel of the root is copied as suffix vowel.

[9] Central Malayo Polynesian (CMP) and Eastern Malayo Polynesian (EMP) languages together form the Central Eastern Malayo Polynesian (CEMP) subgroup, a daughter node of Proto Malayo-Polynesian (PMP), which in turn is a daughter node of Proto Austronesian (PAN). PMP includes all the languages of Indonesia. The CMP node (or 'linkage') was proposed by Blust (1993), and Lamaholot is assumed to be affiliated to it. The existence of the CMP node is the topic of a debate (Donohue & Grimes 2008, Blust 2009b), which I will not go into here. The proto-forms cited here are taken from the online Austronesian Basic Vocabulary Database (Greenhill, Blust & Gray

Many modern Alor and Lamaholot words do not reflect these proto forms, but those that do (such as 'hand/arm', 'mouth' and 'eye') contain a non-etymological final nasal. However, the body part nouns in (9d) do contain an etymological final nasal. In LMH, the suffix is optional (9a), obligatory (9b), or absent (9c). This suggests that in LMH the suffix did not lexicalize regularly. In Alorese, the suffix has been completely lexicalised.

(9) Body part nouns with (fossilized) possessive suffixes in Alor and LMH-Lewoingu

	Alor	LMH-Lewoingu (N&K 2007: 174)	PCMP	PMP	Meaning
a.	lima**ng**	lima(**n**)	*lima	*[qa]lima	'hand/arm'
	fofa**ng**	wəwa(**n**)	*babaq	*baqbaq	'mouth'
	rata**ng**	rata(**n**)	*buq, *qulu	*buhek	'hair'
	fuli**ng**	wuli(**n**)	(no data)	*liqeR	'neck'
b.	kotu**ng**	kotə**n**	*qulu	*qulu	'head'
	ale**ng**	kola**'an**	*mudi	*likud	'back'
	lei**ng**	lei**n**	*wai	*qaqay	'foot, leg'
c.	mata**ng**	mata	*mata	*mata	'eye'
	fefele**ng**	wewel	*l(ə/a)ma	*dilaq	'tongue'
d.	tilu**ng**	tilu**n**		*tali**ng**a	'ear'
	niru**ng**	iru**n**		*(i/u)ju**ng**	'nose'
	ulo**ng**	ipə(**'ən**)	*nipə**n**	*(n)ipe**n**	'tooth'

The modern LMH-Lewoingu possessor suffix (listed in (5) above) is in complementary distribution with the fossilized suffix (inalienable) suffix *–n* in (9a). This is shown by the pair (10a-b) (adapted from N&K 2007:11). It is not possible to combine both suffixes, (10c). Note also that the fossilized nasal suffix has not been integrated completely into the nominal root form: it can attach to the adjective and have scope over the nominal phrase, compare (10d-e).

2008), which lists the source author as Blust (1993).

(10) LMH	a.	*mata-n* eye-POSS	'eye'
	b.	*mata-kən* eye-1SG.POSS	'my eye'
	c.	**mata-n-kən* eye-POSS-1SG.POSS	
	d.	*mata belə* eye big	'big eye'
	e.	*mata belə-n* eye big-POSS	'big eye'

In sum, Alorese and LMH-Lewoingu both distinguish inalienable body part nouns from alienable nouns by the presence of a final velar nasal suffix. In LMH-Lewoingu the nature of this element varies between a suffix and a clitic, and it may be replaced by a modern possessor suffix. In Alorese, however, it is a completely and regularly fossilized final root consonant.[10] In LMH-Lamalera, inalienable nouns lack a possessor suffix entirely, or have a non-segmental possessor.

Unlike any of the LMH varieties, an additional strategy has been innovated in Alorese to mark the alienable-inalienable distinction by choice of free pronoun: alienable nouns take *ni* as 3SG possessor, while inalienable nouns take *no*. This is illustrated in (11).

(11) Alor	a.	*ni* 3SG. AL 'his house'	*uma* house		b.	*no* 3SG.INAL 'his father'	*amang* father

I consider *ni* as cognate with LMH 3rd singular possessor pronouns *na'en / nae*, while *no* is an innovation (possibly harmonizing the vowel with the vowels in 1st singular *go* and 2nd singular *mo*) as a dedicated form to mark a 3SG inalienable possessor.

2.3.3. Possessor-possessed order. The third non-Austronesian feature in the nominal domain is the relative order of possessor and possessed in Alorese and Lamaholot. The Papuan order [possessor-possessed] (see (4) above) is the reverse of the [possessed-possessor] order typically found in Austronesian languages, for instance Indonesian *rumah-ku* 'house-1SG' 'my house'.

In LMH-Lamalera, a possessor may be expressed as a free pronoun and replace the possessor suffix (Keraf 1978:95). A free possessor pronoun follows the possessed, rendering the order [possessed-possessor], as in *langoe goe* 'house 1SG' 'my house' (Keraf 1978:95). In other words, LMH-Lamalera consistently displays the Austronesian order.

[10] This analysis also implies that not all inalienables end in a velar nasal, as only those inalienable nouns whose historical root ends in a vowel could take the *ng* as suffix.

By contrast, Alorese only allows the reversed [possessor-possessed] order, as illustrated in (12). If the possessor is expressed as a proper name, as in (13), it must be accompanied by a pronoun, and both name and possessor pronoun precede the possessed. The Alorese order thus mirrors the possessor-possessed order of Papuan languages, as exemplified by Teiwa in (14).

(12) a. *Ni* *uma*
Alor 3SG.AL house
'his house'

b. **uma* *ni*; **uma-ni*; **uma=ni*
house 3SG.POSS

(13) [*Bapa John* *ni* *uma*] *being.*
Alor Mr John 3SG. AL house big
'Mr John's house [is] big.'

(14) [*Kri John* *ga-yaf*] *uwaad.*
TEI Mr John 3SG-house big
'Mr John's house [is] big' (Klamer, n.d.)

The position between LMH-Lamalera and Alorese is taken up by LMH-Lewoingu, which allows either order of possessor and possessee, and employs free possessor pronouns as well as possessor affixes. The Austronesian [possessed-possessor] order is the unmarked order in LMH-Lewoingu (cf. N&K 2007: 27) and is illustrated in (15). Various kinds of possessors may follow the possessed noun: free possessor pronouns (15a), possessor suffixes (15b), or lexical possessors (15c). A suffix and free possessor cannot co-occur, as shown in (15d), which suggests that they have the same referential function. On the other hand, a nominal and a pronominal possessor can co-occur, as shown in (15e).

(15) a. *Lango* *go'en*
LMH house 1SG
'My house.' (N&K 2007: 23)

b. *Lango-kən*
house-1SG.POSS
'My house.' (N&K 2007: 23)

c. *Lango* *guru*
house teacher
'A teacher's house.' (N&K 2007: 24)

d. * *Lango-kən* *go'en*
house-1SG.POSS 1SG.POSS

e. *Lango* *guru* *na'en*
house teacher 3SG
'The teacher's house.' (N&K 2007: 24)

In addition to the [possessed-possessor] order, LMH-Lewoingu also exhibits the 'reversed' [possessor-possessed] order. This order is used when the possessor is encoded as a suffix and the NP contains a coreferential noun. In that case, the noun is preposed, as illustrated in (16):

(16) LMH a. *Guru* *lango-nən*
teacher house-3SG.POSS
'A teacher's house.' (N&K 2007: 23)

b. *Guru* *lango-ka*
teacher house-3PL.POSS
'The teachers' (PL) house(s)/faculty residence.' (N&K 2007: 25)

Of the two possessor marking strategies, the free possessor pronoun is more regular and productive in LMH-Lewoingu than the possessor suffix. For example, N&K (2007:23) note that some Lamaholot speakers cannot use possessor suffixes with words like *oto* 'car' and *bapa* 'father'. Loan words (like *oto*) and frequently used words (like *bapa*) thus appear to prefer free possessors to bound ones.[11] This suggest a development where the possessor suffixing strategy is losing ground to the free pronoun strategy in LMH-Lewoingu.

In conclusion, the Lamaholot varieties and Alorese share some Papuan structural features in the possessive domain. First, in LMH-Lewoingu, a *pre*nominal possessor pronoun strategy is replacing possessive suffixing. This change has been finalised in Alorese, which has only free possessor pronouns left. Both languages mark inalienable body part nouns as a distinct class by means of a fossilized nasal suffix (and Alorese innovated an additional dedicated 3rd person singular inalienable possessor pronoun *no*). Both Alorese and LMH-Lewoingu (but not LMH-Lamalera) show the [possessor-possessed] order that is typical for Papuan languages. In LMH-Lewoingu this is a marked order, while in Alorese it is the only order allowed. The Papuan features which are present in the Lamaholot varieties and in Alorese have thus developed to a further stage in Alorese.

2.4. [NOUN-LOCATIONAL] ORDER IN LOCATIVE EXPRESSIONS. In Alorese and LMH-Lewoingu, locative expressions are constructed of a noun, followed by a locational lexeme of nominal origin (which may function as postposition in certain contexts). An example are the locational nouns *unung* 'inside' (Alor), illustrated in (17), and *ono'on* 'inside' (LMH),

[11] This explanation differs from the one suggested by N&K (2007:23), who refer to *oto* as a "less familiar" word, and *bapa* as a "respectful kinship term". To characterise these words as such does not seem to be true to fact: *oto* is a loan from Indonesian (which borrowed it from Dutch *auto* < French 1897 *auto* 'car') and is known to everyone. *Bapa* 'father; Mr' is not only a kinship term but also used frequently as the polite term of address for male adults (cf. Indonesian *Bapak* 'Mr').

illustrated in (18).

(17)	*Pa*	*ru*	*oro*	*uma*	*unung?*
Alor	what	FOC	LOC	house	inside
	'What is in(side) the house?'				

(18)	*Busan*	*to'u*	*pe*	*dos*	*ono'on.*
LMH	cat	one	at	box	inside
	'There is a cat in the box.' (N&K 2007: 90)				

Both *unung* and *ono'on* are cognate to *'oné* in Keo, spoken in Central Flores (Baird 2002: 141). In Keo, this lexeme is synchronically a preposition. In Indonesian, too, locational nouns occur in prenominal position (cf. Indonesian *di dalam rumah* 'LOC inside house' versus **di rumah dalam* 'LOC house inside'). In line with these observations, I assume that the position of the lexeme *unung/ono'on* in the Austronesian languages of Flores was originally prenominal, and that it moved to postnominal position in Lamaholot and Alorese because of Papuan influence. A Teiwa example of a Papuan noun-locational noun order is given in (19), where the locational noun *gom* 'its inside' follows the noun *yaf*:

(19)	*Na*	[*yaf*	*g-om*]	*ma*	*gi.*
TEI	1SG	house	3SG-inside	LOC	go
	'I go inside the house.' Lit. 'I go into [the house's inside]' (Klamer, n.d.)				

2.5. Focus particle. Alorese and LMH-Lewoingu both have an information structure particle, *ru* and *ke* respectively. This particle functions to mark contrastive focus. The contrast between an unfocused constituent and a focused one in Alorese is illustrated in (20a-b), another illustration is (21).

(20)	a.	*No*	*lelang*	*batang.*
Alor		3SG	make	break
		'He broke [them].'		

	b.	*No*	*ru*	*lelang*	*batang.*
		3SG	FOC	make	break
		'HE broke [them] (not me).'			

(21)	*No*	*maring*	*aleng*	*keleng*	*maring*	*mo*	*ru*	*hela.*
Alor	3SG	say	back	slender	say	2SG	FOC	climb
	'He said to Slender Back: "YOU climb it" [not I].'							

The particle *ke* marks contrastive focus in Lamaholot, as illustrated in (22):

(22) LMH
a. *Go-ke* *hope* *buku* *pi'in.*
1SG-FOC buy book this
'It's me who bought this book.' (N&K 2007: 129)

b. *Go* *hope-ke* *buku* *pi'in.*
1SG buy-FOC book this
'I BOUGHT this book.' (N&K 2007: 129)

c. *Go* *hope* *buku* *pi'in-ke.*
I buy book this-FOC
'I bought THIS BOOK.' (N&K 2007: 129)

Many Papuan languages have particles marking contrastive focus; an illustration from a Pantar language is Teiwa *la* 'FOC', illustrated in (23):

(23) TEI
a. *Rai* *na-soi* *ga-kamadal* *ga-buxun* *tas.*
king 1SG-order 3SG-belt 3SG-guard stand
'The king ordered me to guard his belt.' (Klamer 2010:409)

b. *Rai* *la* *na-soi* *ga-kamadal* *ga-buxun* *tas.*
king FOC 1SG-order 3SG-belt 3SG-guard stand
'The KING ordered me to guard his belt.'

c. *Rai* [*na* *la*] *soi* *ga-kamadal* *ga-buxun* *tas.*
king 1SG FOC order 3SG-belt 3SG-guard stand
'I was ordered by the king to guard his belt.'

Focus particles encode new information, and are typically followed by propositions that are pragmatically presupposed. In many languages, relative clauses are instrumental in coding presupposed propositions. The focus marker thus functions in a way that is similar to a relative clause marker. It is plausible that because they have a focus marker, Alorese and Lamaholot lack a dedicated, indigenous relative clause construction. Under influence of Indonesian, however, both languages have borrowed a relative clause construction that is marked with Indonesian *yang* 'relative marker'. Borrowed *yang* is used optionally, in addition to the focus marker (see N&K 2007:126-127).

2.6. Absence of a passive voice verb and construction. A passive construction is defined here as a clause where the verb carries special morphology to mark the promotion of the verb's underlying patient argument to become the grammatical subject, while demoting the original agent subject into an oblique phrase.

While the languages of Taiwan and the Philippines have fully developed systems with more than two voices, the western Malayo-Polynesian languages of Indonesia usually

have two (Ross 2002: 52). In eastern Indonesia this voice system is reduced,[12] and many languages lack both passive morphology and a dedicated passive construction. Examples include Taba, Alune, Leti, Roti, Tetun Fehan, Bima, Kambera and Keo (cf. Klamer 1996, 2002: 374).

In the Papuan languages of Alor and Pantar a passive is also generally lacking; examples include Klon (Baird 2008), Abui (Kratochvíl 2007) and Teiwa (Klamer 2010a). In Teiwa, the functional equivalent of a passive is a clause with a fronted P followed by a generic noun *hala* 'someone, unknown person' expressing the (backgrounded) Agent; compare (24), with basic A-P-V constituent order, with (25), with P-A-V order and Agent *hala*:

(24)	P		A	V
TEI	*Uy*	*ga'an*	*yivar*	*ga-far.*
	person	that	dog	3SG-kill
	'That person killed a dog / dogs.'			

(25)	P		A	V
TEI	*Uy*	*ga'an*	*hala*	*ga-far.*
	person	that	someone	3SG-kill
	'That person was killed.' (lit. 'That person someone killed.'			

Alorese and Lamaholot, too, lack a passive (N&K 2007:126, Nagaya 2009). Both languages have basic Agent-Verb-Patient (AVP) constituent order, as in (26) and (28). A functional equivalent to a dedicated passive is the fronting of P, as in (27) and (29).

(26)	A		V	P
Alor	*Ama*	*kali*	*g-ang*	*fata.*
	father	that	3SG-eat	rice
	'That man eats rice.'			

(27)	P	A	V	
Alor	*Ume*	*ape*	*g-ang*	*mungga.*
	house	fire	3SG-eat	while
	'The house is on fire.'			

(28)	A	V	P	
LMH	*Na*	*həbo*	*ana'*	*pe'en.*
	3SG	bathe	child	the
	'She bathes the child.' (N&K:79-80)			

[12] Note that the west and centre of Indonesia are more variegated (in particular Borneo and Sulawesi).

(29)			P	A	V	
LMH	*Nolo*	*pe'en*	*tahan*	*tite*	*gəta*	*hala'*
	past	that	rice	we	harvest	NEG

'In the past rice wasn't a crop.' (lit. 'we didn't harvest rice' (N&K:127)

In neither of the languages does the fronting of P involve a change in the verbal morphology; nor does the original A become part of an oblique constituent and all the nominal constituents retain their original shape. In sum, Lamaholot and Alorese lack the passive constructions and voice morphology found in most of the western Austronesian languages, which are similarly lacking in the Papuan languages of Alor and Pantar.

2.7. Summary and discussion. Lamaholot and Alorese share a number of features that are atypical for Austronesian languages in general, but do exist in Papuan languages of the region: they lack a passive, place the negation in post-predicate position, have [possessor-possessed] order, a formal distinction between alienable and inalienable (body part) nouns, a [noun-locational noun] order in locative expressions, and a focus particle.

The hypothesis I submit is that these features arose in Lamaholot and Alorese as a result of intensive contact with one or more Papuan languages. As similar structural features arose in both Lamaholot and Alorese, I assume that they did not arise independently, but were part of their shared ancestor language, Proto-Lamaholot. This implies that most of the Papuan features found in today's Alorese are not due to contact with its current Papuan neighbors on Pantar and Alor, but rather entered the language before it split from Lamaholot.

No written or oral records exist of a history of contact between Lamaholot speakers and speakers of (a) Papuan language(s). Neither do (written or oral) records exist of Papuan languages spoken in east Flores, where Lamaholot is spoken today .[13] However, there is general consensus among linguists that Papuan (non-Austronesian) populations predated the Austronesians, who arrived in the eastern Indonesian region some 3,500 years ago (Pawley 2005, Ross 2005, Donohue & Grimes 2008, Ewing & Klamer 2010). The Papuan structural features I have reconstructed here for Proto-Lamaholot constitute further evidence that Austronesian and Papuan speakers were once in contact in the Lamaholot homeland. This homeland may have been any location west of Pantar; it could have been Solor, Lembata and/or east Flores, but also another location (see section 5).[14]

[13] Although Donohue (2007) argues that extinct Tambora was a Papuan language spoken on Sumbawa, west of Flores island.

[14] While the Lamaholot speakers currently live in east Flores, Solor and Lembata, the homeland of Proto-Lamaholot could also have been somewhere else. As one reviewer remarked, the oral traditions of most communities in East Flores record that they originally came from elsewhere, although it remains unclear from where exactly.

3. Contrasting Lamaholot and Alorese

3.1. Introduction. This section investigates to what extent Alorese and Lamaholot are different syntactically (section 3.2) or morphologically (section 3.3) and what these differences suggest about the history of the Alorese (section 3.4).

3.2. Syntactic differences. The syntactic differences between Alorese and Lamaholot are minimal. Firstly, the order of [possessor-possessed] is a marked order in Lamaholot, while it is the fixed order in Alorese; this was discussed in section 2.3. Secondly, Lamaholot has only clause-initial conjunctions, e.g. *kədin* in (30), while Alorese has at least one conjunction-like element that is clause final, the sequential marker *mu* in (31).

(30)	*Na*	*səba*	*laran*	*nənən ga'e*	*nələ*	*bisa*	*ai*	*topi*	*pe'en.*
LMH	3SG	search	way	how	so	can	get	hat	the
	'She wondered how to get that hat.'								
	kədin	*Mince*		*mari*	*hi*	*topi*	*pe'en*	*məko*	*pe.*
	then	Mince		say	ah	hat	that	ugly	that
	'Then Mince said, "Ah that hat is ugly"' (N&K 2007:170)								

(31)	*Tiba-tiba*	*aho*	*ning*	*kotung*	*maso*	*toples*	*unung*	*mu,*
Alor	suddenly (IND)	dog	POSS	head	enter	jar	inside	SEQ
	'Suddenly the dog's head got into the jar then							
	no	*goka*	*oro*	*tana*	*lulung.*			
	3SG	fall	LOC	earth	on			
	'he fell on the ground.'							

Thirdly, time expressions follow the predicate in Lamaholot, and precede it in Alorese. This is illustrated with the cognate forms *wia/fiang* 'yesterday' in (32)-(33). The Indonesian example in (34) illustrates the typical head-initial order that is typical for an Austronesian language. This is the order found in Lamaholot (32).

(32)	*Ra*	*səga*	*wia.*	
LMH	they	come	yesterday	
	'They came yesterday.' (N&K2007:86)			
(33)	*Ama*	*kali*	*fiang*	*ho.*
Alor	father	that	yesterday	come
	'That man came yesterday.'			
(34)	*Mereka*	*datang*	*kemarin.*	
IND	they	come	yesterday	
	'They came yesterday tomorrow.'			

In sum, I have not found evidence that the syntactic differences between Lamaholot and

Alorese relate to more than just a few small differences in word order.

3.3. Morphological differences

3.3.1. Introduction. Most Austronesian languages of eastern Indonesia and the Pacific have morphological systems that are less elaborate than the Austronesian languages spoken in Taiwan, the Phillipines or western Indonesia (cf. Blust 2009a:343, 347). Some extreme morphological empoverishment is found in languages spoken in central and eastern Flores, including Manggarai, Ngada, Lio, and Keo (Baird 2002). However, not all the Flores languages underwent such massive morphological loss, Lamaholot being a case in point.

The morphological system of Lamaholot has productive reflexes of a significant number of Proto Austronesian / Proto Malayo Polynesian morphemes. In this section, I present a summary of Lamaholot inflectional morphology, compared to Alorese (section 3.3.2); Lamaholot derivational morphology compared to Alorese (section 3.3.3), followed by a summary and discussion (section 3.3.4).

3.3.2. Inflectional morphology. Lamaholot has quite a lot of agreement morphology: subject (A and S) agreement is marked on verbs, adverbs as well as on the conjunctive element *o'on* 'and, with',[15] while adjectives agree in person and number with the (pro)noun they modify (N&K 2007).

There are two different subject paradigms, one is a set of consonantal prefixes, the other a set of suffixes, as given in (35). LMH-Lewoingu and LMH-Lamalera use the same A prefixes, but different S suffixes. Below I discuss subject marking in LMH-Lewoingu; similar (though not identical) observations can be made for LMH-Lamalera, which is not discussed here for reasons of space (see Keraf 1978).

(35) Subject affixes in Lamaholot

	A prefix	S Suffix LMH- Lewoingu (N&K 2007:13)	 LMH-Lamalera (Keraf 1978:73,76)
1SG	k-	-kən	-ka
2SG	m-	-ko, -no[16]	-ko, -o
3SG	n-	-na, -nən	-fa/ra, -a
1PL.EXCL	m-	-kən	-kem
1PL.INCL	t-	-te	-te
2PL	m-	-ke/-ne	-kre, -re
3PL	r-	-ka	-ri, i

In LMH-Lewoingu, the A prefix obligatorily marks the agent (A) of vowel-initial transitive verbs (N&K 2007: 98). Examples include *-a'an* 'make', *-itə* 'sleep with', *-olin* 'improve' (N&K 2007: 32). However, there are also vowel-initial verbs which cannot take an

[15] This suggests that this element may be analyzed as a verb rather than a conjunction.

[16] N&K 2007 list both forms on p. 13, but only *-ko* on p. 31.

agreement prefix (e.g. *opən* 'lie to someone', N&K 2007: 98), so that the use of the A prefix is not purely phonologically conditioned, but also lexically stipulated.

Many Lamahalot verbs can be used both transitively and intransitively with no difference in verb form (N&K 2007:77). When they are used in a transitive construction, A and P are expressed as free noun phrases; when they occur in an intransitive construction, S is encoded as a verbal suffix (N&K 2007: 75-76, 77-78).[17] An S-suffix is also found on adjectives in predicative or adverbial function, in which case the adjective gets an excessive interpretation (N&K 2007: 98-99).

In sum, Lamaholot S and A are often expressed as pronominal affixes on verbs. In contrast to this, verbal arguments in Alorese are almost universally expressed as free pronouns. Exceptions are a few frequent verbs with a fossilised A prefix that are used in combination with an (obligatory) free subject pronoun. Examples pointed out to me by speakers are *-oing* 'to know' and *-enung* 'to drink', as in *go g-oing* 'I 1SG-know' and *mo m-enung* 'you 2SG-drink'.

3.3.3. Derivational morphology. LMH-Lewoingu has seven derivational affix forms, as listed in (36). LMH-Lamalara has six derivational affixes, as listed in (37). The lists summarize the derivations and their semantics presented in N&K 2007.

Some of the LMH derivational affixes are regular and productive, while others are lexicalised to a small or large extent. Often, a single prefix has developed more than one meaning. In all cases, the semantic relation between the base and the derived form is transparent enough to establish at least a generic core meaning of the derivational morpheme. Note that the many nominalizing prefixes derive different semantic types of nominals, and I refer to the original sources for additional descriptive details about individual derivations. Anticipating a reconstruction of Proto-Lamaholot morphology, I provide the possible PAN /PMP affixes alongside their modern Lamaholot reflexes as a hypothesis about the likely etymological relation between them.

(36) Derivational morphology in LMH-Lewoingu

- Prefix *be(C)-* 'nominalizer',[18] *e.g. linon* 'reflect' > *be-linon* 'mirror' (N&K 2007:49-51) < PMP *paŋ 'instrumental noun' or *paR 'deverbal noun' (Blust 2009a: 359, 366)
- Prefix *pə-* 'verbalizer', e.g. *tua* 'palm wine' > *pə-tuak* 'taste like palm wine' (N&K 2007:51) < PMP *pa-ka- 'treat like X' (Blust 2009a:359);
- Prefix pə- 'nominalizer', e.g. *tutu'* 'speak', *pə-nutu'* 'speaker, speaking' (N&K 2007:51) < PMP *paR 'deverbal noun' (Blust 2009a:359)
- Prefix *kə-*, e.g. *pasa* 'swear' > *kə-pasa* 'oath' 'nominalizer' (N&K 2007: 52-53) < PMP *ka- 'formative for abstract nouns' (Blust 2009a:359, 362)
- Infix *-ən-* 'nominalizer', e.g. *tali* 'add' > *t-ən-ali* 'added thing' (N&K: 53-54) < PAN *-um- 'Actor voice' (Blust 2009a:370)

[17] There are also intransitive verbs that cannot be used as transitives, and they express S as a free noun phrase (N&K 2007:63).

[18] N&K 2007: 50-51 refer to this prefix as *beN-* which is realised as *b-, be'-*, *ben-* or *ber-*.

- Prefix *mən-* 'nominalizer',[19] e.g. *ba'at* 'heavy' > *mən-a'at* 'something heavy' (N&K 2007:54) < PAN *ma 'stative' (Blust 2009a:363-364)
- Prefix *gəN-* 'nominalizer',[20] e.g. *balik* 'to return' > *gə-walik* 'return (N)' (N&K 2007:49) < PMP *ka- 'abstract noun formative' (Blust 2009a: 362)
- Consonant replacement, e.g. *pet* 'bind' > *met* 'belt' 'result nominalizer' (N&K 2007: 48-49) < PAN *ma- 'stative' (Blust 2009a: 363-364)

(37) Derivational morphology in LMH-Lamalera

- Prefix *b-/be-* 'deverbal nominalizer', e.g. *udur* 'push' > *b-udur* 'pusher' (Keraf 1978:188), *doru* 'rub' > *be-doru* 'someone rubbing' (Keraf 1978:193); *fai* 'water' > *be-fai* 'have water' (Keraf 1978: 212) < PMP *paŋ 'instrumental noun' or *paR 'deverbal noun' (Blust 2009a: 359, 366)
- Prefix *n-* 'deverbal nominalizer', e.g. *hau* 'sew' > *nau* 'something sewn' (Keraf 1978:192) < unclear etymology
- Circumfix *pə-k,* e.g. *tana* 'earth' > *pe-tana-k* 'taste like earth' (Keraf 1978:210) < PMP *pa-ka- 'treat like X' (Blust 2009a: 359)
- Infix *-en-* 'instrumental nominalizer', e.g. *tika* 'divide' > *t-en-ika* 'instrument to divide' (Keraf 1978:195-196) < PAN *-um- 'Actor voice' (Blust 2009a: 370)
- Prefix *me-* 'nominalizer', e.g. *nange* 'swim' > *me-nange* 'swimmer' (Keraf 1978:197) < PAN *ma 'stative' (Blust 2009a: 363-364)
- Consonant replacement, e.g. *pota* 'add' > *mota* 'addition' 'result nominalizer' (Keraf 1978:190) < PAN *ma 'stative' (ibid.)

In contrast to Lamaholot, Alorese has no derivational morphology at all. The only productive word formation process in Alorese is reduplication: verbs and adverbs undergo full reduplication to indicate iterative or intensive activity, as in (38); while nominal reduplications denote plural diversity, 'all sorts of N'. Similar reduplication takes place in Lamaholot.

(38)	*No*	*geki-geki*	*sampai*	*no*	*neing*	*aleng*	*bola.*
Alor	3SG	RDP-laugh	until (IND)	3SG	POSS	back	break
	'He laughed and laughed till his back broke.'						

The loss of derivational morphological categories in Alorese can be seen as a kind of formal simplification or regularization: affixes that do not show a regular and transparent form-meaning relationship are lost.

3.4. Conclusions. While modern reflexes of PAN / PMP morphology appear in abundance in the Lamaholot varieties, and the Lamaholot varieties do not show a gradual decline of morphology that is related to a geographical West-East spread, Alorese has lost all of its morphology. As morphemes are more easily lost than gained, I assume that

[19] With non-homorganic nasalization of initial root consonant; the process may involve extra final nasal or syllable (see N&K 2007:54).

[20] The nasal in the prefix changes p/b>m, b>w, h>n, and is unrealized before r/l.

Proto-Lamaholot, the shared ancestor of Alorese and Lamaholot, had at least the amount of morphology of today's Lamaholot varieties. This implies that Proto-Lamaholot (i) had subject and possessor affixes, (ii) distinguished agreement of A (prefix) and S (suffix), and (iii) had at least seven different derivational prefixes. After the Lamaholot-Alorese split, Alorese lost all of this morphology. Such massive reduction of morphology is often taken to suggest that a language has gone through a stage of imperfect or second language learning.

4. Alorese and its Papuan neighbors

4.1. Introduction. This section investigates the lexical and syntactic change that occurred in Alorese after it split from Lamaholot, by comparing the Alorese lexicon and grammar with the lexicon and grammar of the Papuan languages in the neighborhood. Lexical borrowing is investigated in section 4.2, followed by a syntactic comparison, focussing on the expression of three types of predicate-argument relations in section 4.3.

4.2. Lexical comparison. In order to estimate the amount of lexical borrowing in Alorese I compared a 270-item basic word lists of Alorese with published lexical data from LMH-Lamalera, LMH-Lewoingu, LMH-Solor, LMH-Lewolema.[21] I focussed on the Alorese words that are formally dissimilar to their semantic equivalent in all four of the Lamaholot varieties. Fifty-five such words were found. Three of these are reflexes of an Austronesian or Proto Malayo-Polynesian word (which has not been retained in the Lamaholot varieties). The remaining 52 words in which Alorese differs from any Lamaholot variety could be lexical innovations or loan words. Of these, 14 words were identified as loans from an Alor Pantar language (see (39)), 5 are Malay/Indonesian loans (see (40)), and 33 have an unknown etymology or source. The donor language of the 14 identifiable loans was found through the Alor Pantar Lexical Database, which contains (220+) basic lexical data from 18 Papuan varieties of the Alor Pantar family (Holton et al. 2012). For comparison, words of the source language(s), LMH-Lamalera, LMH-Lewoingu and PMP are included in (39).[22]

[21] As the focus of this article is on the changes that took place in Alorese, I do not investigate lexical borrowing in the Lamaholot varieties. Doyle (2010) presents an initial compilation and analysis of comparative lexical data of the Lamaholot varieties.

[22] The items for Proto-Malayo Polynesian (PMP) are from the online Austronesian Basic Vocabulary Database (Greenhill, S.J., R. Blust & R.D. Gray 2008), which lists the source author as Blust (1993).

(39) Alorese words with identified Alor-Pantar donor language(s)

Alorese	Meaning	Source		LMH-Lamalera	LMH-Lewoingu	PMP
tor	'road'	tor	W Pantar	larã	laran	*zalan
baling	'axe'	bali	W Pantar, Sar	hepe	soru	no data
duri	'knife'	duir	Adang	hepe	hepe	no data
kondʒo	'clothing'	kondo	Blagar	alelolo	no data	no data
bire kari	'children'	biar kariman	Teiwa	ana	anaʔ	*anak
haʔã	'this'	ħaʔa	Teiwa	pi	pi, piʔn	*i ni
kar-to	'ten'	Proto AP	Reflexes across AP[23]	pulo	pulo, pulu rua	no data
kar-ua	'twenty'	*qar				
ele	'wet'	qaloʔ	Sar	sə'nəbə	dəman	*ma-baseq
		kalok	Teiwa			
		xolo	Kaera			
kari	'thin'	kira	Blagar, Kaera, Teiwa	mə'nipi	mə'nipi	*ma-nipis
laming	'to wash'	laming	W Pantar	ba, pu	baha	no data
kalita	'dirty'	klitaʔ	Teiwa	milã	milan	*cemed
		klitak	Blagar			
tobang	'to push'	tobung	Kaera	uruk	gehan	no data
doho	'to rub'	dahok	Blagar	doru	dosuʔ	no data

(40) Alorese words borrowed from Indonesian/Malay

Alorese	Meaning	Source	
rekiŋ	'to count'	reken	Malay[24] < Dutch
kali	'river'	kali	Malay/Indonesian
danau	'lake'	danau	Indonesian
buŋa	'flower'	buŋa	Indonesian
hati	'liver'	hati	Indonesian

The data in (39) suggest three things. First, Alorese borrowed words from Alor-Pantar languages from right across the island of Pantar: Teiwa and Sar are spoken in the north-west, Western Pantar is spoken in the west and south, and Blagar and Kaera in the east, see figure 2. That all these donor languages are spoken on Pantar is no surprise given that the Alorese word list investigated here is from the Baranusa dialect, spoken in west Pantar.

[23] See Schapper & Klamer (ms.).

[24] Compare Kupang Malay *reken* 'to count' (Jacob & Grimes 2003).

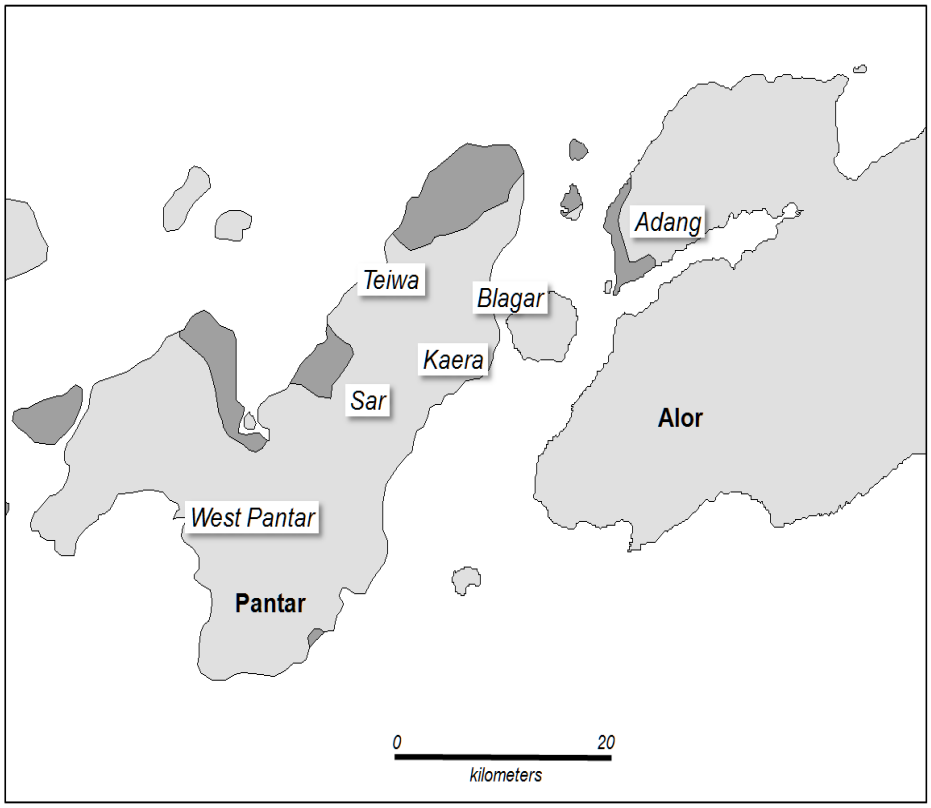


FIGURE 2. Languages from the Alor-Pantar family that are discussed in the text.

Second, among the Alor-Pantar donor languages, there is not one that is particularly dominant. This suggests that contacts of a similar kind existed with different speech communities rather than with one in particular.

Third, of all the donor languages, Malay/Indonesian appears the most dominant one. This is expected of a national language used in education and interethnic communication.

4.3. SYNTACTIC COMPARISON. Alorese and its Alor-Pantar neighbours have a different constituent order: in Alorese the verb precedes the object, as in (41), while the AP languages are all verb final, as illustrated for Teiwa in (42).

(41)	*Aho*	*gaki*	*be*	*kae*	*kali.*
Alor	dog	bite	child	small	that
	'A dog bit that child.'				

(42)	*Yivar*	*bif*	*waal*	*ga-sii.*
TEI	dog	child	that.mentioned	3SG-bite
	'A dog bit that child.'			

The expression of predicate-argument relations is an area where Papuan and Austronesian languages often show grammatical contrasts. In the AP languages, serial verb constructions are pervasive, and do much of the work that is done in (western) Austronesian languages either by causative, applicative, or recipient affixes on verbs (cf. Himmelmann 2005: 170), or by adpositional phrases.

In this section, I compare a small part of the syntax of Alorese with its AP neighbors, to see if there is evidence of syntactic convergence with local Papuan languages after Alorese split from Lamaholot. As a preliminary investigation, I consider three types of predicate-argument relations: 'give' events with an agent (A), recipient (R), and a displaced theme (T) (section 4.3.1); instrumental constructions with an A, patient (P) and instrument (I) (section 4.3.2); and causative constructions where an original S becomes the causee (P) and a new causer (A) is introduced (section 4.3.3).

Alorese is compared with languages spoken in its vicinity: Teiwa (west Pantar), Kaera (east Pantar), Sar (central-west Pantar), Blagar (east Pantar, Pura, Tereweng), and Adang (spoken north of Kalabahi on the Alor peninsula),[25] see figure 2. Contact between Alorese and the Papuan languages spoken in south, central or east Alor is much less plausible, so these languages will not be considered here.

4.3.1. 'Give' constructions. In the Papuan languages investigated here, the verb 'give' is a mono-transitive verb which has the Recipient (R) as its single object, while T is introduced with its own predicate: a verb or a (deverbal) oblique particle. The constituent order is 'A T R-give' , and the pronominal prefix on 'give' encodes person and number of R. Examples are (43)-(47) (data are my own fieldnotes unless indicated otherwise).[26]

(43)	*Uy*	*ga'an*	*u*	*sen*	*ma*	*n-oma'*	*g-an.*
TEI	person	3SG	DIST	money	OBL	1SG-father	3SG-give

'That person gave money to my father.'

(44)	*Ui*	*fo*	*seng*	*ma*	*na-manak*	*g-an.*
Sar	person	that	money	OBL	1SG-father	3SG-give

'That person gave money to my father.' (Baird, survey data 2003)

(45)	*Ui*	*gu*	*gang*	*doi*	*mi*	*na-mam*	*g-eng.*
Kaera	person	that	3SG	money	OBL	1SG-father	3SG-give

'That person gave money to my father.'

[25] Sar data are from a survey carried out by Louise Baird in 2003; Blagar data are from Hein Steinhauer, p.c. 2010, Teiwa and Kaera data are my own fieldnotes (2003, 2007); Adang data are from Haan (2001) unless indicated otherwise.

[26] For further data and discussion of the typology and history of the 'give' construction in Timor Alor Pantar, see Klamer & Schapper (2012).

(46) *Na vet nu metma n-oʔal ʔ-nang.*
Blagar 1SG coconut one OBL 1SG-child 3SG-give
'I give a coconut to my child.' (Steinhauer p.c. 2010)

(47) *Ella seng med ʔ-omang ʔ-en.*
Adang Ella money take 3-father 3-give
'Ella gave money to her father.' (Haan 2001: 377, 139)

In contrast to these, an Alorese 'give' construction employs a ditransitive verb with two bare object NPs, with constituent order 'A give R T' , as illustrated in (48).

(48) *Ama kali ning go bapa seng.*
Alor man that give.(to) 1SG father money
'That man gave my father money.'

In most Austronesian languages, 'give' events involve just a single verb which may be a morphologically simple or derived form, and both objects follow the verb. If one of the objects is part of an oblique constituent, it is R. This is also the pattern attested in Lamaholot, where a bare double object construction like (49) is possible, as well as a construction where R is an oblique constituent (*pe inawae to'u* 'to a girl', N&K 2007:80).

(49) *Go nein inawae to'u bunga to'u.*
LMH 1SG give girl one flower one
'I gave a girl a flower.' (N&K 2007: 80)

4.3.2. Instrumental constructions. Instrumental expresssions involve an agent (A), patient (P) and instrument (I). In the Papuan languages compared here, instruments are either introduced in a serial verb construction with the verb 'take' or hold', or with a deverbal oblique particle. Instruments always precede the main verb.

(50) *Na ped mat ma man taxar.*
TEI 1SG machete take OBL grass cut
'I cut the grass with a machete.'

(51) *Ui nuk peed ma tai gor.*
Sar person one machete OBL tree cut
'Someone cut wood with a machete.' (Baird, survey data 2003)

(52) *Ui gu gang ped mi tei patak-o*
Kaera person that he machete OBL wood/tree cut- FIN
'That person cut wood with a machete.'

(53)	*Na*	*hemering*	*medi-t*	*sal*	*ʔ-u-tukang.*
Blagar	1SG	knife	take-t	rope	3SG-CAU-[27]short
	'I shorten the rope with a knife.'				

(54)	*Name*	*nu*	*sapad*	*puin*	*tiboʔ*	*tatoʔ.*
Adang	person	one	machete	hold	wood	cut
	'Someone cut wood with a machete.' (Baird, survey data 2003)					

Alorese, in contrast, marks instruments with the preposition *nong* 'and, with', in a prepositional phrase following the main verb, as in (55).

(55)	*Ama*	*to*	*tari*	*kaju*	*nong*	*peda.*
Alor	father	one	cut.down	wood	with	machete
	'A man cut the wood with a machete.'					

Proto-Austronesian and Proto-Malayo Polynesian derived instrumental verbs with an instrumental infix. There are also many Austronesian languages where instruments are encoded by an instrumental prepositional phrase, or as complement of the verb 'use'. Lamaholot employs the latter strategy (56).

(56)	*Go*	*bərin*	*Bala*	*pake*	*mənəngo*	*mi'in.*
LMH-Lewoingu	1SG	hit	Bala	use	stick	this
	'I hit Bala using this stick.' (N&K 2007: 116)					

4.3.3. Causative constructions. Highlighted here are causative constructions based on a non-active intransitive verb, whose original subject (S) becomes the causee (P) of the causative construction, while a new causer agent (A) is introduced.

Two languages of Pantar (Teiwa, Sar) employ lexical causatives, as illustrated in (57) and (59).

(57)	a.	*Wat*	*nuk*	*ba'-an*	*suk.*
TEI		coconut	one	fall-REAL	come.down
		'A coconut fell down.'			

	b.	*A*	*wat*	*u*	*pua-n*	*moxod-an*	*gula'.*
		3SG	coconut	DIST	snap-REAL	drop-REAL	finish
		'He picked and dropped that coconut.' (i.e., he had climbed the coconut tree)					

Teiwa also analytical causative constructions where the verb *er* 'make' takes P as its argument, as in (58). The referent of P is identical to the referent of the S of the following

[27] The Blagar causative prefix is either a copy of the first stem vowel, or it is the vowel a-. For example: the causative of tia 'sleep' is i-tia in north Blagar and a-tia in south Blagar.

verb.

(58)	*Na*	*motor*	*er-an*	**(a)*	*sig.*
TEI	1SG	motorbike (IND)	make- REAL	3SG	live

'I switch on the motorbike.' (lit. 'I make the motorbike live')

The lexical causative of Sar is illustrated in (59). The P is introduced as argument of the verb *ma* 'come', which is part of a serial verb construction. Whether Sar also has an analytical causative like Teiwa remains to be investigated.

(59) a.

Wat	*fo*	*baak.*
coconut	that	fall

'That coconut fell.' (Baird, survey data 2003)

b.

A	*wat*	*ma*	*hod.*
3SG	coconut	come	drop

'He drops coconuts'

In Kaera, a causative verb is derived by suffixing the intransitive verb with a causative suffix *-ng*. The causee is prefixed to the derived verb, as in (60b):

(60) a.

Wat	*nuk*	*ba*	*sero.*
coconut	one	fall	descend

'A coconut fell down.'

b.

Gang	*e-wat*	*ga-ba-ng.*
3SG	2SG-coconut	3SG-fall-CAU

'He drops your coconut.'

Blagar and Adang also employ a causative suffix (*-ng* in Blagar (61b), *-ing* in Adang (63)), while they also have a causative prefix. The causative prefix consists of a single vowel (*a-*). The causative verb may be preceded by an object prefix encoding the causee, as illustrated for Blagar in (61b), and for Adang in (63). While all Adang causatives have a prefix, not all have suffixes, as illustrated in (62b) (for more discussion, see Haan 2001).

(61) a.

Vet	*ʔangu*	*ba-t*	*hera.*
coconut	that	fall-t	down

'A coconut fell down.' (Hein Steinhauer, p.c. 2010)

b.

ʔana	*vet*	*ʔ-a-ba-ng.*
3SG	coconut	3SG-CAU-fall-CAU

'He drops (a) coconut.' (Hein Steinhauer, p.c. 2010)

(62) a. *John ʔol.*
Adang John fall.over
'John falls over.'

b. *John na-ri a-ʔol.*
John 1SG-ACC CAU-fall.over
'John made me fall over.' (Haan 2001: 253)

(63) *Ella Ani ʔ-a-mih-ing-am*
Adang Ella Ani 3-CAU-sit-CAU- PRF
'Ella has made Ani sit down.' (Haan 2001: 257)

In contrast to the lexical and morphological causatives found in the AP languages discussed above, Alorese employs analytical causatives: one with the verb *n(e)ing* 'give' (64), and one with the verb *lelang* 'make' (65).

(64) a. *Tapo goka.*
Alor coconut fall
'A coconut fell.'

b. *No neing goka mo tapo.*
3SG give fall 2SG coconut
'He drops your coconut.'

(65) *Mo lelang bola meja ni leing.*
Alor 2SG make break table POSS leg
'You broke the table's leg.'

In Proto-Austronesian, a causative of dynamic verbs was marked with *pa-* and a causative of stative verbs with the prefixes *pa-ka-* (Blust 2009a: 359). Many modern Austronesian languages still use reflexes of *pa-(ka-)* (sometimes along with other affixes) to derive causative verbs. However, many modern languages also use lexical causatives, or periphrastic constructions with 'give', for example in spoken Indonesian and many Malay varieties. In Lamaholot, causatives are expressed by analytical constructions with *nein* 'give' (N&K 2007: 118) or *–a'an* 'make' (N&K 2007: 82) in patterns identical to those found in Alorese.

4.3.4. Summary of syntactic comparison. The structural contrasts discussed above are represented in (66). (A = agent, T = displaced theme, R = recipient, I = instrument, P = patient).

(66) a. Give construction 'A gives T to R'

Teiwa	A	T	OBL	R	give
Sar	A	T	OBL	R	give

Kaera	A	T	OBL	R	give			
Blagar	A	T	OBL	R	give			
Adang	A	T	TAKE	R	give			
Alorese	A				give	R	T	
Lamaholot	A				give	R	T	

b. Instrument ‘A cuts/shortens P with I’

Teiwa	A	I	OBL	P	cut			
Sar	A	I	OBL	P	cut			
Kaera	A	I	OBL	P	cut			
Blagar	A	I	take	P	shorten			
Adang	A	I	hold	P	cut			
Alorese	A				cut	P	with	I
Lamaholot	A				cut	P	use	I

c. Causative ‘A causes P to V’

	Lexical	Morphological	Analytical
Teiwa	yes		[A P make] [S V]
Sar	yes		
Kaera	yes?	Suffix	
Blagar	yes?	Prefix & suffix	
Adang	yes?	Prefix & suffix; prefix	
Alorese			[A give V P]
			[A make V P]
Lamaholot			[A give V P]
			[A make V P]

While the AP languages compared here all express ‘give’ and instrument constructions in a similar way, Alorese employs different constructions. In the expression of causatives, the AP languages show much internal variation, but the pattern used in Alorese does not appear to be related to any of the constructions found in the AP languages. (Note that both Teiwa and Alorese have an analytic causative with ‘make’, but the word orders are different.) In general, Alorese does not diverge from the patterns found in Lamaholot.

In sum, the data presented here provide no evidence that Alorese borrowed grammatical constructions from its Papuan neighbors (and neither did the neighbors borrow from Alorese). At the same time, we find that the Alorese constructions are virtually identical with Lamaholot, suggesting that Alorese retained the syntax of Lamaholot.

4.4. Conclusions. A comparison of the Alorese lexicon and grammar with the lexicon and grammar of neighbouring Papuan languages suggests: (i) Alorese borrowed a small set of words from the basic vocabulary of different AP Papuan languages across Pantar island, no language being more dominant the others; (ii) Alorese did not borrow any of the grammatical constructions to express ‘give’ events, instrumentals or causatives (and neither did the neighbours borrow from Alorese). Instead, Alorese has retained the syntax of Lamaholot.

5. Historical and ethnographic notes. In this section I summarize the historical and ethnographic evidence from which we may infer (i) that the speakers of Alorese moved away from the area where Lamaholot is spoken today (and not the other way around), and (ii) the date before which the split must have occurred.[28]

In Anonymous (1914:75-78)[29] a distinction is made between the mountain populations of Alor and Pantar and the populations on the coast. The coastal people are considered 'niet inheemsch' ('non-indigenous', p. 77). The paper also reports the local legend that Pandai, in north west Pantar, was the first coast to be populated by these non-indigenous coastal people.

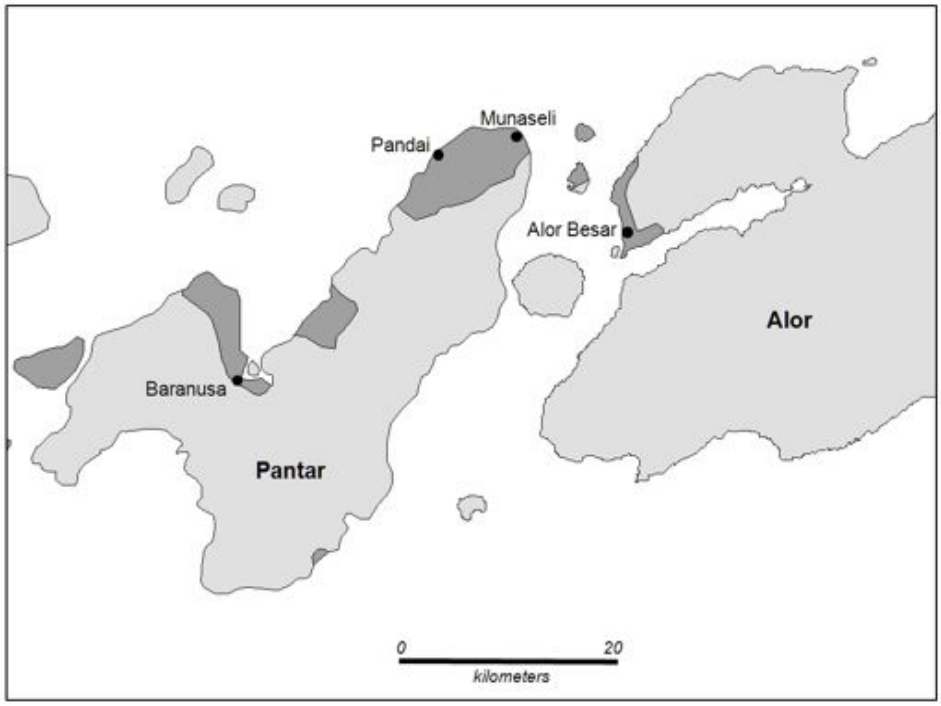


Figure 3. Pantar island with the location of Baranusa, Pandai, Munaseli and Alor Island with the town Alor Besar

Today, Alorese speaking communities are only found in coastal areas of Alor and

[28] The proposed date is not an absolute date but a 'terminus ante quem': the split may have occurred any time before this date.

[29] Major sources of this article were (i) the 'Militaire Memories' (reports on military expeditions that took place on the islands in 1910 and 1911, and (ii) a report of a geological expedition by R.D.M. Verbeek in 1899, published 1908 as 'Molukken Verslag' in J*aarboek van het Mijnwezen in Ned. Oost-Indie.*

Pantar. They are sea-oriented, and for subsistence they traditionally rely on fishing (the men) and weaving (the women). They currently adhere to the Islam religion.

In contrast, speakers of the Papuan languages on Alor and Pantar are inland-oriented, have their traditional villages up in the mountains rather than the coast,[30] adhere to the traditional animist religion, or are Christians. They are farmers and do not rely on fishing or weaving for subsistence.

Traditionally, the coastal Alorese clans exchanged fish and woven cloth for food crops with the Papuan inland populations (cf. Anonymous 1914:76, 81-82). The Alorese clans were, at least initially, quite small. As an example, Anonymous (1914:89-90) mentions clans of 200-300 people. As newcoming clans inhabiting coastal locations geographically remote from each other, many Alorese clans must have been outnumbered by their Papuan neighbors, and it is plausible that they acquired their spouses from the exagamous Papuan clans in their immediate vicinity, rather than from the Alorese clans that were more remote.[31]

According to a legend reported in Anonymous (1914:77), a "colony of Javanese" settled on Pandai, in north west Pantar, some "500 to 600 years ago" [as the article appeared in 1914, this would now be 600 to 700 years ago, i.e. the colony settled on Pandai around 1300-1400 AD]. However, the same source includes a footnote (p. 89) which explains that the notion *orang djawa* (lit. 'Javanese people') applies to everyone who comes from other parts of the archipelago.[32] In other words, the "Javanese" coastal settlers mentioned in the legend were people from "overseas", but not necessarily from Java. Instead, I propose that the close linguistic and cultural ties between today's Alorese and Lamaholot speakers suggest that the colony of *orang djawa* that settled on Pandai according to the legend were in fact Lamaholot speakers from Flores, Solor, and/or south Lembata.

The legend of the founding of Pandai in north Pantar referred to in Anonymous (1914) is also reported in Lemoine (1969) and cited in later sources such as Barnes (1973:86, 2001:280) and Rodemeier (2006). It recounts that two Javanese brothers, Aki Ai and his younger brother Mojopahit, sailed to Pantar, where Aki Ai treacherously abandoned Mojopahit. Mojopahit's descendants eventually colonized Pandai, Baranusa, and Alor Besar. A second legend in Lemoine (1969) recounts of another kingdom on Pantar, the kingdom of Munaseli, located more eastwards on the northern coast. In the legend, Javanese

---

[30] Although many have now moved to villages on the coast for practical purposes.

[31] Clans exchanged wives, but people were also sold or given away as slaves. For example, Teiwa (north-west Pantar) has a word *yu'al* which is translated as 'to give away (people)' (cf. Teiwa *'an* 'to sell'), and it refers to an "old custom" of "sending or giving away people that are useless to the clan". Speakers noted that formerly, yu'al was also used to refer to selling people (including women) to the Baranusa people (Klamer 2010a:41, fn. 2.). Baranusa is an Alorese speaking area.

[32] Compare Kambera (Sumba) *tau Jawa* 'stranger' (lit. 'Javanese person') and *tau Jawa bara* 'westerner' (lit. 'white Javanese') where *Jawa* also denotes 'stranger' (Onvlee 1984: 115).

immigrants who are allied to the kingdom of Pandai, kill the king of Munaseli and destroy his kingdom sometime between 1300-1400 AD. The defeated Munaseli population fled to Alor Besar, on the Alor peninsula (see figure 2).

Other sources confirm that around 1300-1400 AD the influence of the Hindu-Javanese kingdom Majapahit indeed extended to Pantar: the Javanese Nagarakertagama chronicles (1365) contain a list of places in the east that were in the Majapahit realm, including 'Galiyaho' (Hägerdal 2010).[33] The name Galiyahu or Galiyao occurs in a number of 16th and 17th century maps and descriptions by Europeans, and general consensus exists that Galiyahu/Galiyao refers to Pantar (see Le Roux 1929:47, Barnes 1982:407, Dietrich 1984, Rodemeier 1995, Barnes 2001:277, Rodemeier 2006, Hägerdal 2010). Recent linguistic research by Gary Holton and myself on Pantar island revealed that Galiyao is used in various local languages as the indigenous name to refer to the island of Pantar; the name originates from Western Pantar language *Gale Awa*, literally 'living body' (Holton 2010).[34]

Today, Pandai and Munaseli are Alorese speaking areas in northern Pantar. *Tanjung Muna* ('Cape Muna') in North Pantar is still considered the location of the mythical kingdom Munaseli. The language spoken there is referred to in Indonesian as *Bahasa Muna* 'the Muna language', an abbreviation of *Munaseli*. Speakers refer to their own language as *Kadire Senaing* 'Speech we Understand' (Rodemeier 2006:49), and the *Bahasa Muna* or *Kadire Senaing* reported in Rodemeier 2006 is (a dialect of) Alorese.

Alorese is currently also spoken along the coast of the Alor Bird's Head peninsula, and the ancestors of these speakers are probably related to the Muna(seli) population that fled to Alor after their defeat in Pantar by early 1400.

In sum, from historical, ethnographic and linguistic observations we can infer that Galiyahu was Pantar, that Pantar was under the influence of the Majapahit kingdom in 1300-1400 AD which is evidence that the island was known far beyond its immediate neighboring territories. Both the Pandai and Munaseli kingdom in Pantar were in place around 1300-1400 AD in North-Northeastern Pantar, having been established by immigrants speaking an Austronesian language. In the early 15th century, at least one group fled from Pantar to Alor to settle in Alor Besar, on the Alor peninsula. Today the settlements Pandai, Munaseli, Alor Besar and Baranusa still exist, and all of them coincide with locations where Alorese is spoken, so we can safely assume that today's Alorese populations are descendants from clans that settled on Pantar.

---

[33] The influence of Majapahit in the Lesser Sunda Islands did not imply actual political or cultural involvement, as no Majapahit archeological remains have been found in the area.

[34] "The appropriateness of this name is evidenced by the presence of an active volcano which dominates southern Pantar. This volcano regularly erupts, often raining ash and pyroclastic flows onto villages of the region. Even when it is not erupting, the volcano ominously vents sulfur gas and smoke from its crater. In a very real sense, the volcano is a living body." (Holton 2010).

Given the close linguistic and cultural ties between Alorese and Lamaholot, I conclude that the ancestors of the Alorese were Lamaholot speakers from Solor, Lembata, Adonara and/or east Flores. They arrived at the coasts of Pantar before or around 1300-1400 AD.

6. Summary and discussion. A number of shared syntactic features which signal Papuan influences are found in both Lamaholot and Alorese, and must have been part of Proto-Lamaholot. This suggests (prehistoric) Papuan presence in the Lamaholot homeland, which may have been located in east Flores and/or the islands Solor, Lembata and Adonara. The Papuan influence on Proto-Lamaholot was strong enough to increase the complexity of Proto-Lamaholot: an increase in word order patterns, the introduction of an inalienable noun distinction and variable possessor marking structures, as well as a new functional item, the focus marker. Where language contact leads to an increased linguistic complexity with additive features, the language is likely to have been spoken in a community with high degrees of outside contacts (Trudgill 2010: 304). The contact must have been long-term, and have involved language acquisition of pre-adolescents ('pre-critical threshold contact situations', Trudgill 2010: 304, 315).

Proto-Lamaholot had a fairly rich morphology, including possessor suffixes, distinct pronominal affixes for A and S, and at least seven derivational prefixes. After it split from Lamaholot, Alorese underwent a process of simplification: it lost all of the Proto-Lamaholot derivational and inflectional morphology, including the marked distinction between A and S; the variable possessor marking structures were regularized, and the final nasal morpheme on inalienable nouns was reinterpreted as a root-final consonant segment.

After they arrived on Pantar island, either before or during the 14th century, the Alorese did not borrow much vocabulary from their Alor-Pantar neighbours. The limited number of identified loans come from different AP languages across Pantar, none of which appears to have been dominant. Alorese retained the syntax of Lamaholot, simplifying and regularizing some of its irregularities, and the influence of local AP syntax on Alorese appears to have been minimal: Alorese moved its time adverb to postverbal position, and adopted a clause final conjunction-like element.

The limited lexical congruence and virtual absence of syntactic influences suggests a contact scenario that neither involved prolonged stable bilingualism, nor Papuan speaking communities shifting to Alorese. However, the morphological and syntactic simplification of Alorese suggests that the language went through a stage of second language learning. This combination of facts is indeed puzzling.

There is evidence that Alorese was spoken as non-native language: it was used as a regional trade language (Anonymous 1914, Stokhof 1975:8); and intensive trade relations existed between the coastal Alorese and the Papuan populations living in the Pantar mountains, exchanging e.g. woven cloth for food (cf. Anonymous 1914:76, 81-82).

As the Alorese settlements on the coasts of Pantar and Alor were initially quite small, and geographically remote from each other, it is likely that, initially, the Alorese men acquired their spouses from one of the various exogamous communities in their vicinity where an AP language was spoken. As a result, women speaking AP languages were brought into a community that spoke a language similar to Proto-Lamaholot. Trying to learn this language as adults, the women simplified its morphology, and their learner's omissions became part

of a morphologically simplified variety that developed into the morphologically isolating Alorese language as acquired by their children. Inflectional morphology is known to be seriously problematic for post-adolescent second language learners who have passed the 'critical threshold' (Lenneberg 1967) for language acquisition (Kusters 2003:21, 48, citing Clahsen and Muysken 1996, Meisel 1997). And derivational morphology, being partly lexicalised, irregular and semantically opaque, represents arbitrary grammatical patterns which must be learned without any generalization possible, which is equally difficult for post-threshold language learners.

The loss of inflectional and derivational morphological categories in Alorese can thus be seen as an instance of simplification that occurred as a result of non-native adult language learning (Trudgill 2010: 310-313). In general, simplification is most likely to occur in intense contact situations that are short-term and post-critical threshold (Trudgill 2010: 310-315).

The questions that are not answered by this scenario include the following. Did the Papuan mothers introduce more of their Papuan words and syntax into the Alorese they spoke as second language? If they did, why did their children not aquire this along with their morphologically simplified Alorese? Or was there community pressure to speak Alorese in its lexically and syntactically 'pure' form, while omitting its morphology was allowed? Additional sociolinguistic research on the social position and language attitude as well as studies of actual speech of newcomers into Alorese communities may help to shed some light on this.

In the history of Alorese reconstructed here, we see that at different time depths, different language contact situations had different consequences for the structure of the language. Prehistoric, deep time contact between a Papuan substrate and Proto-Lamaholot resulted in a complexification of Proto-Lamaholot, while later, post-migration contact resulted in a simplification. While both outcomes suggest that the contact was intense, the sociolinguistic situations were presumably different: prehistoric contact with Papuan languages in the Flores area was long-term and involved pre-adolescents, while the post-migration contact that took place after settlement on Pantar was short-term, and involved post-adolescent learners. There is no evidence that since that period, linguistic contacts between Alorese and the speakers of AP languages around them have been any more than superficial.

References

Anonymous, 1914. De eilanden Alor en Pantar, Residentie Timor en Onderhoorigheden. *Tijdschrift van het Koninklijk Nederlandsch Aardrijkskundig Genootschap* 31. 70-102.

Alor Pantar Lexical Database. A compilation of 200+ item survey words lists of 18 languages of Alor and Pantar, collected between 2003-2010 by Louise Baird, Gary Holton, Marian Klamer, František Kratochvil, Laura Robinson & Antoinette Schapper. Leiden University and University of Alaska Fairbanks.

Arndt, Paul P. 1937. *Grammatik der Solor-Sprache*. Ende, Flores: Arnoldus-Drukkerij.

Baird, Louise. 2002. *A grammar of Keo: An Austronesian language of East Nusantara.* Canberra: ANU PhD thesis.

Baird, Louise. 2008. *A grammar of Klon.* Canberra: Pacific Linguistics.

Barnes, Robert H. 1973. Two terminologies of symmetric prescriptive alliance from Pantar and Alor in Eastern Indonesia. *Sociologus; Zeitschrift für Völkerpsychologie und Soziologie* 23. 71-88.

Barnes, Robert H. 1982. The Majapahit Dependency Galiyao. *Bijdragen tot de Taal-, Land-, en Volkenkunde* 138(4). 407-12.

Barnes, Robert H. 1996. *Sea hunters of Indonesia: Fishers and weavers of Lamalera*. (Oxford Studies in Social and Cultural Anthropology). Oxford: Oxford University Press.

Barnes, Robert H. 2001. Alliance and warfare in an Eastern Indonesian principality. Kedang in the last half of the nineteenth century. *Bijdragen tot de Taal-, Land-, en Volkenkunde* 157. 271-311.

Blust, Robert. 1993. Central and central-eastern Malayo-Polynesian. *Oceanic Linguistics* 32. 241–293.

Blust, Robert. 2009a. *The Austronesian languages.* Canberra: Pacific Linguistics.

Blust, Robert. 2009b. The position of the languages of Eastern Indonesia: A reply to Donohue and Grimes. *Oceanic Linguistics* 48(1). 36-77.

Clahsen, Harald & Pieter Muysken. 1996. How adult second language learning differs from child first language development. *Behavioural and Brain Sciences* 19. 721-723.

Clark, R. 1990. The Austronesian languages. In Bernard Comrie (ed.), *The major languages of East and South East Asia*, 173-184. London: Routledge.

Dietrich, Stefan. 1984. A note on Galiyao and the early history of the Solor-Alor Islands. *Bijdragen tot de Taal-, Land-, en Volkenkunde* 140(2/3). 317-25.

Donohue, Mark. 2004. Typology and linguistic areas. *Oceanic Linguistics* 43(1). 221-239.

Donohue, Mark. 2007. The Papuan language of Tambora. *Oceanic Linguistics* 46(2). 520-537.

Donohue, Mark & Charles E. Grimes. 2008. Yet more on the position of the languages of eastern Indonesia and East Timor. *Oceanic Linguistics* 47(1). 114-158.

Doyle, Matthew. 2010. Internal divisions of the Flores-Lembata subgroup of Central Malayo-Polynesian. Leiden: Leiden University MA thesis.

Ewing, Michael & Marian Klamer (eds.). 2010. *East Nusantara: Typological and areal analyses* (Pacific Linguistics 618). Canberra: Pacific Linguistics.

Florey, Margaret. 2010. Negation in Moluccan languages. In Ewing & Klamer, *East Nusantara,* 227-250.

Foley, William A. 2000. The languages of New Guinea. *Annual Review of Anthropology* 29. 357-404.

Gordon, Raymond G., Jr. (ed.). 2005. *Ethnologue: Languages of the world*, *15th edn*. http://www.ethnologue.com. (1 August, 2011.)

Greenhill, Simon. J., Robert Blust & Russell D. Gray. 2008. Austronesian Basic Vocabulary Database. http://language.psy.auckland.ac.nz/austronesian. (August 2011.)

Grimes, Charles E. 1991. The Buru language of Eastern Indonesia. Canberra: ANU PhD thesis.

Grimes, Charles E., Tom Therik, Barabara Dix Grimes & Max Jacob. 1997. *A guide to the people and languages of Nusa Tenggara.* Kupang: Artha Wacana Press.

Haan, Johnson W. 2001. The grammar of Adang: A Papuan language spoken on the Island of Alor, East Nusa Tenggara, Indonesia. Sydney: University of Sydney PhD thesis.

Hägerdal, Hans. 2010. Cannibals and pedlars. Economic opportunities and political alliance in Alor, 1600-1850. *Indonesia and the Malay World* 38. 217-246.
Himmelmann, Nikolaus P. 2005. The Austronesian languages of Asia and Madagascar: Typological characteristics. In K.A. Adelaar and N.P. Himmelmann (eds.), *The Austronesian languages of Asia and Madagascar.* 110-181. London: Routledge.
Holton, Gary. 2010. *An etymology for Galiyao*. Alaska: University of Fairbanks MA thesis.
Holton, Gary, Marian Klamer, Frantisek Kratochvíl, Laura Robinson & Antoinette Schapper. 2012. The historical relation of the Papuan languages of Alor and Pantar. *Oceanic Linguistics* 51(1). 87-122.
Jacob, June & Charles E. Grimes. 2003. *Kamus Pengantar Bahasa Kupang.* Kupang: Artha Wacana Press.
Jones, Russell (ed.). 2007. *Loan words in Indonesian and Malay*. Leiden: KITLV Press.
Keraf, Gregorius. 1978. *Morfologi dialek Lamalera.* Jakarta: Universitas Indonesia PhD dissertation.
Klamer, Marian. 1996. Kambera has no passive. In Marian Klamer (ed.), *Voice in Austronesian* (NUSA Linguistic Studies of Indonesian and Other Languages in Indonesia 39), 12-30. Jakarta: Universitas Atma Jaya.
Klamer, Marian. 2002. Ten years of synchronic Austronesian linguistics (1991-2002). *Lingua* 112. 933-965.
Klamer, Marian. 2010a. *A grammar of Teiwa*. Berlin: De Gruyter.
Klamer, Marian. 2010b. Ditransitives in Teiwa. In Andrej Malchukov, Martin Haspelmath & Bernard Comrie (eds.), *Studies in ditransitive constructions*. Berlin: De Gruyter.
Klamer, Marian. 2011. *A short grammar of Alorese (Austronesian)*. Munich: Lincom Europa.
Klamer, Marian. n.d. *Teiwa corpus*. Leiden: Leiden University.
Klamer, Marian. Forthcoming. From Lamaholot to Alorese: Simplification without shift. In David Gil & John McWhorter (eds.), *Austronesian undressed: How and why languages become isolating.*
Klamer, Marian & Michael Ewing. 2010. The languages of East Nusantara: An introduction. In Ewing & Klamer, *East Nusantara*, 1-24.
Klamer, Marian & Frantisek Kratochvíl. 2010. Abui Tripartite Verbs: Exploring the limits of compositionality. In Jan Wohlgemuth & Michael Cysouw (eds.), *Rara & Rarissima: Documenting the fringes of linguistic diversity* (Empirical Approaches to Language Typology 46), 209-233. Berlin: De Gruyter.
Klamer, Marian, Ger Reesink & Miriam van Staden. 2008. East Nusantara as a linguistic area. In Pieter Muysken (ed.), *From linguistic areas to areal linguistics*, 95-149. Amsterdam: Benjamins.
Klamer, Marian & Antoinette Schapper. 2012. The development of 'give' constructions in the Papuan languages of Timor-Alor-Pantar. *Linguistic Discovery* 10.3. 174-207.
Kratochvíl, Frantisek. 2007. *A grammar of Abui*. [Leiden University PhD thesis.] Utrecht: LOT.
Kusters, Wouter. 2003. *Linguistic complexity.* [Leiden University PhD thesis.] Utrecht: LOT Publications.
Lenneberg, Eric. 1967. *Biological foundations of language*. New York: Wiley.
Lemoine, Annie. 1969. Histoires de Pantar. *L'Homme* 9(4). 5-32.

Le Roux, C.C.F.M. 1929. De Elcano's tocht door den Timorarchipel met Magelhaes' schip Viktoria. *Feestbundel Koninklijk Bataviaasch Genootschap van Kunsten en Wetenschappen 1778-1928, II*. 1-70. Batavia: Kolff.

Lewis, M. Paul (ed.). 2009. *Ethnologue: Languages of the world, 16th edition*. http://www.ethnologue.com. (18 August, 2011.)

Lewis, E. Douglas & Charles Grimes. 1995. Sika: Introduction and word list. In Tryon, *Comparative Austronesian dictionary*, 1:601-9.

Lynch, John, Malcolm Ross & Terry Crowley. 2002. *The Oceanic languages.* Richmond, Surrey: Curzon.

Meisel, Jurgen. 1997. The acquisition of the syntax of negation in French and German: contrasting first and second language development. *Second Language Research* 13. 227-63

Nagaya, Naonori. 2009a. Space and motion in Lamaholot. Paper presented at the 11th International Conference on Austronesian Linguistics, Aussois, France.

Nagaya, Naonori. 2009b. Subject and topic in Lamaholot. Paper presented at the 11th International Conference on Austronesian Linguistics, Aussois, France.

Nishiyama, Kunio and Herman Kelen. 2007. *A grammar of Lamaholot, Eastern Indonesia: The morphology and syntax of the Lewoingu dialect*. Munich: Lincom Europa.

Onvlee, Louis.1984. *Kamberaas-Nederlands woordenboek*. Leiden: KITLV.

Pampus, Karl-Heinz. 1999. *Koda Kiwa: Dreisprachiges Worterbuch des Lamaholot (Dialekt von Lewolema)* (Abhandlungen fur die Kunde des Morgenlandes 52.4). Stuttgart: Steiner.

Pampus, Karl-Heinz (with help of Yohanes E. Lamuri). 2001. *Mue Moten Koda Kiwan: Kamus Bahasa Lamaholot, Dialek Lewolema, Flores Timur*. Frankfurt: Frobenius-Institut Frankfurt am Main.

Pawley, Andrew K. 2005. The chequered career of the Trans New Guinea hypothesis. In Pawley et al., *Papuan pasts*, 67-107.

Pawley, Andrew, Robert Attenborough, Jack Golson & Robin Hide (eds.). 2005. *Papuan pasts: Cultural, linguistic and biological histories of Papuan-speaking peoples.* Canberra: Pacific Linguistics.

Reesink, Ger P. 2002. Clause-final negation: Structure and interpretation. *Functions of Language* 9. 239-268.

Robinson, Laura & Gary Holton. 2012. Reassessing the wider genealogical affiliations of the Timor-Alor-Pantar languages. *Language and Linguistics in Melanesia* (Special issue 2012 Part 1). 59-87.

Rodemeier, Susanne. 1995. Local tradition on Alor and Pantar. An attempt at localizing Galiyao. *Bijdragen tot de Taal-, Land- en Volkenkunde* 151. 438-442.

Rodemeier, Susanne. 2006. *Tutu Kadire - Erzählen und Erinnern lokalgeschichlicher Mythen am Tanjung Muna in Ostindonesien.* Leipzig: University of Leipzig PhD dissertation.

Ross, Malcolm. 2002. History and transitivity of western Austronesian voice. In Fay Wouk & Malcolm Ross (eds.), *The history and typology of Western Austronesian voice systems.* 17-62. Canberra: Pacific Linguistics.

Ross, Malcolm. 2005. Pronouns as preliminary diagnostic for grouping Papuan languages. In Pawley et al., *Papuan pasts*, 15-65.

Samely, Ursula. 1991. *Kedang, (Eastern Indonesia): Some aspects of its grammar.* Hamburg: Buske.

Schapper, Antoinette, Juliette Huber & Aone van Engelenhoven. 2012. The historical relations of the Papuan languages of Timor and Kisar. *Language and Linguistics in Melanesia* (Special issue 2012 Part 1). 192-240.

Schapper, Antoinette & Marian Klamer. ms. Cardinal numerals in the Papuan languages of Alor-Pantar: History and typology. Leiden University. Unpublished ms.

Steinhauer, Hein. 1993. Bahasa Blagar Selayang Pandang. Penyelid Bahasa dan Perkembangan Wawasannya I, 639-659. Jakarta: Masyarakat Linguistik Indonesia.

Stokhof, W.A.L. 1975. *Preliminary notes on the Alor and Pantar languages (east Indonesia)*. (Pacific Linguistics B 43). Pacific Linguistics, Research School of Pacific Studies, The Australian National University.

Trudgill, Peter. 2010. Contact and sociolinguistic typology. In Raymond Hickey (ed.), *The handbook of language contact,* 299-319. Malden, MA: Wiley-Blackwell.

Tryon, Darrell T. 1995. The Austronesian languages. In Tryon, *Comparative Austronesian dictionary*, 1.5-44.

Tryon, Darrell T. (ed.). 1995. *Comparative Austronesian dictionary: An introduction to Austronesian studies* (Trends in Linguistics: Documentation 10). 4 parts. Berlin & New York: Mouton De Gruyter.

Wurm, Stephen A., C.L. Voorhoeve & Kenneth A. McElhanon. 1975. The Trans-New Guinea phylum in general. In Stephen A. Wurm (ed.), *Papuan languages and the New Guinea linguistic scene* (Pacific Linguistics C 38), vol. 1 of *New Guinea area languages and language study*, 299-322. Canberra: Pacific Linguistics.

Marian Klamer
M.A.F.Klamer@hum.leidenuniv.nl

5

Language Documentation & Conservation Special Publication No. 5 (December 2012)
Melanesian Languages on the Edge of Asia: Challenges for the 21st Century,
ed. by Nicholas Evans and Marian Klamer, pp. 109–149
http://nflrc.hawaii.edu/ldc/sp05/
http://hdl.handle.net/10125/4562

Even more diverse than we had thought: The multiplicity of Trans-Fly languages

Nicholas Evans
Australian National University

Linguistically, the Trans Fly region of Southern New Guinea is one of the least known parts of New Guinea. Yet the glimpses we already have are enough to see that it is a zone with among the highest levels of linguistic diversity in New Guinea, arguably only exceeded by those found in the Sepik and the north coast. After surveying the sociocultural setting, in particular the widespread practice of direct sister-exchange which promotes egalitarian multilingualism in the region, I give an initial taste of what its languages are like. I focus on two languages which are neighbours, and whose speakers regularly intermarry, but which belong to two unrelated and typologically distinct families: Nen (Yam Family) and Idi (Pahoturi River Family). I then zoom out to look at some typological features of the whole Trans-Fly region, exemplifying with the dual number category, and close by stressing the need for documentation of the languages of this fascinating region.

1. Introduction.[1] The distribution of linguistic diversity is highly informative, about

[1] My thanks to two anonymous referees and to Marian Klamer for their usefully critical comments on an earlier version of this paper. I gratefully acknowledge the support of the Australian National University (Professorial Setup Grant) and the Australian Research Council (Discovery Project 'Languages of Southern New Guinea') for supporting my fieldwork in Southern PNG, the Linguistics Society of America for support to teach a Field Methods course on Idi at the Boulder Linguistics Institute in 2011, as well as the ANRC for funding enabling me to attend the Manokwari conference, to members of the audience there for their helpful discussion, and Jeff Siegel, Christian Döhler, Grahame Martin and Garrick Hitchcock for access to unpublished materials drawn on here. Most importantly I thank my Nen and Idi teachers, particularly Michael Binzawa, †Aramang Wlila, Jimmy Nébni and Wasang Baiio, for their insightful and dedicated efforts to teach me their languages. Material on Idi comes predominantly from two sources: recordings made with Wasang Baiio during a Field Methods course at the LSA Institute in Boulder, Colorado in July-Aug 2011, and material recorded from Mr Gus Iamatta (Ymta) in 2010, who at that time was the school principal at Bimadbn community school. I would also like to thank Ewelina Wnuk, Kate Miller, Rebecca Defina and Grant Aiton who during the Field

history, social configurations, and ideologies of language use. Over the last four decades of scholarship, New Guinea's position as the most linguistically diverse region of the planet has not changed, but received views of where the most deep-level diversity lies within New Guinea have moved substantially. Various versions of the Trans-New Guinea hypothesis have led to hundreds of languages centred on the cordillera being joined into a single family of (sometimes only distantly) related languages, whereas the progression of research on the Sepik has found a mosaic of small families and isolates – a pattern taken to be more representative of New Guinea as a whole before the spread of Trans-New Guinea languages.

Southern New Guinea – and more particularly that part of it known as the Trans-Fly (fig. 1)[2] –has not yet figured prominently in assessments of where the most diversity lies. Though it has sometimes been mentioned (e.g. Pawley 2008:51) as 'a smaller region of high diversity', existing assessments tend to lump together several families on little evidence: both Pawley (2005) and Ross (2005:30-31) essentially reproduce Wurm's earlier lumper classification of what I will argue are several distinct families in the Trans-Fly region.

Methods course elicited some of the material cited here in small-group sessions. Material on Warta Thundai comes from a field methods course taught in February 2011 with Sembara Dibara whom I thank for his enthusiastic participation.

[2] There is no universally accepted definition of the extent of 'Trans-Fly': it tends to be well-defined at its eastern and northern extremities (by the Fly River) and to the south by the Torres Strait but, as one moves west, geographical boundaries give way to political ones, as in Williams' (1936) 'the south-west corner of Papua' (where Papua meant the [then] Australian territory of Papua). From the ecological point of view, however, it makes sense to consider the Trans-Fly Region as extending somewhat further west, taking in Kolopom Island, as is done in Fig. 1, and this is the term that has been adopted by conservationist groups like the World Wildlife Fund. For present purposes I will take it to extend across the (modern) national boundary to the Merauke River, in traditional Marind territory. Southern New Guinea is of course bigger than this, with many other linguistic groupings which I do not discuss here, such as Yelmek-Maklew, which Ross (2005) treats as non-TNG but part of a 'South-Central Papuan' family without adducing any evidence of formal cognacy across the three branches. A full discussion of these languages is beyond the scope of this article, but the existence of further groups to the west of the Trans-Fly merely amplifies the point I am making here.

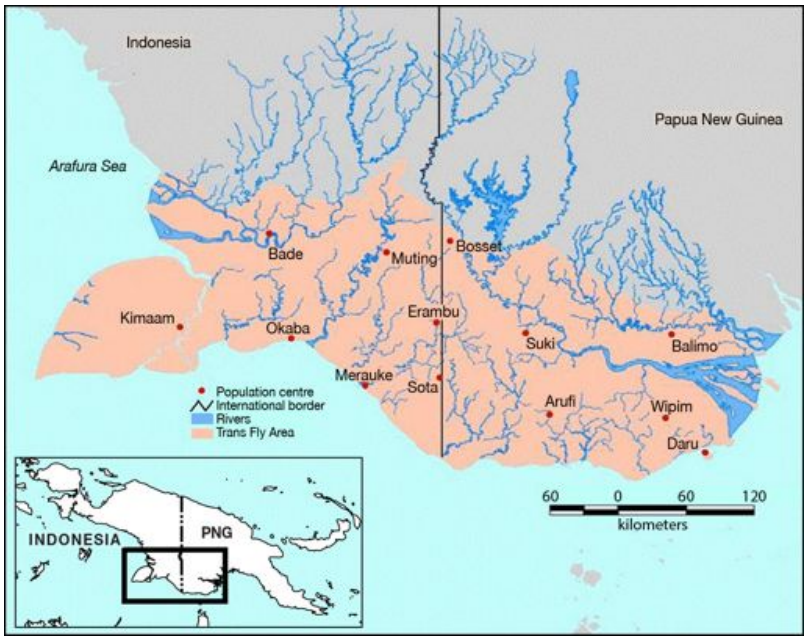


FIGURE 1. The Trans-Fly Region, as defined by the WWF (WWF Transfly Team 2006)

In this paper I will argue that Southern New Guinea in fact contains more deep diversity than has hitherto been realised, with somewhere between five and eight unrelatable families taking in forty or so languages in an area about the size of the Netherlands. On top of that, there are major typological differences between the languages of these families, and many of them (such as the Yam and Pahoturi River families) diverge significantly from the picture of a 'typical Papuan language' that has arisen from studies centred in the Highlands, the Sepik or the islands to the east of the New Guinea mainland. Taken together, data from Southern New Guinea significantly amplifies our view of the overall level of diversity in New Guinea.

This diversity is even more astonishing given that the region did not even exist in its present form until recently and large parts of it were underwater following mid-Holocene sea-level rises until rebuilt by progradation from sediments brought down by the Fly and Digul rivers. It is thus unlikely that all language differences currently found in Southern New Guinea developed in situ. What seems more likely is that they represent the interaction of a number of unrelated groups entering the region from different regions as it became habitable land, combined with specific features favouring diversification such as the pattern of direct sister-exchange between small groups to be discussed in §2.1.2, which is likely to have created high levels of diversity in a multilingual population coupled with a valuation of very local markers of linguistic allegiance.

I structure the paper as follows. In §2 I give a basic description of the geography and ethnography of the region. In §3 I review the main linguistic groupings, as currently

understood, then in §4 go on to give brief portraits of Nen and Idi, two languages which, although close geographical neighbours bound by relations of marriage exchange and multilingualism, diverge significantly on a wide range of measures – not only are they in different phylogenetic groups, but their typological profiles also differ markedly. But divergence of this type does not mean there are no significant areal features across Southern New Guinea, and in §5 I illustrate this point with one such feature – three-valued number systems – while emphasising that the means of composing the dual value vary significantly from one language group to another. I close the article in §6 by summarising the key scientific challenges facing linguists as we confront a zone that is simultaneously one of the most diverse and one of the least-known regions of the logosphere.

2. Southern New Guinea as a geographical and cultural region. In its biota, such as its vegetation of eucalypts, melaleuca, acacia and banksias combined with wallabies, bandicoots, goannas, taipans and termite mounds, Southern New Guinea is more like northern Australia than like the rest of New Guinea.

Geographically, much of it is new, low land, a kind of tropical Netherlands built up over the last few millennia as the giant Fly River to the east (fig. 2) and the Digul and other rivers to the west have carried down and deposited sediment from the central cordillera. Compared to most of present-day New Guinea and Australia (except for the Sepik), it has had a turbulent geomorphological past over the last 10,000 years. The ancient land-bridge to Australia was severed by the rising seas around 9,000 b.p., and for a while higher sea-levels than today meant that some of what is now land was then submerged, before being rebuilt by progradation.

The northern parts are characterised by vast tracts of rainforest, with only the occasional clearing for a village, swidden garden or sago (fig. 3). Moving south, this gives way to eucalytus and melaleuca savannah reminiscent of northern Australia (fig. 4), and – around rivers like the Bensbach – extensive floodplains supporting massive populations of birds, wallabies and (now) deer. There is a marked monsoonal cycle, with a long dry season (July-November) alternating with an intense wet season (December–June). The length of the wet season increases as one heads north.

Staple foods vary somewhat across the area. In the Morehead district yams and other root crops predominate, based on swidden (slash-and-burn) agriculture which yields one year of fertile soils, followed by one or two years for less demanding crops like cassava and pineapples, then gradual reversion of the cleared area to forest over around 17 years, with mature coconut trees then the only sign of prior cultivation. Languages of the region contain numerous terms for different phases of cultivation – in addition to the generic word *kkp* for ‘garden’, Nen distinguishes *ḡayag* ‘new garden’, *kkp get kr* ‘old garden’ and *du* ‘abandoned overgrown garden’.

FIGURE 2. Aerial view of the Fly River, taken from the southwest, with the central cordillera visible far to the north (Photo: N. Evans)

FIGURE 3. Sago clearing in rainforest between Kiriwo and Fly River (Photo: N. Evans)

FIGURE 4. Jimmy Nébni in open savannah country with mixed melaleuca and eucalyptus vegetation, southern part of Nen-speaking area (Photo: N. Evans)

In this region great social value is placed on the accumulation of yams through expert gardening, with large traditional yam-feasts (Williams 1936) and counting-ceremonies based on powers of six, along with social stipulations also reckoned in powers of six, such as that a household needs to have 1,296 (6^4) stored in its yamhouse to feed it from one year to the next. Senary power terms, representing powers of six up to 6^5 or 6^6 are found throughout the Yam family (table 1; Evans 2009) but their extremely limited occurrence outside it[3] suggests that the development of this senary system is a linguistic innovation within the Yam family – either at proto-Yam level or, as Hammarström (2009) argues, at the level of the Tonda branch. We will not be able to resolve this question until better comparative data on sound correspondences is accumulated.

[3] Restricted to some very limited-use terms in Agöb and Idi which appear to be borrowings.

Value	Power	Nen (base)	Keraakie	Arammba	Kanum exponential term	Agöb (Buzi village)
6	6^1	*pus*	*(eembru) for*	*nimbo*		*put*
36	6^2	*prta*	*ferta (eembru) [peta]*	*feté*	*ptae*	*purta*
216	6^3	*taromba*	*taromba [tarumba]*	*tarumba*	*tarwmpao*	*tarumba*
1,296	6^4	*damno*	*daameno [dameno]*	*ndamno*	*ntamnao*	*damuno*
7,776	6^5	*wärämaka*	*werameka*	*wermeke*	*wrmaekr*	*waramakai*
46,656	6^6	*[]*	*wi*	*wi*		
279,936	6^7		*meemee wemb*			

TABLE 1. Base-six power in some languages of the Yam family, as well as from adjoining Agöb

In the swampier, more low-lying areas around the Bensbach and Torassi Rivers, there is evidence for the earlier use of mound-and-ditch agriculture to cultivate taro (Hitchcock 2010). And as conditions get wetter to the north and northwest, making burning off more difficult, yam gardens give way to sago as the main staple. Hunting is also important throughout the region, with cassowary, wallabies, bandicoots, wild pigs and (in modern times) deer all present in large numbers; in the savannah areas fire-drives were used to hunt wallabies in much the same way as in northern Australia. According to local tradition some peoples, such as some Pahoturi River groups, were until recently hunter-gatherers rather than gardeners.

In addition to the great cordillera-fed rivers, there are numerous shorter rivers running south into the Torres Strait from the low-lying Trans-Fly plateau. Historically these were important as supplementary waterways permitting war-canoes to penetrate far into the interior, thus playing a key role in depredations effected on speakers of the smaller language groups by huge war-parties of Marind from the west (as well as Kiwais to the east and Torres Strait peoples to the south).

2.1. PRECONTACT. Colonial contact began late in the region, and it was only early in the twentieth century that the respective colonial powers (at that time the Netherlands in the west and the British in the east) began to assert some control over large and ferocious armed groups such as the Marind (aka Tugere) to the west, the Kiwai to the east, and the Suki to the north. Indeed, it was British demands to the Dutch that they take responsibility for pacifying raids carried out by peoples within the latter's territory that led to the joint Anglo-Dutch expedition in 1893 which fixed the border that has divided the island of New Guinea ever since.

There were clear discrepancies in the size of social units in the region which opposed relatively large and complex polities (numbering up to 10,000 or more) employing

expansionistic military policies to small units numbering in the hundreds at most. The Marind – described in detail in Van Baal's (1966) classic ethnography of Dema – were the most successful of the former groups, in demographic and military terms, able to muster parties of scores of war canoes each containing 50 or more warriors. Their policies included the forming of alliances with immediate neighbours to allow them safe passage to raid groups beyond, the assimilation of non-Marind neighbours (such as the Marori and the Kanum) into an expansive system of allied clans aligned with Marind cultural norms, and the full social assimilation of children captured in headhunting raids to Marind ethnicity.

It is not hard to see how the power imbalances in this situation would have driven demographic and linguistic expansion of Marind at the expense of their smaller neighbours. The greater retention of Marind with respect to highly endangered smaller languages like Marori and Kanum in the modern era is simply a continuation of a much older dynamic. Without us yet being able to put details to this scenario, it suggests a situation where rapid expansion of some larger groups at the expense of smaller ones was interrupted by the intervention of European colonial powers – and we may not be exaggerating to say that without the arrival of colonial governments (and missionary endeavours eliminating headhunting and overt warfare) many of the small languages of the Trans-Fly may not have survived in the way they have.

A further, fascinating element in this dynamic comes from the linguogenetic affiliations of the groups involved. All of the large, expansive groups have been classified to be members of the Trans-New Guinea grouping. These include Marind (7,000 speakers), Kiwai (9.700) and Suki (3,500), though both Marind and Kiwai deviate significantly from typical Trans-New Guinea languages typologically (see footnote 4 for an elaboration of this point as it pertains to Marind), likely reflecting prior substrate linguistic influence from autochthonous Southern New Guinea languages. All of the above languages boast speaker populations an order of magnitude higher than languages in the Morehead district, with populations like 710 (Nambu), 250 (Nen) or – at the larger end, 1,600 (Idi). This is not to say that there are not also small Trans-New Guinea languages – Marori (Arka, this vol.) is a clear case, with a current population of under 40 probably reflecting a long period of restricted demography. But all the big languages in the region are Trans-New Guinea[4],

[4] Particularly in the case of Marind, there is evidence for significant typological assimilation to their Southern New Guinea neighbours, so it is useful to say a little more here about the Marind case.

Along with Kuni and other languages around the southern end of Lake Murray, with which it forms a clear subgroup, Marind has been considered by most investigators to be a branch of the Trans-New Guinea family (e.g. Pawley 2005, Ross 2005). Though these sources based the claim primarily on free pronouns, supplemented by a few possible lexical cognates, their argument has recently been strengthened by Suter's (2010) findings of cognacy within the bound pronominal object system as well, on a subset of transitive verb. Suter originally based his reconstructions of this subsystem on languages of the Huon Peninsula, but has more recently extended it upward to a probable pTNG level. He reconstructs 1sgO *na-*, 2sgO *ga-*, 3sgO *wa-* and 3pl *ya-* for proto Huon Peninsula; in Marind the corresponding forms are *na-*, *ha-*, *wa-* and *e-* as illustrated by the verbs *n-esov* 'follow me', *h-esov* 'follow you', *w-esov* 'follow him/her' and *y-esov* 'follow them' (Drabbe 1955:77). As in the Huon Peninsula languages investigated by Suter, as well as in many

suggesting that this area will be a particularly fruitful place to look at the question of why and how speakers of Trans-New Guinea languages have expanded across much of New Guinea, carpeting what was presumably once a much more diverse region with relative linguistic homogeneity.

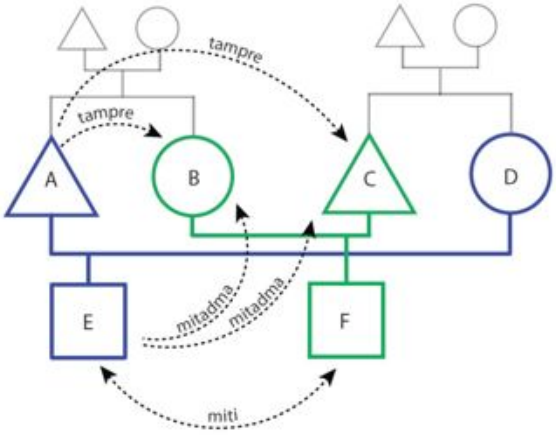


FIGURE 5. The special affinal terms that result in Nen from direct sister-exchange. Following the consummation of a full exchange, brothers stop calling their sisters 'sibling' and instead call them *tampre*, the term for 'sibling-in-law'. Special terms *mitadma* and *miti* also exist for the simultaneously affinal and consanguineal relatives produced by such an exchange – *mitadma* denotes both parents' opposite sex siblings, just in case they were a party to a direct sister exchange, and the term *miti* denotes just those cross-cousins born to such an exchange.

other TNG languages, it is only a subset of transitive verbs that take object prefixes (Drabbe 1955 lists 31).

On the other hand, two important publications (Reesink et al 2009 and De Vries 2004) place Marind outside TNG on the basis of its typological profile. De Vries (2004) suggested a link with the Inanwatan family. And Reesink et al (2009), using the Bayesian tree-building algorithm Structure, single out Marind as one of four languages in their sample (along with Inanwatan again, but also Klon and Abui of the Timor-Alor-Pantar group) which had been considered as TNG in existing classifications but which do not pattern with TNG in a profile of 160 typological characters.

The most likely reconciliation for these conflicting affiliations is that Marind is in fact a Trans-New Guinea language phylogenetically, but has undergone extensive typological reconfiguration as its ancestral speakers moved into Southern New Guinea. This would make it an interesting case of a Trans-New Guinea language assimilating structurally to substrate Papuan languages from other families. In fact, Wurm (1982:95) already suggested something along these lines: he considered Marind and its relatives, while members of the 'Trans-New Guinea phylum', to 'display a number of aberrant features which are probably attributable to a strong substratum, with several of these aberrant features comparable to characteristics of languages of the Trans-Fly Stock'.

FIGURE 6. Youths from the adjoining villages of Bimadbn (Nen-speaking) and Dimsisi (Idi-speaking) transferring a load onto bicycles at an intermediate point between the two villages. They converse in an easy mixture of Nen and Idi. (Photo: N. Evans)

2.1.2 Sister-exchange and multilingualism in the Morehead Region. The Morehead region is famed anthropologically for its practice of direct sister-exchange resulting in virilocal residence (see Williams (1936) and Ayres (1984) for classic anthropological accounts). Figure 5 shows how such direct exchanges impact on aspects of the kinship terminology in Nen. Since exchanged women should come from different clans, and there is a strong chance that different clans will speak different languages, this makes it highly likely that a child's mother will have married into the village from another language background, adopting her husband's language after marriage (though possibly knowing it fairly well before through prior exposure). Since different generations in a lineage typically exchange women with different clans, this regularly brings a large set of languages into the household, and into the experience of the growing child. For example, a Nen-speaking man U may have a Nen-speaking father V who married an Idi-speaking woman W, and in turn marries a Nambu-speaking wife X. U would be expected to have good mastery of Nen (the language of his father's clan), Idi (the language of his mother's clan, whom he would visit regularly) and Nambu (the language of his wife, with whose clan he needs to maintain regular contact). It is evident that, by continually creating multilingual households in a stable and recurring way, direct sister-exchange engenders conditions that favour language contact and mutual influence (see figure 6) – we look at some of the consequences in section 5.

2.2. IMPACT OF MODERN POLITICAL UNITS ON LANGUAGE USE. The impact of modern polities on the Southern New Guinea region has had very different effects on the two sides of the border, so that it is now one of the steepest economic and demographic gradients across a national boundary to be found anywhere in the world (figs. 7, 8).

FIGURES 7 AND 8. Views looking west into Indonesia and east into PNG, from the border point at Sota in Indonesia (Photos: N. Evans)

On the PNG side, the Trans-Fly is a forgotten region – perhaps the poorest and most isolated in the country. Yet, balancing this, people retain full control of their land, according to traditional laws, and their lives depend almost entirely on subsistence activities. Though some languages appear to have become extinct in living memory, or are down to just a few speakers (e.g. Len and Rema within the Yam family), people claiming descent from these speakers have typically shifted to another language of the region rather than to an outside lingua franca.

The language ecology of typical individuals involves a substantial portfolio of languages. A man in the village of Bimadbn, for example, might speak Nen (daily language), Nambu and Idi (neighbouring languages and probably those of his wife or mother-in-law), Motu (for wider communication) and English. Tok Pisin is starting to appear at the fringe of people's repertoire, either through the church or through residence elsewhere (e.g. Port Moresby, Ok Tedi mine). Young men, in particular, have a growing interest in adding some form of basic Indonesian to this repertoire, as they travel by bicycle across the border to

acquire trade goods not available in the Morehead district itself. The overall picture, then, is of solid retention of traditional language as part of a subsistence economy, traditional land rights, and a culture of multilingualism in both local languages and those of wider communication.

FIGURE 9. Rice paddy in area of cleared melaleuca forest, between Merauke and Wasur (Photo: N. Evans)

On the Indonesian side, rapid economic development and environmental change accompanying the influx of transmigrants is proceeding at a rapid pace, and Merauke is a booming local centre. Much land has been cleared for rice cultivation (fig. 9) by transmigrants from Java and other parts of Indonesia; roads have been established and are now lined with tokos (Indonesian-style roadside stores); there are police posts in every village and houses in villages like Wasur or Poo are now mostly built by the government rather than by locals themselves. Speakers of traditional languages of the area are now significantly outnumbered by transmigrants from elsewhere in Indonesia. On the other hand, access to education, electricity, health care and the means of earning money are all far ahead of what is available on the PNG side, so much so that some young Papua New Guineans are undertaking courses, such as in agriculture, on the Indonesian side of the border. In terms of the effect on language, Yei and Kanum are both yielding to Indonesian, at different rates in different villages (e.g. when I visited Poo in 2008 the youngest Yei speaker I could find was in late middle age, whereas in Erambu there were fluent Yei speakers in their late twenties). Marind, however, seems to be holding its ground much better, reflecting the traditional dominance of the Marind-Anim in the region and this is visible in the public symbolic use of written Marind alongside Indonesian in some public signage (e.g. in the Wasur National Park), on the side of aircraft flying to Merauke, etc.

Overall language shift, then, is already reaching a critically advanced state in many languages on the Indonesian side of the border (Yei, Kanum, Marori) as young members of the community shift to Indonesian as the dominant language. On the PNG side, by contrast, the situation is currently one of stable multilingualism with a strong presence of traditional languages in all age groups.

3. Main linguistic groupings in Southern New Guinea. The Southern New Guinea region is home to around 40 languages split between some nine language families – representing, on our current knowledge, five or six maximal clades (i.e. unrelatable units). An indication of the relevant families (though not including all members of each family) is given in figure 10, along with a listing of sources in table 2. The spatial distribution of families suggests a sort of historical pincer movement by which Trans-New Guinea languages came down the Fly River to the north and east, and the Digul to the West, trapping the much more diverse languages of the Trans-Fly region between these rivers and the southern coast. Thus Suki/Gogodala and Tirio to the north, Kiwai to the east, and Marind (and Marori) to the west are all plausible branches of the Trans-New Guinea family.

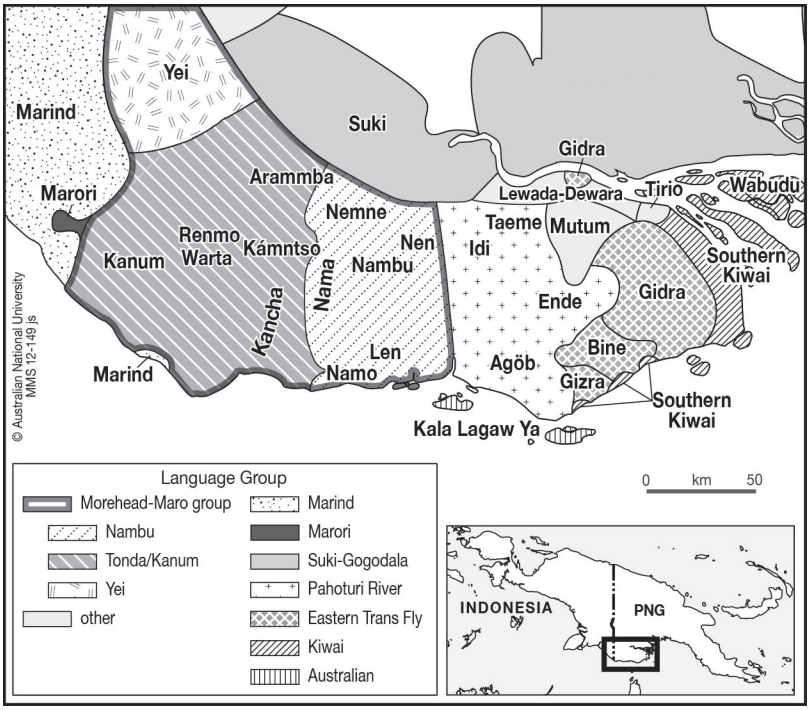


Figure 10. The (focal) Trans-Fly region, showing the main language groups and selected languages from each. Kanum, Yei, Tonda and Nambu are branches of the Yam (Morehead-Upper Maro) family. (Family boundaries are indicative only and need further checking)

Family	Main members	Affiliations and counter-claims	Main sources on affiliation
Marind	Marind, Yaqay, Kuni-Boazi (c. 6 languages in 3 branches)	Claimed member of TNG by numerous authors, though typological classifications place outside TNG, with Inanwatan	De Vries (2004), Pawley (2005), Ross (2005), Reesink et al (2009), Suter (2010)
Marori	Marori (isolate)	Claimed member of TNG	Ross (2005), Pawley & Hammarström (f/c)
Yam (Morehead-Maro)	Around 15 languages across three branches (Tonda, Nambu, Yei) including Nen, Kanum	Claimed subfamily of 'South-Central Family' but better regarded as independent family	Ross (2005)
Pahoturi River	4 closely-related languages or perhaps even one dialect chain: Idi, Taeme, Ende, Agob	Claimed subfamily of 'South-Central Family' but better regarded as independent family	Ross (2005)
Eastern Trans-Fly	4 languages: Bine, Gidra, Gizra, Meriam	Independent family	Ray (1923), Ross (2005)
Tirio	Up to 5 languages: Tirio, Lewada-Dewala, Atulu, Abom, Baramu	Claimed branch of TNG	Ross (2005) Pawley & Hammarström (f/c)
Suki-Gogodala	2 languages (Suki, Gogodala)	Claimed branch of TNG	Voorhoeve (1970), Ross (2005) Pawley & Hammarström (f/c)
Kiwai	Dialect network divisible into about 6 closely-related languages	Claimed branch of TNG	Ross (2005), Pawley (2005), Pawley & Hammarström (f/c)
Western Torres Strait	Dialect chain with a number of dialects (Kala Kawaw Ya, Kala Lagaw Ya, etc.)	Member of Pama-Nyungan family, Australia	Latham (1852), Alpher et al (2008)

Table 2. Main linguistic groupings in Southern New Guinea

To the south, in the western part of the Torres Strait, is the language known in its dialectal variants as Kala Kawaw Ya (on Saibai and other island) and Kala Lagaw Ya (on the more southerly islands), as well as simply 'the Western Torres Strait language'. This is clearly an Australian language (Alpher et al 2008; Evans 2005), though particularly in its phonology it has undergone a significant restructuring away from Australian norms.

Between the Trans-New Guinea languages to the north, west and east, and the Australian languages to the south, lie three language families which on best current evidence appear to be unrelated either to each other or to the languages which adjoin them.

The largest of these, with around 15 languages depending on how the language/dialect boundary is negotiated, has traditionally been called the 'Morehead-Upper Maro family', but in this article I will refer to it by the more compact term 'Yam family'. This term is triply motivated: it recognises the importance of a significant paradigmatic alternation in establishing the relatedness of the family (3sg of 'be' is *yəm*; 3 nsg is *yæm* in Nen and there are cognates across the family – see Evans 2009). The lexical item *yam* or similar words is a widespread word for 'law' or 'culture' in languages of the family (e.g. Nen *yam* 'law, tradition, culture'). And the language-family name pays tribute to the central role of yam-cultivation in the economy of most of the region. This family divides into three branches – Nambu to the east, Tonda in the middle and west (including Kanum), and Yei to the northwest.

Moving east we encounter the second family, Pahoturi River, with four very closely related varieties – Idi, Taeme, Ende and Agob – which may turn out to be a single dialect chain, or else two or more very closely related languages.

Even further east lie the languages of the Eastern Trans-Fly family (also known as the Oriomo River family) – Bine, Gidra and Gizra on the mainland, along the southern coast, up the Oriomo River and abutting the western side of the Fly River, and Meryam Mir on Murray Island in the Torres Strait, inside the Australian political boundary.

To complete our brief survey of the language families of the region, two further languages to the west of Marind bear mention – Yelmek and Maklew. Though Ross (2005) grouped these as a third branch of a putative 'South-Central Family' – along with Morehead-Upper Maro and Pahoturi River – it is not at all clear what this decision is based on and until we know more about these languages it seems safer to regard them as a separate and unrelated family.

The existence of so many languages and families in such a small area, namely of 4-7 maximal clades[5] (i.e. currently unrelatable phylogenetic units), makes southern New Guinea one of the most diverse parts of Melanesia, outstripped only by the Sepik and the central North Coast. As we shall see in the next section, the diversity is not simply phylogenetic – there are major typological differences as well, even in languages spoken by interconnected neighbouring communities.

4. Nen and Idi: divergent neighbours. To give a feeling for how languages of the region work, as well as the balance of sameness and difference across neighbouring language families, I will briefly sketch the functioning of two languages – Nen and Idi – which belong to different non-TNG families in the region, yet are spoken in neighbouring villages and linked by close ties of intermarriage and multilingualism. Nen is the easternmost member of the Yam family, and is spoken in just one village (Bimadbn) by around 250 people, though this village represents a colonial-era aggregation of what were formerly a number of hamlets scattered over a relatively wide area. Idi belongs to the

[5] I.e. Yam, Pahoturi, Eastern Trans-Fly, Yelmek-Maklew, Trans-New Guinea and Australian, with Marori also a possibility if it turns out not to be part of TNG. This gives a high range of 7 maximal clades, and a low range of 4 if one were to follow Ross (2005) in putting Yam, Pahoturi and Yelmek-Maklew into a single grouping.

Pahoturi River family and has around 1,600 speakers in several villages, such as Dimsisi and Sibidiri.

There are close ties between speakers of these languages, reinforced by sister-exchange across the language boundary which produces widespread knowledge of each other's languages and other interesting manifestations including place-names that are said to mix Nen and Idi elements, such as *Sugäl* (said to be comprised of Nen *su* 'belly' plus Idi *gäl* 'canoe') or *Dudumae* (Nen *Dudu* [old garden place name] plus Idi *mae* 'house'). As is the case more widely in the Morehead district, these languages are named after their respective word for 'what' (*nen* in Nen, *idi* in Idi), as if English were called *Whattish*, German *Wasisch*, French *quoiais*, and Russian *štoskij*. A variant version of these names is to use the form for 'what is it', such as *Nen Ym* [what 3sg:be] or *Idi Da* [what 3sg:be], some of the language names reported in Ray (1923) are renditions of names of this type, such as 'Nenium' (Ray 1923:334) for *Nen Ym*. The use of such shibboleth-naming is only one manifestation of a sophisticated metalinguistic awareness of structural, phonological and lexical differences found quite widely over the region.

Despite this, the languages differ significantly on many dimensions indeed, so that if Nen's relationship to its westerly neighbour Nambu is like Spanish to Portuguese or German to Dutch, its relationship to its easterly neighbour Idi is like Spanish to Basque or German to Hungarian. I will illustrate this first with a brief sketch of how each language looks on its own, then compare a number of relevant typological features more systematically.

4.1. Nen (Ethnologue Code NQN).[6] Nen's phonological inventory is given in tables 3 and 4. It has relatively few places of articulation, no velar nasal, a voicing contrast, and eight vowels (including a couple of short vowels, plus two marginal nasal vowels). The only somewhat unusual phonemes are the labial-velars, which are coarticulated at labial and velar places of articulation, though phonemes of this type are of course found in many other parts of Melanesia (e.g. Huon Peninsula, Onin Peninsula, Vanuatu). As in a number of other Papuan languages such as Kalam (Blevins & Pawley 2010, Donohue 2009) many syllables lack specified vowel nuclei; these are filled in with brief epenthetic schwas which are not shown in the practical orthography.

[6] Data presented here were gathered by the author over 5 fieldtrips, totalling 15 weeks, between 2008 and 2012.

	Bilabial		Alveolar /dental		Palatal		Velar		Labial-velar		Glottal
Voiceless stop	p	<p>	t̪	<t>			k	<k>	k͡pʷ	<q>	
Voiced stop	b	<b>	d	<d>			g	<g>	g͡bʷ	<g̃>	
Prenasalised stop	mb	<mb>	nd	<nd>	ndʒ	<nz>	ŋg	<ng>	ŋ͡mg͡bʷ	<ng̃>	
Nasal	m	<m>	n	<n>	ɲ	<ñ>					
Voiced fricative					z ~dʒ	<z>					
Voiceless fricative			s	<s>							h <h>
Lateral			l	<l>							
Trill			r	<r>							
Semi-vwl					j	<y>			w	<w>	

TABLE 3. The Nen Phoneme inventory: consonants

	Front		Back	
	Non-short	Short	(Short)[7]	Non-short
High	i (i)	ɪ (é)		u (u)
Mid	e (e)		ɐ (á)	o (o)
Low	æ ~ ɛ (ä)			a (a)

+ marginal ẽ in ẽ 'yes' and gẽhẽ 'over there'

TABLE 4. The Nen Phoneme inventory: vowels

In terms of its grammatical typology, Nen has the following features:

(a) preference for verb-final

(b) no verb-chaining but widespread use of true subordinate constructions using a nominalised verb usually inflected for case, as in (1).

[7] This vowel can almost be eliminated as a phoneme, except in a couple of words, *má* and *mái* 'still' where the presence of *á* cannot be motivated by epenthesis.

(1) *Ynd* *yergb-at* *one-s-t* *w-ng-m.*
1ABS river-AL fish.with.net-NLZR-AL 1sgU[8]:α-away-be
'I'm going to the river to net fish.'

(c) use of suffixes on the final NP element to mark an absolutive-ergative case system, plus another dozen or so case distinctions. These suffixes (and also free pronouns) also encode a singular-non-singular distinction in all but the absolutive case (2). Note that ND stands for 'non-dual' (more on this below), and by not glossing the number of *ynd*, i.e. writing it as '1ABS', I indicate that it is unspecified for number.

(2) *togetoge-yäbem* *ynd* *w-aka-ta-t* /
children-PL.ERG 1ABS 1sgU:α-see-ND-3nsgA

yn-aka-ta-t
1nsgU:α-see-ND-3nsgA

'The children see me / us (3 or more).'

(d) complex verb morphology involving both prefixes and suffixes (2), and including double agreement (actor suffixes and undergoer prefixes, though sometimes a particular combination of actor and undergoer will be shown at just the suffixal or prefixal site), direction (towards, away, neutral), and diathesis (a range of valency-changing prefixes to the root). A complex TAM system combines information from the verbal suffixes (9 distinctions, 3 each for perfective, imperfective and neutral aspect), the verbal prefixes (3 distinctions coded by different series of undergoer prefixes) and preverbal particles.

(e) Monovalent verbs split in their agreement patterns, though not their case, according to whether the predicate is static or dynamic. The subjects of static verbs use undergoer prefixes (3a) and the subjects of dynamic verbs use actor suffixes and a person-invariant 'middle' prefix (3b).

(3a) *Ynd* *w-aki-ngr*
1ABS 1sgU:α-be.standing-STAT
'I am standing.'

(3b) *Ynd* *n-owab-ta-n*
1ABS M:α-talk-ND:IPFV-1sgA
'I am talking.'

The undergoer prefixes have three series, whose semantics is too complex to capture with a gloss, and for which I use the Greek letters α, β, γ. If we just look at the imperfective series, α, β, and γ work backwards from today into the future: the α-form *nowabtan* is 'imperfective non-past' (roughly) and refers to me talking any time from this morning's dawn onwards (with finer specification by preverbal particles), the β-form *k-owab-ta-n* [M:β-talk-ND:IPFV-1sgA] is 'imperfective yesterday past' and refers to me talking

[8] A = Actor (subject of transitive or of dynamic intransitive), U = undergoer (object of transitive, subject of stative).

yesterday or a few days ago, while the γ-form *g-owab-taw-n* [M:γ-talk-ND:REM.IPFV-1sgA] is 'imperfective remote past' and refers to me talking at any time before that (last month, last year etc.).

If that was all there was to the three series, glossing them would be easy. But if we look at other functions of each series the picture becomes muddier: in addition to its imperfective non-past use, the α-series is used for perfectives in the past (i.e. the direction of time-reckoning flips over in the perfective), for future imperatives (do it later!), and two of the 'neutral aspect' categories (which include a couple more remote pasts). The β-series, in addition to its yesterday past use in the imperfective, is used for present imperatives (do it now!), and with a few verbs for perfectives denoting unexpected occurrences. The γ-series, used for the remote past in the imperfective, is used in the perfective for futures (another time flip), as well as for mediated imperatives transmitted via a messenger (X should do it! (convey my command to X)) and for the irrealis.

Given the semantic disparities between these uses, the best treatment is to regard the choice of prefixal series plus the TAM suffix as forming a single circumfixal sign (see remarks later on circumfixal paradigms) and once we adopt that treatment the glossing difficulties vanish since the prefix series are not required to have any meaning of their own. For further remarks on this problem see Evans (forthcoming b).

(f) the existence of a large set of positional verbs (around 30), which in addition to meanings like 'be standing' in (3a) often have very specific semantics (e.g. 'be in a tree fork', 'be immersed'), and which form the lion's share of the stative predicates. From these, transitives ('cause to be in position X') and middles ('become in position X') are then derived. All positional verbs are prefixing verbs, in the sense of using only prefixes to signal person-agreement information.

(g) an unusual 'constructive'[9] number system within the verb which obtains three values[10] by crossing the singular vs non-singular contrast of the agreement morphology with a dual vs non-dual contrast on the root (Fig. 11a) or the verb thematic (Fig. 11b).

[9] The original term used for this type of system was 'constructed' (Corbett 2000:169) but in more recent publications (e.g. Arka 2011) the term 'constructive' has been used and I follow that variant here.

[10] There is also an incipient but much more irregular system of distinguishing small vs large, or partial vs exhaustive plurals, which I don't discuss here, but which underlies my reluctance to use 'plural' here as if it were an unproblematic term.

	U Prefix	Root patterning	Inflected form	
sg	*w-*	*m*	*w-m*	'I am'
pl	*yn-*	*m*	*yn-m*	'we (more than two) are'
du	*yn-*	*ren*	*yn-ren*	'we two are'

FIGURE 11A. Unification of affixal singular vs non-singular agreement values with dual vs non-dual suppletive root of 'be' to give a three-valued basic number system

	U Prefix	Thematic patterning	Inflected form	
sg	*-n*	*nowabta*	*nowabtan*	'I talk'
pl	*-m*	*nowabta*	*nowabtam*	'we (more than two) talk'
du	*-m*	*nowab*	*nowabm*	'we two talk'

FIGURE 11B. Unification of affixal singular vs non-singular agreement values with dual vs non-dual thematic forms of √*owab* 'talk' to give a three-valued basic number system; *n-* is a person/number invariant 'middle prefix'

Though constructive number systems are not all that unusual cross-linguistically (see Corbett 2000:169-70; Arka 2011, this volume) the use of a pervasive dual vs non-dual opposition is, as far as I know, unique to Nen and its close relatives.

(h) a general tendency to exploit *distributed, paradigmatic*, and *constructive/ unificational* architectures to give complete grammatical feature specifications.

It is *distributed* because there is a strong tendency to underspecify information at one site (e.g. giving person but not number in the absolutive pronoun forms) which is then filled in by unification with information at another site (e.g. the verb contributes number information, while the pronoun contributes person information). Complete feature value sets are not present until material from both affix positions, and from free pronouns has been unified (table 5). As can be seen, the absolutive pronouns only show person, not number – *ynd* '1st person abs. (any number)', *bm* '2nd person abs. (any number)',[11] *bä*

[11] A peculiarity of Nen is that the 1sg and 2sg forms, *ynd* and *bm* respectively, neutralise the absolutive vs ergative case distinction found everywhere else in the system. This appears to result from a recent sound change by which the original ergative singular pronominal suffix *-o* was lost from these pronouns as part of a general loss of word-final *o* – cf. Nama which contrasts absolutive *yənd* and *fəm* to ergative *yəndo* and *fəmo,* and Nambu which contrasts absolutive *yənd* and *bəm* with ergative *yənd* and *bəmo*. (Note in passing that the loss of final *-o* is one of the main sources of the coarticulated labial-velar phonemes in Nen – cf. Nambu *məngo* 'house', Nen *mnḡ*; Nama *frango-* 'leave', Nen *branḡ-* 'leave', Nambu *ingo*, Nama *injo-* 'catch sight of, see', Nen

'3rd person abs. (any number)'. Conversely, affixes reliably show number but not person: syncretisms merge the 2nd and 3rd persons in the A and U affix positions: *ya-~yä-* is '2|3nsgU:α' and *-e* is '2|3sgA'. Once free pronouns and inflected verbs are unified all ambiguities are eliminated:

	Free pronoun	talk (2|3sgA) *nowabte*	talk (2|3du) *nowabt*	talk (2|3pl) *nowabtat*
2	*bm*	*bm nowabte* 'you (sg) talk'	*bm nowabt* 'you two talk'	*bm nowabtat* 'you (3(+)) talk'
3	*bä*	*bä nowabte* '(s)he talks'	*bä nowabt* 'they two talk'	*bä nowabtat* 'they (3(+)) talk'

Table 5. Unification of underspecified pronoun and agreement information to give precise person/number specification

It is *paradigmatic* (and sometimes even *circumparadigmatic*[12]) because the information from prefix and suffix often needs to be treated as part of a single paradigm, with forms having very different values according to their place in the paradigm. Thus with 'neutral aspect' TAM suffixes, the suffixal pair *-nd* vs *-t* contrasts 2/3pl vs 2/3du, but their values are swapped (i.e. 2/3du vs 2/3pl) with perfective aspect TAM suffixes. Likewise the γ-series of undergoer prefixes indicates remote past when combined with imperfective verb suffixes, but future when combined with perfective ones.

And – intimately linked to the preceding characteristics – it is *constructive/ unificational* because the full range of categories once combinations are taken into account is much greater than that found at any contrast site. Note that such unification needs to take place both within the word (e.g. between the prefixing and suffixing sites of the verbs) and between the verb and free pronouns (e.g. in working out the full person/number specifications for undergoers).

Note that these characteristics create difficulties for interlinear glossing (as they do in Idi) and I adopt the following two non-standard conventions in the examples that follow. First, I use the pipe (|) to join disjunct feature values (which are then disambiguated through feature unification) such as 2|3sg for 'second or third person singular' in (4.1). Second, as already mentioned above, I use Greek letters (α, β, γ) for contrasting prefix series without clearly specifiable semantics of their own, where this is only 'cashed in' after unifying this information with other parts of the paradigm (such as the suffixes).

A further salient feature of Nen, particularly important for historical and comparative

inḡ- 'see, catch sight of'. However this only occurs with final velars (the prenasalised palatal affricate *nj* in Nama results in this case from palatalisation after the preceding *i*). After other segments, such as /nd/ or /m/, final /o/ simply disappeared without trace.

[12] By which I mean that prefixes and suffixes need to be combined into a single paradigm that is only partially factorisable into separate prefixal and suffixal paradigms.

purposes, is the almost total disconnect[13] between the form of free pronouns and that of agreement morphology for verbs: table 6 compares the absolutive and possessive free pronouns with the three series of undergoer prefixes and the actor suffixes (basic imperfective and past perfective sets). In fact, when one looks right across the Yam family there is good agreement on the form of the undergoer prefixes (see Evans 2009), less so for the actor suffixes (where contrasts are attenuated or lost the further west one goes) and little agreement on the free pronominals.

	Abs.	Poss.	U-prefix (α)	U-prefix (β)	U-prefix (γ)	A-suffix (imperf.)	A-suffix (past perf.)
1sg	*ynd*	*tande*	*w-*	*q-*	*g̃-*	*-n*	*-n*
1nsg	*ynd*	*tbende*	*yn-*	*tn-*	*dn-*	*-m*	*-m*
2sg	*bm*	*bende*	*n-*	*k-*	*g-*	*-e*	*-ø*
2nsg	*bm*	*bbende*	*ya-~yä-*	*ta-~tä-*	*da- ~dä-*	*-t*	*-t/-nd*[14]
3sg	*bä*	*yande*	*y-*	*t- ~*	*d-*	*-e*	*-a*
3nsg	*bä*	*ybende*	*ya-~yä-*	*ta- ~ tä-*	*da- ~ dä-*	*-t*	*-t/-nd*

Table 6. Free pronouns and corresponding verbal agreement forms in Nen

We will mention a few further typological features below (see also Evans forthcoming a,b), but this is now a good point to give a global overview of the language by tackling the following mini-text, which can be heard on http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4562/NenBlacksnakeExcerpt.wav. It is an account of the dangers of being bitten by a Papuan Black Snake, recorded by the present author from the late Aramang Wlila (then aged in his early 60s) in September 2008 .

According to the story, when a Papuan Black Snake bites you, blood starts pouring out of your eyes, and people check how bad you are by asking:

(4.1)

snamb	*bnz*	*aba*	*ya-wakae-w-ng*
how_many	fire	Imm.Pst	2|3nsgU:α -see-IMPF.DU-2|3sgA>DU:IMPF
snamb	*är*	*aba*	*ya-wakae-w-ng*
how_many	person	Imm.Pst	2|3nsgU:α -see-IMPF.DU-2|3sgA>DU:IMPF

‘ “How many fires did you see? How many people did you see?’
[perhaps better translated as ‘Did you see so many fires - two?’ Did you see so many people - two?]

[13] Of the forms given in table 4, only the 1nsg undergoer prefix shows any plausible formal connection to the free pronouns.

[14] Just in a couple of the perfective series the 2nd and 3rd person actor suffixes distinguish dual (here *-nd*) from plural (here *-t*).

(4.2)
a *snamb* *kesär* *ya-waka-t-e*
and how_many sun 2|3nsgU:α-see-IMPF:ND-2|3sg:IMPF
‘”and how many suns can you see?”’

(4.3)
dene *geä* *g-owab-ta-ng-a* *deneya[més]*
thus COND M:γ-talk-IMPF-ND:FutPf-3sgA like_this
‘If he says like this:’

(4.4)
“sombes-ngama, *sombes-ngama* *y-ng-aka-ta-n”*
two-ABL two-ABL 3sgU:α-AWAY-see-ND:IMPF-1nsgA
‘“I see two of each.”’

(4.5)
aa *gn-anma-ng-a…*
um 2sgU:γ-(BEN)call-ND:FutPf-3sgFutPfA
‘Um, the man will call out to you…’

(4.6)
är-t *da-w-anma-nga* *de<neyamés>*
man-PL.OBL 2|3nsgU:γ-BEN-call-ND:FutPf-3sgA like.this
‘He will call out to the people like this’

(4.7)
yna *är-äm* *geym* *ti bä* *da-w-anma-ng-a*
this man-ERG FOC ? 3ABS 2|3nsgU:γ-BEN-call-ND:FutPf-3sgA
‘The person will call the people.’

(4.8)
“tä-n-m, *wgd* *zer-s*
2|3nsgU:β-hither-be:ND proper bite-INF

aba *y-ze-n-e,*
IMM.Pst 3sgU:α-bite-IMPF-ND-2|3sgA
‘“Come, it has bitten him good and proper’

(4.9)
“a *kr* *kaka* *y-m”*
and death near 3sgU:α-be
‘”and is about to die.”’

These few lines of text illustrate many of the salient features of Nen morphosyntax,

reinforcing the simplified examples given above with real textual material:

(a) ***the existence of two alignment systems*** – an absolutive/ergative system for case and a split-S system for agreement (on the basis of a stative vs dynamic contrast rather than an agency contrast; fig. 12). The case system is ergative/absolutive, opposing an ergative form for the agent of transitives (*är-äm* 'man-ERG' in (4.7)) to an (unmarked) absolutive form for the patient of transitives (*är* in (4.5)) and the sole argument of intransitives, whether dynamic or stative; there is also a dative for recipients/beneficiaries. The ergative is fully specialised for this function and does not mark any oblique function (e.g. instrument or source).

The verbal indexing system employs an 'undergoer' prefix for patients of transitives (e.g. *y-* in *yzene* 'it bit *him*') in (4.8)), and the sole argument of statives (e.g. *y-* in *ym* 'h*e* is' in (4.9)), and an 'actor' prefix for agents of transitive (e.g. *-e* '2|3sg' in *yawakate* in (4.2) and in *yzene* in (4.8)) and of dynamic monovalent verbs (e.g. *-a* '3sgFPfA in (4.3)). The 'undergoer' prefix (obviously the term is not perfect) is also used for the recipient or beneficiary of ditransitive verbs.

Ditransitive	Case marking Verbal indexing	Dative U-: IO	Ergative -A: A
Transitive	Case marking Verbal indexing	Absolutive U-: O	Ergative -A: A
Intransitive	Case marking Verbal indexing	Stative: Absolutive U-: S_{stat}	Dynamic: Absolutive -A: S_{dyn}

FIGURE 12. Role splits and mergers: case-marking and verbal indexing. U- and -A represent the undergoer prefix and actor suffix respectively; syntactic roles are represented by A, $S_{stat,}$ $S_{dyn,}$ O and IO. In addition to the roles shown here, in ditransitives there is an O, marked with the absolutive case, which is not indexed on the verb.

(b) ***a split in morphological organisation between prefixing verbs*** (monovalent, stative, e.g. *ym* in (4.9))[15] and ***ambifixing verbs*** (I use this term for verbs which take both prefixes and suffixes[16]). The latter may be divalent like *dawanmanga* 'he will call out to them' in

[15] The stative characterisation leaks slightly. It holds of the base form 'be', plus around thirty 'posals' giving position (be in a tree fork) or posture (be sitting). But three verbs defy the characterisation of this category as stative – 'come' and 'go', which are the 'towards' and 'away' forms of 'be' and hence may simply be inheriting the morphology of their source verb, but also *utan* 'walk'.

[16] A note on this terminological choice: the reason I don't use 'circumfixing' here is that ambifixing allows for the possibility that choices in the prefix and suffix are independent, i.e. represent orthogonal categories, whereas circumfixing implies that material from prefix and suffix gets integrated into a single semantic value. Of course, the morphological fact of a verb being

(4.7), or monovalent and dynamic like *gowabtanga* 'he will say' in (4.3), as well as trivalent like 'give' (no examples in this text). Nen has an unusually large number of middle verbs (Evans forthcoming a), assigning virtually all dynamic one-place predicates to this class (e.g. talk, work, ascend), as well as more typical middles like (derived) reflexives and reciprocals.

(c) ***Four sites for encoding TAM:***
Coding of tense/aspect/mood is split across

Time adverbs, all of which are bidirectional, e.g. *kae* 'yesterday, tomorrow; one day from today'

Preverbal particles, which are unidirectional, e.g. *aba* 'just now, very recently' (4.1), *geä* 'if, when' (4.3).

Undergoer-prefix series, which have three sets encoding TAM. The semantics of these is not straightforward, and cannot be specified until they combine with TAM suffixes and preverbal particles. In our sample text, the 2|3nsgU prefix is exemplified with all three values: α form *ya-* in (4.1) and (4.9), β form *tä-* (an allomorph of *ta-*) in 4.8, and γ form *da-* in 4.6 and 4.7. As these forms illustrate, the α-series are glides or nasals, the β-series are the corresponding voiceless stops, and the γ-series are the voiced correspondents of the β-forms.

In these examples the α-series is associated with present and recent past, the β-series with the imperative, and the γ-series with the future. But things are not always so straightforward: with imperfective inflections, the γ-series signals remote past rather than future, and the β-series signals the past of yesterday or a couple of days ago.

Suffix series, expressing TAM + number + actor person/number (it is usually possible to split these further into a 'thematic' followed by a 'desinence' (see Evans forthcoming b). For ambifixing verbs, these form nine sets divisible into three aspect series (perfective, imperfective and neutral) each containing three values. (For prefixing verbs the possibilities are much more limited). The current text exemplifies some of these: the (basic) imperfective (4.1, 4.2, 4.4, 4.8), which covers all imperfective indicatives except the remote, and the future (4.3, 4.5, 4.6, 4.7).

(d) ***employment of infinitive forms***. Nen does not have any form of verb-chaining or switch reference. Rather, it makes frequent use of infinitive forms for a whole range of functions, such as complement clauses of various types, as well as a sort of emphatic construction, exemplified in (4.8), in which the infinitive form of the verb (*zers* 'to bite') is combined with an inflected form (*yzene* 'he bit it') to mean something like 'he really bit him' (lit. 'he

ambifixing does not preclude that some or all of the prefix + suffix combinations function as circumfixes, but it also leaves open the possibility that they are independent.

bit him to bite'). Infinitives are formed by adding *-s* to the verb root.[17]

An important use of the infinitive in Nen, not illustrated in the text fragment, is as the complement of phasals such as 'begin to V' or 'finish V-ing', expressed by combining the infinitive (suffixed with an appropriate case) with a phasal auxiliary. The auxiliary carries all inflectional material that the lexical verb would have borne had it been finite – middle prefix plus actor suffix with 'return (itr.)' in (5a), undergoer prefix plus actor suffix with 'stand up (tr.)' in (5b), and undergoer prefix, benefactive prefix and actor suffix with 'give' in (5c).

(5a) *Ynd* *angḡ-s-t* *n-opap-nd-m.*
1ABS return-NLZR-AL M:α-begin-ND:PFV:PST-1nsgA
'We are about to return.'

(5b) *Ynd* *bä* *w-nḡi-s-t* *y-a-pap-nd-n.*
1sgA 3ABS TR-stand.up-NLZR-AL 3sgU:α-CAU-begin-ND:PFV:PST-1sgA
'I am beginning to / about to/ trying to stand him up.'

(5c) *Ahã* *Gbae* *ynd* *begta* *tande* *yép*
here.you.are [name] 1sgA 2sg:DAT1 1sgPOSS bag(ABS)

räm-s-t *n-ng-a-wa-pap-nd-n.*
give-NLZR-AL 2sgU-away-BEN-CAU-begin-ND:PFV-1sgA
'Here, Gbae, I'm about to give you my bag.'

This concludes our short sketch of Nen. For comparison, we now travel about 25 km east, from Bimadbn to the neighbouring village of Dimsisi. Since there is negligible published material on languages of the Pahoturi River family, this will also give a chance to give the public at least a small glimpse of how languages in that family work.

To give an initial idea of the degree of difference between the languages, we can compare their paradigms of free pronouns, which show negligible[18] resemblances; for comparison the free pronouns are also given (in blue and red respectively) for Nama and Nambu[19], 40 km and 20 km to the west of Nen (table 7).

[17] The 3sgA form in (4.8), *yzene*, replaces the *r* with *n*. This is a regular process with verbs whose stems end in *-r*, before non-dual. But the *r* of the imperative can be seen clearly in imperfective non-dual forms, e.g. *yzert* 'the two of them bit him', and in perfective imperatives, e.g. *tzer* 'bite him! (newly initiating the action)'.

[18] One could seize on the presence of *b-* in 2nd and 3rd person forms as a vestige of possible relatedness. In other cases apparent similarities (e.g. Nen 2nsg abs. *bm*; Idi 2nsg acc. *bibim*) are coincidental in the sense that the *m* in Nen is part of the root whereas the *-m* in Idi marks accusative.

[19] I thank Jeff Siegel for supplying me with these forms.

	Nen (with Nama in small blue and Nambu in small red)			Idi		
	Abs	Erg	Poss	Nom	Acc	Poss
1sg	*ynd* yənd yənd	*ynd* yəndo yəndo	*tande* tane tande	*ŋən*	*bom*	*bo*
1nsg	*ynd* yənd yənd	*yndbem* yəndfem yəndvem	*tbende* təfene təvende	*bi*	*ba*	*ba*
12nsg	*ynd* yənd yənd	*yndbem* yəndfem yəndvem	*tbende* təfene təvende	*ybi*	*yba*	*yba*
2sg	*bm* fom bəm	*bm* fəmo bəmo	*bende* fene bende	*be*	*babom*	*béna*
2nsg	*bm* fom bəm	*bmbem* fəmofem bəmovem	*bbende* fəfene bəvende	*be*	*bibim*	*béna*
3sg	*bä* fæ bæ	*ymam* yəmo yəmo	*yande* yæne yænde	*bo*	*obom*	*obo*
3nsg	*bä* fæ bæ	*ymabem* yəmofem yəmovem	*ybende* yəfene yəvende	*bo*	*ubim*	*oba*

TABLE 7. Free pronouns in Nen, Nama (small blue font), Nambu (small red font) and Idi.

4.2. IDI (ETHNOLOGUE CODE IDI). Idi is spoken in the three villages of Dimsisi, Sibidiri and Dimiri by a population of around 1,600 people. Together with three other named varieties – Ende, Agöb and Taeme – it forms the Pahoturi River Family. Compared to the Yam family, where there are substantial differences across different branches, the current (extremely limited) data suggests that all the Pahoturi River varieties are extremely close, possibly even sister dialects.

Comparing Idi and its neighbour Nen, one is immediately struck by a number of salient differences in both consonant and vowel inventories. Idi has a retroflex series of stops (/ʈ/ and /ɖ/), which are generally realised with significant affrication, at least two laterals (certainly /l/ and /ʎ/, possibly also /ɭ/), and a velar nasal (lacking in Nen).

It has a smaller vowel inventory than Nen (though this part of Idi phonology is still not well understood) – cf. the contrasting phonemes /e/ and /ä/ in Nen which fall within the allophonic range of a single /ɛ/ phoneme in Idi. Current analysis suggests a six-vowel system – i, ɛ, a, ə, o, u.

The status of labial-velars is problematic. Some Idi-Nen bilinguals use labial-velar articulations in certain Idi words, which may turn out to be Nen loans. But if we limit

ourselves to phonemes used by all speakers then there do not appear to be labial-velars, though there are velars with a rather lax rounded release. The consonant inventory is shown in table 8.

Manner / Place	Bilabial	Alveolar	Retro-flex	Lamino-Palatal	Velar	Labio-velar	Co-articulated Labial-velar
Voiced stop	b <b>	d <d>	ɖ <ḍ>		g <g>	gw <gw>	g͡bʷ < g̱ >
Voiceless stop	p <p>	t <ṭ>	ʈ <ṭ>		k <k>	kw <kw>	k͡pʷ <q>
Affricate/ fricative				dʒ~z <z>			
				s <s>			
Nasal	m <m>	n <n>		ɲ <ñ>	ŋ <ŋ>		
Lateral		l <l>	ɭ <ḷ>	ʎ <ʎ>			
Rhotic		r <r>					
Continuant				j <y>		w <w>	

Table 8. Idi Consonant inventory (with proposed orthographic symbols in angle brackets)

In terms of grammar, there are some gross typological similarities with Nen. Both are verb-final, both inflect transitive verbs with both prefixes and suffixes, both have TAM-senstive forms of the prefix series, and both have infinitive plus auxiliary constructions in which the auxiliary indexes all arguments of the infinitive verb. However, there are no verbs which use prefixes alone to signal subject agreement, in the way that is found with 'prefixing verbs' like the copula or the positional verbs in Nen: all intransitive verbs in Idi, including the intransitive auxiliary and the copula, make exclusive use of suffixation for agreement purposes.

The complex architectural relationship between free pronouns and agreement morphology also shows typological similarities to Nen: there is a severe disconnect between both the forms and the categories of free pronouns and verbal agreement, with widespread but non-correlated syncretisms in each system which require the unification of information from both free pronouns and inflected verbs before the precise feature values can be known, as we shall see from examples to be given below.

Table 9 gives the free pronoun forms plus intransitive auxiliary forms for two tenses (present and far past); note again the lack of any formal connection between the free pronoun forms and the inflected auxiliaries. Note also the lack of any formal similarity between the person/number forms of the auxiliary in Idi and those given for Nen verbs in table 5.

	Nominative	Accusative	Possessive	Intrans. auxiliary: Pres	Intrans. auxiliary: Past
1sg	*ŋən*	*bom*	*bo*	*wala*	*wagən*
1du	*bi*	*ba*	*ba*	*waŋama/ walala*	*gwaga*
12du	*ybi*	*yba*	*yba*	*waŋama*	*gwagma*
1pl	*bi*	*ba*	*ba*	*waŋama*	*gwaga*
12pl	*ybi*	*yba*	*yba*	*waŋama*	*gwagma*
2sg	*be*	*babom*	*béna*	*walale*	*gwege*
2du	*be*	*bibim*	*béna*	*walala*	*gwaga*
2pl	*be*	*bibim*	*béna*	*waŋama*	*gwagma*
3sg	*bo*	*obom*	*obo*	*wala*	*gwaggen*
3du	*bo*	*ubim*	*oba*	*walalo*	*gwago*
3pl	*bo*	*ubim*	*oba*	*waŋamo*	*gwagmo*

Table 9. Comparison of Idi free pronouns and inflected forms of the intransitive auxiliary (present and past forms)

In Idi, the infinitive plus auxiliary construction is much more widespread than in Nen. In Nen it is used for phasal constructions 'begin to V; finish Ving', and this is also the case in Idi (examples to be given later, in (12)). But in Idi its use is extended further – it is the normal construction in the present tense, for example (6a,b) – and it is only in a subset of TAM values (e.g. past perfective settings) that the main verb is directly inflected (cf. 6c,d). Note that valency alternations shown by auxiliary choice in the periphrastic construction are shown by the choice of prefix on the finite verb.

(6a) *pelaʈ-a* *paldab* *wala*
plate-COR break:INF INTR.AUX:1|3sg:PR
'The plate is breaking.'

(6b) *ʈiʈim-e* *pelaʈ-a* *paldab* *yera*
girl-COR plate-DIR break:INF TR.AUX:1|3sg>3sg:PR
'The girl is breaking the plate.'

(6c) *ʈiʈim-e* *pelaʈ-a* *ya-paldab-en*
girl-COR plate-COR PST:sgO-break-1|3sg>npl
'The girl broke the plate.'

(6d) *pelaʈ-a* *wa-paldab-en*
plate-COR PST:RR-break-1|3sg
'The plate broke.'

Examples (7) and (8) compare intransitive clauses, using the intransitive auxiliary *wala*, with transitive clauses using the transitive auxiliary *yera*. Sometimes (as in the case of (7c) vs (8c), or (7d) vs (8d)) this effects the difference between intransitive/causative or reflexive/transitive doublets. These examples also illustrate another interesting feature of Idi. A 'core case' marks all core nominal arguments – subjects (transitive or intransitives) and objects – even though nouns used in isolation (e.g. in nomination) appear without it, e.g. *ged* 'child' or *ʈiʈim* 'girl' (in an elicitation context). It is only personal pronouns which distinguish core arguments, via a nominative vs accusative case distinction (7b) – contrast *ŋən* '1sgNOM' vs *bom* '1sgACC'; *bo* '3sgNOM' vs *obom* '3sgACC'.

(7a) *ged-e* *méλ* *wala*
child-COR scream INTR.AUX:1|3sgS:PR
'The child is screaming.'

(7b) *ʈiʈim-e* *wala-ŋgawa* *bisi* *wala*
girl-COR forest-ALL go INTR.AUX:1|3sgS:PR
'The girl is going to the forest.'

(7c) *lu-e* *zaŋg* *wala*
tree-DIR burn:INF INTR.AUX:1|3sgS:PR
'The tree is burning.'

(7d) *ʈiʈim-e* *oboobo tetu* *wala*
girl-DIR 3sgRR wash INTR.AUX:1|3sgS:PR
'The girl is washing herself.'

(8a) *ged-e* *lu-e* *kakλ* *yera*
child-DIR tree-DIR climb:INF TR.AUX:1|3sg>3sg:PR
'The child is climbing the tree.'

(8b) *ŋən* *obom* *ɖəndɖəg* *yera*
1sgNOM 3sgACC bite:INF TR.AUX:1|3sg>3sg:PR
'I am biting him/her.

(8c) *lu-e* *ged-e* *zaŋg* *yera*
tree-DIR child-DIR burn:INF TR.AUX:1|3sg>3sg:PR
'The child is burning the tree.'

(8d) *ʈiʈim-e* *obo* *ged-e* *tetu* *yera*
girl-DIR 3sgPOSS child-DIR wash TR.AUX:1|3sg>3sg:PR
'The girl is washing her child.'

Other case morphology includes locative *-me* (*kələm-me* 'in the swamp'), allative *-awa* (*kələm-awa* 'to the swamp'), ablative *-(a)ɻ* (*walaŋg-aɻ* 'from the forest'), dative *-ble* (*gəd-ble* 'to the boy') instrumental *-enda* (*sabor-enda* 'with a spade' (*sabor* < Eng. 'shovel').

As with the verbal morphology and the free pronouns, there are no formal resemblances between the forms of any of the case suffixes and those in Nen or other languages of the Yam family (the respective forms in Nen would be locative *-an*, allative *-ta*, ablative/instrumental *-ngama*, and dative *-eita* or *-eipap*).

As in Nen, Idi organises its agreement morphology in a way that requires unification of featural information from free pronouns and inflected verb before all feature combinations are resolved. For example the present tense form of the intransitive auxiliary includes such syncretisms such as *wala* [1|3sgSubj], which is resolved once combined with the free pronouns: *ŋən bisi wala* 'I go', *bi bisi wala* 'he/she goes' (cf. 7b). Likewise the transitive auxiliary *yera* 'to do to something' includes many forms with a large syncretic range such as *ñerala* '1nsg|12nsg|2nsg>du; 2nsg>1pl; 1nsg>2pl'.

Syncretisms in the Idi paradigm extend much further than in Nen, collapsing large sets of combinations in underspecified blocks. Consider the immediate past, as it applies to finite transitive verbs. Prefixes simply distinguish singular object (*na-*) vs non-singular subject (*ña-*), while suffixes distinguish a range of categories defined by person and number. Examples in (9), from the near past (same day) paradigm illustrate how the combinations get disambiguated once free pronouns are added. (The time adverb *sisiri ektende* 'earlier today' could optionally be added to any of these.) As these examples show, the inflected verb forms *na-nḍəg-la* (singular object) and *ña-nḍəg-la* (non-singular object) are compatible with a very wide range of subject/object combinations for person/number – in these combinations, the second person needs to be non-plural (i.e. singular or dual) whereas first persons need to be non-singular (i.e. dual or plural). (9a-c) illustrates some of these possibilities with a singular object, signalled by the prefix *na-*, and (10a-10d) with a non-singular object, signalled by *ña-*. (To avoid over-complex glossing here I use one value set for 2nd person and another for non-2nd, allowing for prior disambiguation by the free pronoun.)

(9a)	*bi*	*komblebe*	*bom*	*na-nḍəg-la.*
	2NOM	two	1sgACC	TOD.PST.sgO-see-2nplA>sgO
	'You two saw me (earlier today).'			
(9b)	*be*	*komblebe*	*obom*	*na-nḍəg-la.*
	2nsgNOM	two	3sgACC	TOD.PST.sgO-see-2nplA>sgO
	'You two saw him (earlier today).'			
(9c)	*ybi*	*ʈayebibi*	*obom*	*na-nḍəg-la.*
	12NOM	many	3sgACC	TOD.PST.sgO-see-1nsgA>sgO
	'We (you, me and others) saw him/her (earlier today).'			
(10a)	*bi*	*komblebe*	*bibim*	*ña-nḍəg-la.*
	1nsgNOM	two	2nsgACC	TOD.PST.nsgO-see-1nsgA>nsgO
	'We two (excl.) saw you (non-singular) (earlier today).'			

(10b) *bi* *komblebe* *obim* *ña-nḍəg-la.*
1nsgNOM two 3nsgACC TOD.PST.nsgO-see-1nsgA>nsgO
'We two (excl.) saw them (earlier today).'

(10c) *be* *komblebe* *bim* *ña-nḍəg-la.*
2nsgNOM two 1nsgACC TOD.PST.nsgO-see-2nsgA>O
'You two saw us (exclusive) (earlier today).'

(10d) *ybi* *komblebe* *obim* *ña-nḍəg-la.*
12nsgNOM two 3nsgACC TOD.PST.nsgO-see-1nsgA>nsgO
'We two (inclusive) saw them (earlier today).'

As in Nen, diathetic changes such as reflexive/reciprocal are signalled by verbal prefix. The verb *boku* 'cut' (far past stem *kon*), for example, normally takes various forms of prefix according to object values (e.g. *gakon* for 'I cut you (sg)', *bekon* for 'I/her cut him/her'). But the reflexive/reciprocal employs a person/number invariant prefix form *gwa-*, along with a person-sensitive reflexive pronoun formed by the possessive pronoun plus *ḍagəmende*, e.g. *oba ḍagəmende* 'themselves', or a reciprocal/reflexive pronoun formed by reduplicating the possessive pronoun (e.g. *baba* 'ourselves (exc.)/ each other'). Examples are:

(11a) *ŋən* *bo-ḍagəmende* *gwa-ko-n* *tətəm*
1sgNOM 1sgPOSS-REFL RR:RemPst-cut-1|3sgA yesterday
'I cut myself yesterday.'

(11b) *be* *bene-ḍagəmende* *gwa-ko-ya* *tətəm*
2NOM 2sgPOSS-REFL RR:RemPst-cut-2sgA yesterday
'You cut yourself yesterday.'

(11c) *bi* *baba* *gwa-ko-ma* *tətəm*
1sgNOM 1nsgRR RR:RemPst-cut-1nsgA yesterday
'We (exclusive) cut each other yesterday.'

(11d) *bo* *komblebi* *obaoba* *gwa-ko-yo* *tətəm*
3NOM two 3nsgRR RR:RemPst-cut-3duA yesterday
'They two cut each other / themselves yesterday.'

To conclude this brief sketch we illustrate the use of infinitive verbs inflected for case in phasal complements, which parallel Nen in their structure. The phasal auxiliary agrees with both arguments of the verb, and the infinitive, placed before it, is inflected for an appropriate case, such as the allative in constructions meaning 'to be about to' (12a,b).

(12a) *Bi* *babom* *koko-awa* *deada* *nalala*
1plNOM 2sgACC cut(INF)-ALL be.about.to Tr.AUX:1nsg>2sg
'We two are about to cut you.'

(12b) *ŋən* *bibim* *komblabe*
1plNOM 2nsgACC two

koko-awa *deada* *ñere*
cut(INF)-ALL be.about.to Tr.AUX:1sg>2du
'I am about to cut you two.'

As stated earlier, though claimed as related to Nen and the other Yam languages by such earlier classifications as Wurm (1982:182-4, inside his 'Trans-Fly Stock'), and Ross (2005), a more sober assessment of the present evidence does not find support for this position, and it seems more prudent to consider the Pahoturi and Yam families as unrelated (as always, pending evidence to the contrary). None of the morphological paradigms which are probative of genetic relationship show significant resemblances between Nen and Idi – free pronoun, bound pronominal affixes to the verb, or case suffixes.

4.3. Nen and Idi: a brief typological comparison. Nen and Idi, as mentioned above, belong to totally distinct language families, but are linked by strong ties of intermarriage and bilingualism. They show an interesting mixture of typological convergence and divergence which I briefly summarise here.

Firstly, there are significant convergent features. These include:

(a) the employment of both prefixing and suffixing on transitive verbs, with the prefix basically used for the undergoer and the suffix basically used for the actor, though with some leakage. The use of both prefixes and suffixes on the verb is in fact widespread though the Southern New Guinea region, being found in Eastern Trans-Fly languages like Meryam Mir (Piper 1989), in Marind (Drabbe 1955), and in Marori (Arka this volume), as well as throughout the Pahoturi River and Yam families.[20]

However, the functions of the prefix and suffix slots in these languages are different from what we find in Nen and Idi. In Marind both actor and undergoer arguments are generally cross-referenced by prefixes only (leaving aside one specialised construction which uses undergoer suffixes). In Marori only suffixes are employed for agreement on lexical verbs – it is just the auxiliary that uses both prefix and suffix slots. And in Meryam both arguments are cross-referenced by the prefixes, except for some number marking effected by suffixation.To the south and north, the Western Torres Strait language (Pama-Nyungan; Australian) and Suki (Trans New Guinea) are exclusively suffixing. In this sense, the shared pattern of U-prefixation and A-suffixation between Nen and Idi (and more generally between the Yam family and Pahoturi River languages) is significant.

(b) the existence of underspecified or disjunctive semantic values for these verbal affixes, which means that the verb plus free NPs need to be unified before person/number values are resolved. The level of underspecification, however, is much greater in Idi than in Nen.

(c) the location of coding site for argument agreement alternates between finite main

[20] The use of prefixes and suffixes is also found elsewhere in New Guinea – for example, in Goroka-Kainantu languages of the Trans-New Guinea family.

verb in simple constructions and auxiliary verb in non-finite constructions. Auxiliary constructions are more extensive in Idi than in Nen. This reflects the fact that in Idi they are the basic construction in the present, and the auxiliary indicates ongoing aspect (as well as serving as a light verb for many verb lexemes) whereas in Nen the auxiliary is reserved for phasal constructions (begin to, finish).

(d) both languages are verb-final, but this is so widespread in New Guinea that it has little or no distinctive value.

Passing now to divergent features, which are much more numerous, the most significant among them are:

(a) the different organisation of case, both on pronouns and on nouns. Nen has an absolutive/ergative system throughout (apart from the neutralisation of absolutive and ergative for 1st and 2nd singular pronouns); Idi has a nominative/accusative system for pronouns and a highly unusual system opposing a 'direct' case (used in A, S and O functions) to a zero form (used in nomination, and nominal predicates).

(b) Nen lacks an inclusive/exclusive distinction; Idi has one.

(c) Nen forms its infinitives by suffixation to the stem (e.g. √*esr* 'descend', *esrs* 'to descend'); Idi forms its infinitives either by reduplication (e.g. √*ʈme* 'close', *ʈmeʈme* 'to close, closing'; √*ko* 'cut', *koko* 'to cut, cutting'), by using the bare stem (e.g. √*trem* 'open (tr.)', *trem* 'to open, opening'), or introducing some other modification to the stem (e.g. √*nɖog* 'burn' *ɖoŋg* 'to burn').

(d) Nen has an indigenous power-based senary system; in Idi these are extremely marginal and clearly borrowed

(e) Nen has a rich set of postural/positional verbs – about thirty verbs with meanings like 'be the end of something', 'be up high', 'be wedged', 'be in a tree fork' and so on – which have a cluster of distinct morphosyntactic characteristics and are a central part of the grammatical system. Idi appears to have no such phenomenon.

(f) in the unmarked case – absolutive for Nen, nominative for Idi – Nen doesn't distinguish number for any person, whereas Idi distinguishes number for all persons except second

(g) the dominant person syncretism within the Nen verbal agreement system is second person with third (not unusual in Papuan languages), whereas in Idi it is first person with third (much more unusual), as exemplified in many examples in (6), (7) and (8).

(h) in terms of phonological inventories, Nen has no velar nasal, no retroflexes, a single lateral, and a coarticulated labial-velar series. Idi has a strikingly 'Australian' phoneme inventory, with initial velar nasals, a retroflex series, and two laterals – some speakers have coarticulated labial-velars in some loanwords but otherwise this series is absent.

Short and incomplete as it is, this list should demonstrate how many typological isoglosses separate Nen from Idi, and show that widespread bilingualism and intermarriage between speakers of these two languages has not produced strong convergences of structure (although there are a few, as outlined). At the present stage of research it is too early to tell whether this bespeaks relatively recent contact, or rather indicates that long-standing contact has left the basically different typological profiles of the two languages (and language families) untouched.

5. Areality in Southern New Guinea: the case of the dual. Despite the significant typological variety of the languages found in Southern New Guinea – something illustrated

in a very localised way by the comparison of Nen and Idi in the last section – there are some common typological themes running through the whole region (see also Reesink & Dunn, this issue). In this section I focus on just one – the presence of dual number on the verb, which runs through the region from Marori (Arka this issue) in the west to Kiwai in the east (Ray 1932), though apparently not in Marind, as far as I can determine from Drabbe (1955) who only mentions singular and plural. In fact, most languages of the region have an additional number distinction – adding a trial or paucal, or extending the plural up to a large plural. But for reasons of space I skirt that additional complexity here, since my goal is to focus on the rather different ways that the same result – a grammatical category expressing dual number on the verb – can be put together in interestingly different ways in different families.

One of Greenberg's well-known universals about morphological categories states that:

> No language has a trial number unless it has a dual. No language has a dual unless it has a plural. (Greenberg 1963)

A morphological consequence one might expect from this would be that duals are built up from plurals. This is indeed the case in many languages, e.g. the pronominal object prefix system in Bininj Gun-wok (Evans 2003), and it is found in some languages of the Southern New Guinea region. The Idi copula provides a clear example: the singular form is *da*, the plural is built up from this (*dag*), and the dual in turn is built up from the plural (*dago*).

A second possibility is to first distinguish singular from non-singular, then to distinguish dual from plural in an equipollent way, i.e. there is no obvious way of deriving either non-singular form from the other. Kala Kawaw Ya is an example of this strategy. Taking the perfective form of the verb 'cut oneself' as an example, the singular adds the vowel *-i* to the root (plus final *-z*) whereas the non-singulars add *-e*. The non-singulars then add suffixes from a pair where neither has a claim for priority: dual *-man* vs plural *-min*. This gives the three form series sg *pathiz*, du *patheman*, pl. *pathemin*.

A third possibility is to have a category merging singular and dual (let us call this non-plural) and cross it with a singular vs non-singular distinction. This system is found in Hopi, for example (Hale 1997). Within the southern New Guinea region it can be exemplified from the paradigm of 'to be' in Warta (Thundai), a language of the Tonda branch of the Yam family. In the present tense, the root for 'be' is *-iyene* in the non-plural but *-ərei* (1st) or *-ero* (2nd/3rd) in the plural, while the pronominal prefixes are organised on a singular vs non-singular basis. This is illustrated in table 10.

		Singular				Non-singular	
		1 w-	2 n-	3m s-	3f w-	1 n-	2|3 ø/y-
be:NPl	-iyene	*wiyene*	*niyene*	*siyene*	*wiyene*	*niyene*	*iyene*
		1sg:be	2sg:be	3sg.m.be	3sg.f.be	1du.be	2|3du.be
be:Pl	*-ero*					*nərei*	*yero*
						1pl.be	2|3pl.be

Table 10. Composing the dual of 'be' in Warta Thundai by crossing singular vs non-singular and non-plural vs plural distinctions.

A fourth possibility, already illustrated for Nen in §4.1, is to derive a three-way number system by crossing a singular vs non-singular with a dual vs non-dual system. This is highly unusual typologically, but found in several languages of the Nambu branch of the Yam family. Nama, for example (Siegel 2012), has a very similar system to that found in Nen (table 11). Illustrating with the actor suffixes of the past perfective tense, and using the verb *injoy* 'to catch sight of' prefixed for 3sg undergoer, we obtain the following paradigm. A singular vs non-singular organisation of the actor suffixes crosses with a dual (*-ea*) vs non-dual (*-ø*) organisation of the thematic element appearing between the verb stem and the past tense suffix *-y*.

	1sgA: *-n*	2|3sgA: *-ø*	1nsgA: *-m*	2|3nsgA: *-nd*
nd *-ø-*	*yinjoyn*	*yinjoy*	*yinjoym*	*yinjoynd*
du *-ea-*			*yinjoeaym*	*yinjoyeaynd*

Table 11. Partial verb paradigm for the past perfective of *y-injo-_-y_* [3sgU-catch.sight.of-_-PstPerf-_] 'caught sight of it' (Siegel 2012). Thus *yinjoyn* is 'I caught sight of it', *yinjoyeaym* is 'we two caught sight of it', etc.

Intriguingly, this pattern is not confined to the Nambu languages. Within the Eastern Trans-Fly branch, Meryam Mir (Piper 1989) exhibits a very similar pattern, though the distribution of information is different: the singular vs non-singular contrast is found in the free pronouns, while the dual vs non-dual contrast is found in the pronominal prefixes. There are two further interesting twists: the dual is also used for paucals[21] and many verb stems supplete on a singular|dual vs paucal|plural pattern. Two views of the workings of this system are illustrated in tables 12 and 13 (data from Piper 1989:127); note that *(r)edi* is the present-tense suffix to the verb, *e* and *wi* are the third singular and third non-singular free pronouns, and *(i)mi* and *(e)mr* are the singular|dual and paucal|plural stems of 'sit'.

	sg|pl *ø-*	du|pauc *na-*
sg|du *(i)mi*	*imiredi* 'he is sitting'	*na-miredi* 'they (two) are sitting'
pauc|pl *(e)mr*	*emredi* 'they (pl) are sitting'	*na-mredi* 'they (pauc) are sitting'

Table 12. 'Sit' and number in Meryam Mir. Inflected verb only; all four numbers, showing suppletive stem.

[21] Though these terms are not used in descriptions, it would make sense to talk of an 'outer' vs 'inner' contrast in number, where outer is singular or plural, and inner is dual or paucal.

	sg|pl *ø-,*	du|pauc *na-*
3sg *e*	*e ø-imiredi* 'he is sitting'	
3nsg *wi*	*wi ø-emredi* 'they (pl) are sitting'	*wi na-mredi* 'they (pauc) are sitting'

TABLE 13. 'Sit' and number in Meryam Mir, showing interaction with free pronouns but omitting paucal.

From the examples considered in this section, it is clear that having a dual category on verbs is a clear typological feature of the Southern New Guinea region. However, the means by which languages build this up span a radically varied range of methods (including some extremely rare ones typologically), suggesting a large number of individual convergence pathways brokered by a common semantic target. Further consideration of this question – taking into account more languages, more patterns within each (for the sake of exposition I have picked particular illustrative patterns which are by no means the only ones in a given language), and the further complications brought in by a fourth number – is likely to reveal an even more intricate set of developments. It may also suggest earlier contact scenarios – is it possible that the presence of such similar but typologically unusual ways of constructing the dual in the Nambu branch of the Yam family and in Meryam Mer from the Eastern Trans-Fly family reflects an earlier period of contact between those families, with the Pahoturi River languages being a later intrusion? Until we get more data on the various languages involved it is too early to answer this question.

6. DOCUMENTING THE LANGUAGES OF SOUTHERN NEW GUINEA: CHALLENGES AHEAD. The main purpose of this article has been to give a small taste of how much interest and diversity is presented by the languages of Southern New Guinea, in terms of their structures, sociolinguistic settings and historical and areal trajectories – for more detail than could be given here, the reader is referred to Evans (forthcoming a,b). As pointed out in the introduction, our knowledge of virtually every language of the region is extremely basic, even by the standards of New Guinea in general, which is in its turn the least-documented part of the world linguistically. Getting data on these varied and unusual languages is therefore of the highest scientific priority.

In terms of the urgency of the task, the status of the languages is very different according to the country concerned. In Papua New Guinea most of the languages are reasonably secure and are being transmitted to children despite the small speaker-populations, though there are nonetheless individual languages within the Yam family which are close to extinct (e.g. Len, said to be down to just one speaker) or receding from use (e.g. Rema, around Weam near the Indonesian border). In Australia, the sole Papuan language (Meryam Mir) is only spoken by people of middle age or above. Likewise in Indonesia, many of the languages – Marori, Maklew[22], Yei and Kanum are all clear examples to varying extents

[22] Cf. this quote on the status of Maklew, from Lebold et al (2010:25): 'The people who speak the Maklew language seem to be a small group. They also seem much less proud of their language and culture than the Marind people do. The adults in Welbuti complained to the survey team that their children do not speak their language and sometimes make fun of them for using it. The

– are only spoken by people of middle age or above and are significantly endangered. Marind is often said to be in better shape, but there have been no recent detailed studies of the language which could verify this. As these examples make clear, documentary work on languages on the Indonesian side of Southern New Guinea is a particularly urgent priority.

Beyond that, a closer study of the whole Southern New Guinea region will plainly lead to many discoveries – of a host of undescribed linguistic phenomena, of the dynamics of village multilingualism and its effects on language change, of the forces that drove the expansion of Trans-New Guinea languages, of a complex process of relatively recent colonisation as the land was built up over the last few millennia, of contacts between Papuan and Australian languages across the Torres Strait, linked by an Australian language (Kala Kawaw Ya / Kala Lagaw Ya) in the Western Torres Strait and a Papuan language of the Eastern Trans-Fly family in the Eastern Torres Strait (Meryam Mir).

At present, the only reasonable-sized published grammars we have for the whole region are a bunch of papers for the Western Torres Strait language (see Ford & Ober 1991 for onward references), an unpublished MA Thesis for Meryam (Piper 1989), and relatively complete but now outdated grammars from an earlier era for Kiwai (Ray 1932) and Marind (Drabbe 1955). For the Yam family, the Pahoturi River family, all other members of the Eastern Trans-Fly family, for Suki and other TNG languages along the southern bank of the Fly, we have minimal documentation. Finally, the fact that the languages of Southern New Guinea depart in so many ways from what we have come to regard as 'typical' of Papuan languages will have the salutary effect of making us realise that Papuan languages are even more diverse than we had thought – however difficult it is to grasp these even greater levels of diversity.

There are currently a number of projects under way on languages of the Trans-Fly. These include Marori (Wayan Arka), Nen (this author), Nama (Jeff Siegel), Kámnzo (Christian Döhler), Warta Thundai (Kyla Quinn), Taeme (Philip Tama), Kanum (Matthew Carroll), Ranmo (Jessica Thiessen) and Suki (Charlotte van Tongeren). However, this still leaves a large number of languages in urgent need of research, and I hope this article will lead other scholars to undertake work in this fascinating and little-known region.

References

Alpher, Barry, Geoffrey O'Grady & Claire Bowern. 2008. Western Torres Strait language classification and development. In Claire Bowern, Bethwyn Evans and Luisa Miceli (eds.), *Morphology and language history: In honour of Harold Koch*, 15-30. Amsterdam: John Benjamins.

children themselves reportedly do not speak the vernacular, although they are able to understand it. This was a concern to the adults because they were afraid their language would die out; they said their language would not be spoken anymore in 20 years. Nevertheless, the survey team did notice some older women who only spoke the vernacular language and were not able to speak Indonesian.'

Arka, Wayan. 2011. Constructive number systems in Marori and beyond. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG11 conference*. CSLI Publications. http://cslipublications.stanford.edu/LFG/16/papers/lfg11arka.pdf

Arka, Wayan. This volume. Projecting morphology and agreement in Marori, an isolate of Southern New Guinea.

Ayres, Mary C. 1983. *This side, that side: Locality and exogamous group definition in the Morehead area, southwestern Papua.* Chicago, IL: University of Chicago PhD thesis.

Blevins, Juliette & Andrew Pawley. 2010. Typological implications of Kalam predictable vowels. *Phonology* 27. 1-44.

Corbett, Greville G. 2000. *Number*. Cambridge: Cambridge University Press.

Donohue, Mark. 2009. Syllables, morae and vowels in Kanum. Seminar Handout, Australian National University. Unpublished ms.

Drabbe, P. 1955. *Spraakkunst van het Marind*. (Studia Instituti Anthropos 11). Mödling: Missiehuis St. Gabriël.

Evans, Nicholas. 2003. *Bininj Gun-wok: a pan-dialectal grammar of Gun-djeihmi, Kunwinjku and Kune* (Pacific Linguistics 541). Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University.

Evans, Nicholas. 2005. Australian languages reconsidered: a review of Dixon (2002). *Oceanic Linguistics* 44. 242-286.

Evans, Nicholas. 2009. Two *pus* one makes thirteen: senary numerals in the Morehead–Maro region. *Linguistic Typology* 13. 321-335.

Evans, Nicholas. Forthcoming a. Valency in Nen. In Andrej Malchukov, Martin Haspelmath, Bernard Comrie & Iren Hartmann (eds.), *Valency Classes in the World's Languages*.

Evans, Nicholas. Forthcoming b. Inflection in Nen. In Matthew Baerman (ed.), *The Oxford Handbook of Inflection*.

Evans, Nicholas. Forthcoming c. The languages of southern New Guinea. In Palmer, *Languages and linguistics of New Guinea.*

Ford, Kevin & Dana Ober. 1991. A sketch of Kalaw Kawaw Ya. In Suzanne Romaine (ed.), *Language in Australia,* 118–142. Cambridge: Cambridge University Press.

Greenberg, Joseph. 1963. *Universals of Language*. London: MIT Press.

Hale, Ken. 1997. Some observations on the contribution of local languages to linguistic science. *Lingua* 100. 71-89.

Hammarström, Harald. 2009. Whence the Kanum base-6 numeral system? *Linguistic Typology* 13. 305–319.

Hitchcock, Garrick. 2010. Mound-and-ditch taro gardens of the Bensbach or Torassi River area, southwest Papua New Guinea. *The Artefact* 33. 70-90.

Latham, R.G. 1852. Appendix. On the general affinities of the Oceanic Blacks. In J. McGillivray, *Narrative of Surveying Voyage of H.M.S. Fly,* 313-320. London.

Lebold, Randy, Ron Kriens, and Peter Jan de Vries. 2010. Report on the Okaba subdistrict survey in Papua, Indonesia (SIL Electronic Survey Report 2010-008). http://www.sil.org/silesr/2010/silesr2010-008.pdf

Palmer, Bill (ed.). Forthcoming. *The languages and linguistics of New Guinea: A comprehensive guide*. Berlin: De Gruyter.

Pawley, Andrew. 2005. The chequered career of the Trans New Guinea hypothesis: recent research and its implications. In Pawley, *Papuan pasts*, 67-108.

Pawley, Andrew, Robert Attenborough, Jack Golson & Robin Hide (eds.). 2005. *Papuan pasts: Cultural, linguistic and biological histories of Papuan-speaking peoples*. Canberra: Pacific Linguistics.

Pawley, Andrew. 2008. Recent research on the historical relationships of the Papuan languages, or, what does linguistics say about the prehistory of Melanesia? In Jonathan Friedlaender (ed.), *Population genetics, linguistics and culture history in the southwest Pacific*. Oxford: Oxford University Press.

Pawley, Andrew & Harald Hammarström. Forthcoming. The trans-New Guinea family. In Palmer, *Languages and linguistics of New Guinea.*

Piper, Nick. 1989. A sketch grammar of Meryam Mir. Canberra: ANU MA thesis.

Ray, Sidney H. 1923. The languages of the Western Division of Papua. *JRAI* 53. 332-60.

Ray, Sidney H. 1932. *A grammar of the Kiwai Language, Fly Delta, Papua. With a Kiwai vocabulary by the late Rev. E. Baster Riley*. Port Moresby: Edward George Baker. [This source is cited in the literature with various dates (1931, 1932 and 1933) and determining its exact date of appearance is problematic, since the fly leaf contains no date, while the preface by Ray bears the date 1931. I stick to 1932 as the commonest year with which it is cited.]

Reesink, Ger, Ruth Singer & Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS Biology* 7(11). e1000241.

Ross, Malcolm. 2005. Pronouns as a preliminary diagnostic for grouping Papuan languages. In Pawley et al., *Papuan pasts*, 15-66.

Siegel, Jeff. 2012. The significance of the numbers 2 and 6 in the Nama language (Papua New Guinea). Powerpoint Presentation, Freiburg Institute for Advanced Studies. Unpublished ms.

Suter, Edgar. 2010. Object Verbs in Huon Peninsula languages. Unpublished ms.

Van Baal, Jan. 1966. *Dema. Description and Analysis of Marind-Anim Culture (South New Guinea)*. The Hague: Martinus Nijhoff.

Voorhoeve, C.L. 1970. Some notes on the Suki-Gogodala subgroup of the Central and South New Guinea Phylum. In Stephen A. Wurm & Donald C. Laycock (eds.), *Pacific linguistic studies in honour of Arthur Capell* (Pacific Linguistics C 13), 1247-1270. Canberra: Pacific Linguistics, Australian National University.

Vries, Lourens de. 2004. *A short grammar of Inanwatan, an endangered language of the Bird's Head of Papus, Indonesia* (Pacific Linguistics 560). Canberra: Pacific Linguistics, Australian National University.

Williams, F.E. 1936. *Papuans of the Trans-Fly*. Oxford: Clarendon.

Wurm, Stephen. 1982. *The Papuan languages of Oceania*. Tübingen: Narr.

WWF TransFly Team. 2006. Conservation Scenarios for the Transfly Ecoregion New Guinea. Powerpoint Presentation, Brisbane. Unpublished ms.

Nicholas Evans
nicholas.evans@anu.edu.au

6

Language Documentation & Conservation Special Publication No. 5 (December 2012)
Melanesian Languages on the Edge of Asia: Challenges for the 21st Century,
ed. by Nicholas Evans and Marian Klamer, 150–173
http://nflrc.hawaii.edu/ldc/sp05/
http://hdl.handle.net/10125/4563

Projecting morphology and agreement in Marori, an isolate of southern New Guinea

I Wayan Arka
Australian National University/
Universitas Udayana

This paper is the first detailed investigation on agreement in Marori (Isolate, Papuan, Merauke-Indonesia), highlighting its significance in the cross-linguistic understanding of NUM(BER) expression and in the unification-based theory of agreement. Marori shows PERS and NUM agreement with distributed exponence in DUAL. The paper proposes that DUAL is formed by two basic NUM features (SG, PL) each with its binary values and that DUAL is [-SG,-PL] (unmarked). The novel aspect of the analysis is the idea that the NUM feature is mapped onto a language-specific structured semantic space of NUM. A morpheme is analysed as carrying a feature bundle, with the semantic spaces referred to by the individual features possibly overlapping with each other. The proposed analysis can provide a natural explanation for NUMBER agreement in Marori and can be extended to account for unusual cases of NUM agreement and expression in other languages.

1. Introduction. Marori[1] is a Papuan language (isolate, Trans New Guinea (Ross 2005)). It is spoken by the Marori people in Kampung Wasur, around 15 kilometres east of Merauke, Indonesian Papua.

Marori is under-documented. Previous publications mentioning this language (Boelaars 1950, Wurm 1954) mainly originated from the work of the Dutch missionary Father P. Drabbe, who also published his own work on the languages of southern New Guinea (Drabbe 1954, 1955). Mark Donohue collected a word list and also produced a picture dictionary (Gebze & Donohue 1998). A sociolinguistic survey was undertaken by SIL (Sohn, Lebold & Kriens 2009) on languages around Merauke including Marori.

[1] Alternative names are Morori, Moaraeri, Moraori, and Morari.

Marori language is highly endangered. There are several fluent speakers left, out of a total of 52 families or 119 people. Marori people typically have mixed marriages with Marind and non-Papuan Indonesians such as the Tanimbar people and currently the Javanese.[2] The sociolinguistic survey carried out in 2000 (Sohn, Lebold & Kriens 2009) reports the precarious nature of the language, which I further confirmed when I did my fieldwork in 2008 and 2009. Young Maroris no longer actively speak their langauge. They may, however, still have passive competence of varying degrees. They almost all speak Indonesian or the local variety of Indonesian/Malay, and also Marind.

This paper is the first detailed investigation on agreement in this language, highlighting its significance for the unification-based theory of agreement. Marori shows PERS and NUM agreement at the clausal level (between the predicate and its argument) and at the phrasal level (between a noun and its determiner). Of particular interest is the issue of distributed exponence in DUAL expression and agreement. It is proposed that there are two basic NUM features, each with its binary values ([+/-SG], [+/-PL]), and that they are semantically grounded on language-specific structured semantic space of NUM. A morpheme carries a feature bundle which allows the morpheme to refer to different portions of the semantic space. It is demonstrated that DUAL agreement in Marori can be dealt with in a straightforward manner using a unification-based analysis as compositionally [-SG,-PL] (unmarked). The analysis can be naturally extended to cases of DUAL in other languages with more complex NUM systems such as Nen, Hopi and Larike.

The paper is organised as follows. Section 2 outlines basic facts on clausal organisation and agreement types in Marori. Section 3 starts with the conception of agreement from a lexicalist point of view, followed by the discussion on the nature of agreement in Marori. This section also outlines the proposal pertaining to feature structures with their corresponding semantic space and the analysis of the distributed exponence of NUM. Section 4 demonstrates how the analysis of DUAL in Marori can be extended to account for more complex NUM categories in other languages. Conclusions are given in section 5.

2. Basic facts on Marori syntax

2.1. Marori clausal syntax in brief. The basic clause structure in Marori is shown in figure 1. The verbal structure typically consists of a lexical verb (VERB) and a light or auxiliary (AUX) verb. The AUX verb can be morphologically complex consisting of an AUX root and possibly one or more affixes. This is depicted in figure 2.[3]

[NP* VERB AUX.VERB]CLAUSE

Figure 1

[2] Most of the Javanese people coming to west Papua were originally part of the transmigration program sponsored by the government. They are now the first or second generation born in Merauke, and call themselves Jamer (Jawa Merauke).

[3] Different lexical verbs may take different AUX verbs, depending on lexical classes. This is not discussed in this paper.

PREF-AUX.ROOT-SUFF
(U) (A)
(S) (S)

FIGURE 2

Agreement on the (auxiliary) verb is expressed by portmanteau affixes with the prefix showing undergoer agreement and the suffix actor/subject agreement. Table 1 shows free pronouns and their corresponding agreement affixes on the verb. For simplicity, only the actor suffix paradigm in the Past Tense is given.[4]

	FREE PRONOUN	UNDERGOER PREFIX	ACTOR SUFFIX (PAST)	
1SG	*na*	*i-*	*-men*	'NonPL'
1NonSG	*nie*	*yar-(<i-ar)* 'NonSG.PST/DU.PRES' *yor-(<i-or)* 'NonSG.FUT/PL.PRES'	*-men* *-ben*	'NonPL' 'PL'
2SG	*ka*	*k-*	*-m*	'NonPL'
2NonSG	*kie*	*kar-* 'NonSG.PST/DU.PRES' *kor-* 'NonSG.FUT/PL.PRES'	*n- -m* *n- -b ~ -im*	'2.NonPL' '2.PL'
3SG	*efi*	Ø	*-m*	'NonPL'
3NonSG	*emnde*	Ø	*-m* *-b ~ - im*	'NonPL' 'PL'

TABLE 1. Free pronouns, Undergoer prefixes, and past actor suffixes in Marori

Examples with agreement morphology on the AUX are given in (1).[5] However, certain verbs of high frequency in daily life, e.g. 'sit' and 'bring', often have irregular inflection or the TNS/PERS morphology directly on the verbs, e.g. *-du* '1SG(PRES)'[6] on the verb 'bring/take' in (2).[7]

[4] Abbreviations used in the glosses:1/2/3 'first/second/third person', A 'actor', ABS 'absolutive', AUX 'auxiliary', DET 'determiner', DU 'dual', ERG 'ergative', F 'female', FUT 'future', M 'male', NonPL 'non plural', NonSG 'non singular',O 'object', NrPST 'near past', PERF 'perfective', PL 'plural', POSS 'possessive', PROG 'progressive', REFL 'reflexive', SG 'singular', PRES 'present', PST 'past', U 'undergoer',

[5] There is no standard orthography for Marori yet. This paper follows the Indonesian-like orthography commonly used by my Marori consultants, e.g. *y* represents the approximant /j/ and *ng* the velar nasal /ŋ/. Consonants with prenasals are written with more than one symbol, e.g. *mb*, *nd*, and *ngg*. Bilabial fricatives are written as *f* (voiceless) and *v* (voiced).

[6] The concept PRESENT in Marori can cover the time NOW (today) and often yesterday.

[7] Auxiliaries and lexical verbs are independent words, each can have their own affixes. When they

(1) a. *Ka* *ku=ndo-Ø.*
2SG run=AUX.2/3NonPL-NPST
'you (sg) will run'

b. *Nawa* *payung=i* *nde=ngge-ben.*
1SG umbrella=U buy=AUX-1NrPST
'I bought an umbrella.'

(2) *ujif* *ke=me=na* *fis* *ndon-du* *tamba* *yabah* *ngguo-f.*[8]
bird REL=wish =1SG yesterday bring-1SG PERF die AUX.DU-PST
'The (two) birds that I wanted to take with me yesterday were already dead.'

Marori shows split intransitivity. The patientive S and Patient/Object is (typically) marked by *=i*.

These examples show the contrast where the patientive S argument *na* 'I' is marked by *=i* (3a) whereas the agentive S must not be marked by *=i* (3b).

(3) a. *Na=i* *patar* *yu-nggo-f.*
1SG=U cold 1SG-AUX.1/2-PST
'I suffered from cold.'

b. *Efi* *ramon (*=i)* *ku=ndo-f.*
that woman run=AUX.2/3NonPL -PST
'She/the woman ran off.'

The split appears to be skewed: only patientive S of states as in (3a) is treated as object-like. Patientive S of motion predicates like 'fall' receives suffix agreement:

(4) *Nie* *yanadu* *purfam* *pa=saron-den* *kwi* *uyow* *ngge.*
1NonPL two person soon=fall-1DU.PRES tree above from
'We two are about to fall off from the tree.'

The following examples show that *na* in (5a) functions as Subject appearing without *=i*. In (5b), it functions as object; hence taking *=i.*

(5) a. *Tamba=na* *Albert=i* *keswe=mi-men.*
already=1SG Albert=U hit=3SGM.AUX-1NonPL.PST
'I already hit Albert.'

---

appear together forming phonological words, they are separated by a = (a notation conventionally used for clitics), e.g. *ku=ndo-Ø* in (1). Likewise, the same convention is applied to similar cases such as the free pronoun *na* as in *tamba=na* 'already=1SG' (6) or the beneficiary postposition *na* which can becomes *=n* forming phonological words with other items as in *na=n=du* '1SG=for=REFL' as in (7b).

[8] A more precise gloss for *ngguo-* would be 'AUX.NonSG.NonPL': the three-way distinction of *nggu* 'AUX.SG', *nggo* 'AUX.PL' and *ngguo* 'AUX.DU' suggests that vowel *-u* attached to the root *ngg-* is actually associated with 'NonPL' and *-o* with 'NonSG'.

b. *Efi* *purfam* *na=i* *kaswa=ri-ma-m.*
that person 1SG=U hit=1-AUX-2/3NonPL.PST
'The person hit me.'

In a three-place (ditransitive) structure, the Goal/Recipient argument object is marked by *=i*. The verb is inflected showing agreement with this Goal NP, in addition to the agreement with the actor NP.

(6) a. *Tamba=na* *Robertus=i* *bosik* *nji=me-feri.*
already=1SG Robertus=U pig 3M.give=AUX-1.RPST
'I already gave Robert a pig (a long time ago).'

b. *Tamba=na* *Maria=i* *bosik* *njo=mo-fori.*
already=1SG Maria=U pig 3F.give=(F.)AUX-1.RPST
'I already gave Maria a pig (a long time ago).

c. *Robertus/Maria* *na=i* *bosik* *i=mo-fi.*
Robert/Maria 1SG-U pig 1SG.give=AUX-2/3.RPST
'Robert or Maria gave me a pig.'

A beneficiary participant in a three-place predicate is marked by *=na or =n*. The verb agrees with the theme/patient, not with the beneficiary NP:

(7) a. *Maria* *ka=na* *di* *bosik* *eyew* *Ø-nda-Ø* *tanamba.*
Maria 2SG=for soon pig see 3-AUX-2/3.NPST now
'Maria readily hunts a pig for you now.'

b. *Nawa* *fis* *nandu* *dakai* *tawramon.*
1SG yesterday na=n=du daka=i taw=Ø-ramon
1SG=for=REFL water=U take=3-AUX.1NonPL.PST
'I bailed water out for myself yesterday.'

2.2. Agreement types in Marori. Agreement in Marori is of two types: clausal predicate-argument agreement and phrasal noun-determiner agreement.

In the predicate-argument agreement, the core arguments (subject and object) agree in PERS and NUM with either the AUX or the main lexical verb, or both. The first and most common pattern is the one where the AUX is inflected and the lexical predicate remains constant in its form. In the following examples, the AUX is inflected (*nadam*, *ndamon*) whereas the lexical verb *eyew* is not inflected:

(8) a. *Kie* *tamba* *Maria=na* *bosik* *eyew* *nadam.*
2NSG PERF M=for pig see Ø-n-nda-m
3-2NonSG-AUX-2/3NonPL.PST
'You (DU) hunted a pig for Maria.'

b. *Nawa* *fis* *Maria=na* *bosik* *eyew* *ndamon.*
1SG yesterday Maria=for pig see Ø-nda-mon
3-AUX-NonPL.PST
'I hunted a pig for Maria.'

Note that the agreement morphology may have distributed exponence within the AUX. This is seen in (8a), where the second person dual actor past tense agreement in *nadam* is formed by the discontinuous formatives *n-* and *-m,* added to the auxiliary root *nda* (cf. table 1).[9]

In the second type the verb itself is inflected to show agreement. This verb is typically associated with activity of high frequency in daily life such as 'bring' and 'sit'. In the following examples, the agreement morphology *-du* is affixed to the verbs.

(9) a. *Pa=na* *ka=na* *ujif* *nde-du* *sokodu.*
soon=1SG 2SG=for bird bring -1SG.PRES one
'I (will) bring one bird for you.'

b. *Nawa* *kursi* *uyowé* *kuye-du.*
1SG chair on.top sit-1SG.PRES
'I sit on a chair.'

The third type is inflection on both the lexical predicate and the AUX. This is the case with predicates that encode certain qualities such as 'red' and 'big'. These predicates are inflected for NUM showing opposition of SG and NonSG. The inflection may be morphologically regular (e.g. *para* 'red' → *para-won* 'red.SG' ; *para-nde* 'red.NonSG') or suppletive (e.g., *siel* 'big.SG'; *kofe* 'big.NonSG' *monjun* 'small.SG', *menindum* 'small. NonSG').

Consider the following examples:

(10) a. *Efi* *nam* *pu* *para-won* *te.*
3SG POSS hair red-SG be.(3NonPL.)PRES
'Her/his hair is red.'

b. *Emde* *usindu* *nam* *pu* *para-nde* *te-re(re).*
3NonSG PL POSS hair red-NonSG (3)be-PL.PRES
'Their (PL) hair is red.'

Sentence (10a) shows singular agreement, where the singular suffix *-won* must be used on the lexical predicate *para* 'red' and the auxiliary shows third person NonPL (i.e., Ø) morphology. Sentence (10b) is the counterpart sentence that shows plural agreement. The NonSG *-nde* is used on *para* 'red' and the suffix *-re(re)* on the auxiliary.

In addition to the predicate-argument agreement just outlined, Marori shows agreement between the determiner and the noun head in the noun phrase. The determiner in Marori shows an opposition of SG vs. NonSG: *efi* 'DET.SG' vs. *emnde* 'DET.NonSG'.

(11) a. *efi* *ramon* *sokodu* 'the (one) woman'
DET woman one

[9] The form *nadam* is analysed as having an underlying form *n-nda-m*. The form *nadam* involves epenthesis vowel harmony with the consonant *nd* of the auxiliary *nda* becoming *d.* The prenasal part perhaps becomes the coda of the first syllable (*nan.dam*) which is then weakened and lost (*na. dam)*.

b. *emnde* *ramon* *yanadu* ‘the two women’
DET.NonSG woman two

c. *emnde* *ramon* *usindu* ‘all of the women’
DET.NonSG woman all

3. Projection issues in agreement and proposed analysis. Before discussing the issues posed by agreement in Marori and the proposed analysis, it is useful to have a brief review of the lexicalist approach to agreement.

3.1. A lexicalist theory of agreement. While specific details and mechanisms are different, all theories of agreement operate on the same principles: compatible or same features are allowed to pass through in the formation of larger syntactic structures whereas incompatible ones are not. In a lexicalist non-derivational framework of grammar, e.g. LFG (Bresnan 2001, Dalrymple 2001, Falk 2001) and HPSG (Sag, Wasow & Bender 2003), the mechanism is done via unification of features. Features of the same or compatible values will successfully unify. This can be informally represented in figure 3 for a language like English that requires subject-verb agreement. The NUM feature of [NUM SG] carried by SUBJ NP will become [SUBJ [NUM SG]]. It then unifies with the same feature carried by the verb, as shown figure 3b. If the verb has different, incompatible value, then the unification fails. The feature clashes, indicated by a star in figure 3c.

a. [SUBJNP VP]clause
| |
b. [SUBJ[NUM SG]] U [SUBJ[NUM SG]] = [SUBJ[NUM SG]]
c. [SUBJ[NUM SG]] U [SUBJ[NUM PL]] = *

Figure 3

The challenge is how to develop a linguistically motivated feature structure that can capture the complex patterns of agreement in a particular language and across languages. I argue that NUM features and their structures must be mapped onto the semantic space of NUM which must be established on a language-specific basis. This is further discussed in sub-section 3.3.2 below.

A theory of agreement must also be able to deal with intricacies of different types of agreement including indeterminacy as in case agreement (Dalrymple, King & Sadler 2009) and the domains/relations involved. There are at least three, often inter-related, kinds of agreement: grammatical, semantic and pragmatic (Wechsler & Zlatic 2000, Kathol 1999, Pollard & Sag 1994).[10]

[10] Pollard and Sag (Pollard & Sag 1994) also discussed honorific agreement, e.g. in Japanese and Korean. This type of honorific agreement is also encountered in Balinese, analysed as pragmatic

In grammatical agreement (CONCORD), the agreeing units share grammatical features. Purely syntactic features, e.g. NOM case, are features required by the grammar to encode syntactic relations. For example, grammatical agreement is observed in the Serbo-Croatian NP in (12) where the determiner and adjective agree in case (in addition to gender and number) with the head noun:

(12) *ov-a* *star-a* *knjig-a*
this-NOM.F.SG old-NOM.F.SG book(F)-NOM.SG (Wechsler & Zlatic 2000)

In semantic agreement, the agreeing units share referential indices: PERS, NUM, GEND. These features are essentially semantic because they indicate referents in the external world. However, they are also often grammaticalised in many languages, and are often tied to grammatical features. For example, they are often expressed by a portmanteau morpheme. Thus, the SUBJ-VERB agreement in English is grammatical as well as semantic, because we have cases like *committee are/is ...* or where the noun *committee* can have singular or plural interpretation. The two interpretations allow singular and plural agreement on the verb. The plural agreement shows the agreement with the plural referents of the subject, despite the form of the subject noun being singular.

Note that agreement in English is also grammatical in the sense that the agreeing SUBJ is obligatorily required by the verb and that the agreement verbal morphology makes reference to the syntactic property of subjecthood.

Pragmatic agreement, also called anaphoric agreement (see footnote 11), is a case of co-referential elements which show compatible referential properties. Pragmatic agreement is typically not constrained by certain syntactic domains. Cases showing left dislocation as in English (13a) below belongs to pragmatic agreement. Agreement of the type shown in (13b) from Kambera (an Austronesian language of Sumba, Indonesia) where the free NP subject is optionally present also belongs to anaphoric agreement.[11]

(13) a. *John, I like him very much.*

b. *(I* *Ama)* *na-kei-nja* *ri.*
ART father 3sN-buy-3pD vegetable (Klamer 1996)
'Father buys vegetables for them.'

Predicate-argument agreement in Marori is, as we shall see in the next sub-section basically semantic-pragmatic in nature.

3.2. The nature of agreement in Marori. Agreement in Marori is not grammatical, but semantic-pragmatic in nature. It is not grammatical because the agreement features (NUM and PERS) are essentially referential (hence, semantic) in nature. In Marori, these features are not grammaticalised to become part of an agreement system that makes reference

agreement in Arka (2005). Bresnan & Mchombo (1987) distinguish grammatical and anaphoric agreement.

[11] The distribution of a resumptive pronoun which is a case of anaphoric agreement may be also constrained to a certain degree by termhood/coreness of arguments. For example, resumptive pronouns in Balinese (Arka 2003) are restricted to core arguments (Subject and Object).

to syntactic functions and/or syntactic marking. It should be noted that, while agreement has been described in this paper to involve subject and object, these Subject and Object labels should be understood as macro (semantic) roles. Marori has no syntactic subject/ pivot of the type found in English or certain Austronesian languages such as Indonesian.

Grammatical agreement requires that the agreeing NP be obligatorily present. This is not the case with Marori. The free NP that the verb agrees with is often dropped (i.e. optionally present). This is shown in the following examples, extracted from the Frog story[12] in Marori. Sentence (14c) comes with no free A/U NPs. The agreement morphology on the verb, both of which are zero formatives, anaphorically refers to the NPs mentioned earlier in the texts ('the dog', 'Thomas', and '(the) frog').[13]

(14) a. *Koro Thomas fi njaj uyow ...*
dog Thomas with bed top
'Thomas and the dog were (sleeping) on the bed… .' (FrogStory_Paskalis.009)

(three-four lines later, line 012-0113)

b. *Mar tok reruwo rowae kuya-maf.*
NEG frog jar inside BE.2/3NonPL-PST
'There was no frog inside the jar.'

c. *Mbe tanamba eyew=Ø-nda-Ø-fi.*
PROG now see=3-AUX-2/3NonPL-RPST
'(They were (two)) now looking for (it) (i.e., the frog).'

Further evidence that agreement in Marori is semantic in nature comes from the fact that, when the agreeing NP is present, it is for a functional-semantic reason to create a specific referent. This is the case with DUAL reference. Thus, the 3NonSG pronoun *emnde* 'agrees' with the NonPL Actor suffix *-m* in (15a) to create a dual referent, as the translation shows. When the actor is PL, *-im* is used (cf. table 1) giving rise to *ndim* (15b).

(15) a. *Emnde na=n bosik eyew nda-m.*
3NonSG 1SG=for pig see Ø-nda-m
3-AUX- 2/3NonPL.PST
'They (two) hunted a pig for me.'

b. *Emnde usindu Maria=na bosik eyew ndim.*
3NonSG all Maria=for pig seeØ-nda-im
3-AUX-2/3PL.PST
'They (all, more than two) hunted a pig for Maria.'

It should be noted that the formation of DUAL reference in Marori is achieved by

[12] This is the frog story (*Frog, where are you?*) by Mercer Mayer (1969).

[13] Inflection showing tense in Marori is complex. There is more than one way of doing it, and syncretism adds to the complexity. Past tense for 2/3 Actor, for example, can be expressed by adding the suffix *-f* (PST) or *-fi* (typically remote past (RPST)) as seen in (13c), or adding *–m* 'NonPL' as in (7a) (see also table 1).

combining NonSG and NonPL morphemes in phrasal syntax as well as in word-internal syntax. Example (15) illustrates the formation of DUAL in clausal syntax. Example (8), repeated here as (16a), shows the formation of DUAL within the verb (i.e. the combination of the NonSG actor prefix *n-* and *-m*). Note than the combination of *n- -im* is used when the actor is plural, giving rise to *nedim* as seen in (16b), and no *n-* is used with *-m* when the actor is singular giving rise to *ndam* (16c). Thus, a formal and functional analysis of agreement in Marori must take into account these phrasal and sublexical layers of agreement. The agreement across these layers has to be dealt with in a uniform way. The issue will be further discussed in subsection 3.3.3 below.

(16) a. *Kie* *tamba* *Maria=na* *bosik* *eyew* *nadam.*
2NonSG PERF Maria=for pig see Ø-n-nda-m
3-2NonSG-AUX-2/3.NonPL.PST
'You (DU) hunted a pig for Maria.'

b. *Kie* *usindu* *Maria=na* *bosik* *eyew* *nedim.*
2NonSG all Maria=for pig see Ø-n-nda-im
3-2NonSG-AUX-PL.PST
'You (all, more than two) hunted a pig for Maria.'

c. *Ka* *Maria=na* *bosik* *eyew* *ndam.*
2SG Maria=for pig see Ø-nda-m
3-AUX-2/3.NonPL.PST
'You (SG) hunted a pig for Maria.'

3.3. Projecting morphology

3.3.1. What is projection? The notion of projection is one of the central concepts in modern syntactic theories. It refers to the mechanism by which a (sub-)unit of a structure determines or constrains a larger (syntactic) structure which it is a part of, or a structure it is related to. Thus, one can talk about (lexical-)categorical projection, e.g. a verb (V) (in the lexicon) is projected to verb phrase (VP) in syntax. In Chomskyan terms, the EPP (Extended Projection Principle) is proposed to ensure that the verb which is projected to syntax must have an NP in the subject position (Chomsky 1981). In the LFG model (Dalrymple 2001), the term 'projection' refers to mapping or correspondence between layers of structures.

Projection of morphology to syntax refers to how a morpheme in a sublexical structure determines or constraints the structure of phrasal or clausal syntax of which the word is part. By 'structure' we mean (grammatical) structure of different kinds. The relevant ones for the purpose of the present discussion are semantic (predicate-)argument structure (where A vs. P are relevant), word-internal structure (which agreement affixes are part of), and phrasal and clausal syntax (which the agreeing NPs are part of).

Of particular interest are the projection issues in relation to the agreement patterns presented earlier. Adopting a traditional view where morphology and syntax are two different but related domains of grammar, we have the following questions: how do we maintain the distinction while at the same time capture the idea that the same principle applies across boundary of morphology and syntax? Regarding NUM agreement, what can we learn from Marori in relation to the feature structure of NUM? What is the best analysis, and to what extent is the analysis applicable to other languages?

In what follows, I address these questions. I propose a lexically-based analysis for NUM agreement in Marori where DUAL is not primitive. I sketch how the proposed analysis can be extended to account for complex NUM systems in other languages.

3.3.2. Proposed feature structure and claims. The points of the analysis are the following. First, following Hale (1997), I adopt the analysis that SG and PL features are the most basic NUM features. Each has a binary value (+/-) as shown in figure 4a.

a.	NUM	={[+/- SG], [+/-PL] }	
b.	DUAL	=[-SG,-PL]	(where [-SG] is NSG and [-PL] is NPL)

FIGURE 4

Second, on the basis of figure 4a, the DUAL in a three-way NUM system as observed in Marori (SG, DU, PL) is analysable as being formed out of these basic NUM features, namely [-SG, -PL]. This is shown in figure 4b.

Third, as seen from feature specification in figure 4b, DUAL is unmarked. It is formed out of a combination of two features with negative values.[14] There is evidence from Marori that DUAL is indeed encoded by two underspecified morphemes, e.g. *n- -m* in (16) glossed as NonSG and NonPL respectively. There is also evidence from Nen (a Papuan language of southern New Guinea) where certain verbal stems expressing DUAL are unmarked and the formation of SG/PL is achieved by having additional marking on these stems.

Fourth, while a specific number morphology signals the presence of a number feature, I claim that the absence of number morphology associated with a form does not mean that the form contributes no number feature. What number information is contributed by the form is lexically determined within the larger system of the language. For example, the demonstrative *this or that* in English can be analysed as carrying [NUM [PL -]] because a demonstrative is part of the nominal category in English where plural is morphologically marked. Hence, in our analysis *this/that* carries [PL -][15] (and is compatible with a noun carrying [PL +] such as *children*. The definite article *the*, however, does not enter into

[14] Harley & Ritter (2002) provide an analysis where dual is universally associated with positive specification of both Minimal and Group (semantic) features, roughly corresponding to 'singular' and 'plural' with evidence, for example, coming from Hopi. In this language, dual is expressed by both singular and plural forms. Evidence from Papuan languages as discussed in this paper, however, shows that dual is expressed distributively by two NonSG and NonPL morphemes supporting the analysis that their SG and PL features carry negative values.

[15] One might want to analyse that English singular nouns and demonstratives this/that carry [SG +]. While this is intuitively reasonable, there are good reasons why this analysis is untenable. It would mean that singular is morphologically marked in English (i.e. there is a dedicated morphology to mark singular which is not the case). (The third person singular present tense -s is not solely for number.) In addition, it would lead to an unwanted outcome in the unification process, allowing unacceptable structure with feature unification of [NUM [SG +, PL +] in English such as in **this children*.

number contrast in English and therefore carries no number feature.

The fifth, key and new, proposal is the modelling and conception of NUM system. I argue that the NUM system must be understood as reflecting language-specific categorisations of the semantic space of NUM, and that NUM features are distinguished and structured on the basis of the corresponding structures of the relevant NUM spaces to which they are mapped onto.

I therefore claim that feature operations to establish NUM referents are determined or constrained by the semantic space of NUM of the language. This allows us to provide a natural explanation for certain cases which appear to be unusual, e.g. the coding of exhaustive set/plural or paucal using SG/DUAL morpheme in Nen (discussed in section 4.1 below). In the proposed analysis, different interpretations of PL, which may or may not include the meaning of DUAL/TRIAL as in Larike (discussed in section 4.3) also follow naturally.

To begin with, the simplest model of pairing between (NUM) FORM and its corresponding semantic space (MEANING) is arguably the one shown in figure 5 below. (I will show later that the MEANING part is richly structured with possible overlapping spaces.) The line with an arrow at the end associated with PLURAL is meant to capture the idea that plurality is quantitatively unspecified. In contrast, singularity (i.e. being 'one') is quantitatively specific; hence no arrow is represented at the end of NUM space.[16]

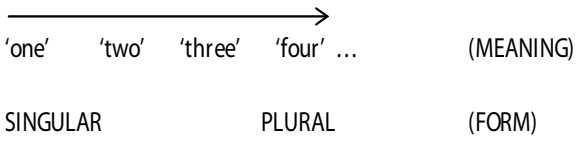


FIGURE 5

Languages differ in the way the space between the two ends is divided and encoded linguistically. To capture the differences and similarities, it is necessary to represent the internal structure of the space explicitly, from which the abstraction of atomic NUM features ([+/-SG] and [+/-PL]) can be postulated. To illustrate the points, the NUM system in Marori is compared with that of English. English is simpler and is discussed first.

English has a two-way NUM system showing SG ('one') vs. PL ('more than one') opposition. It is the PLURAL category that is morphologically marked in this language on nouns. That is, there is a dedicated PLURAL morpheme in English. Given that PL has a binary (+/-) value and that SG is analysable as [-PL], the simplest analysis is that English

[16] Note that we are talking about NUM in natural language semantics. In mathematical sense, one can talk about spaces below one or zero (i.e. minuses), in which case there should be an arrow specified for the line. For simplicity, I also ignore the complication in the conception of NUM in 'mass' nouns, where the FORM is SINGULAR but it does not refer to an individuated 'one' entity, e.g. English *water* and *air*.

employs one NUM feature only, namely the PL feature. The schema figure 6 represents the mapping of the PL feature onto the semantic space of NUM in English. As seen, the feature [+PL] carried by a PLURAL form maps onto a NUM space covering 'more than one', indicated by a thick line in the arrow. A SINGULAR form analysed as carrying the [-PL] feature, which is mapped onto 'one'.

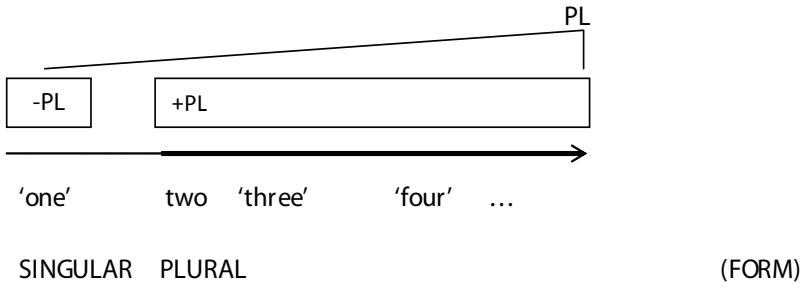


FIGURE 6. The semantic space of NUM in English

Marori, however, has a three-way NUM system: SG, PL and DUAL. Given the fact that both the SG and PL have their marking on the verb in Marori, we can say that Marori activates both SG and PL features in its system. Unlike in English where SINGLUAR is [-PL], Marori's SINGULAR is indeed associated with the SG feature. Thus, the semantic space of NUM in Marori can be represented in figure 7. Both SG and PL features with their respective binary values are present in Marori and mapped onto the NUM space.

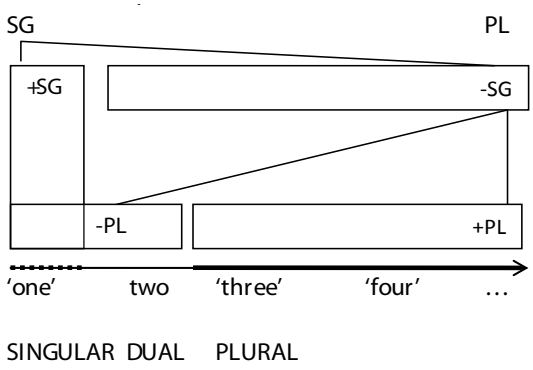


FIGURE 7. The semantic space of NUM in Marori

The NUM space of Marori needs some explanation. The features are hierarchically structured. The +PL feature is embedded in the [-SG] feature value. This is to capture the intuition that the first high-order distinction is between SG vs. NonSG. We have seen that there is morphological evidence for this in Marori: the distinction is expressed by two distinct morphemes (with complexity in the verbal morphology as seen in table 1). There is also evidence from other languages, e.g. Larike (discussed in section 4 below) where PL is understood in its broad sense as 'non-singular' (i.e. also covering 'dual'), in addition to

its narrow sense of 'three or more'.

The analysis as shown in figure 7 allows us to account for facts about distributed exponents in NUM expressions in Marori (and also in other languages). We can straightforwardly account for the economical way of encoding of DUAL in Marori by means of combining the available NonSG ([-SG]) and NonPL ([-PL]) morphemes.

However, it should be noted that the three-way number distinction in Marori can also be encoded by distinct forms, typically for the first person in the present/future tense. These forms are often associated with partially irregular lexically-determined paradigmatic patterns, e.g. *-du*, *-den*, *-men*, and *-ru*, *-ren* and *-men* for the first person singular, dual and plural categories in present and future tenses as shown in (17)(a-b). For the past tense as seen in (17c), *-men* is used for the first person irrespective of the number, in contrast to the second and third persons.

(17) a. The paradigm of the (auxiliary) verb 'be.at/sit' in the present tense in Marori

	1	2	3
Singular	*kuyedu*	*kami*	*kuye*
Dual	*kuyeden*	*kanermi*	*kuye*
Plural	*minggemen*	*kaminenggem*	*minggri*

b. The paradigm of the (auxiliary) verb 'be.at/sit' in the future tense in Marori

	1	2	3
Singular	*miru*	*kami*	*mi*
Dual	*miren*	*kanermi*	*mi*
Plural	*minggemen*	*kaminenggem*	*minggem*

c. The paradigm of the (auxiliary) verb 'be.at/sit' in the past tense in Marori

	1	2	3
Singular	*kuyemen*	*kuyem*	*kuyem*
Dual	*kuyemen*	*norowem*	*kuyem*
Plural	*mingrimen*	*minenggrim*	*minggrim*

While the dual forms *-den* as in *kuyeden* and *-ren* as in *miren* encode number, they are actually portmanteau morphemes that also encode specific person and tense (i.e. present/future) information. Therefore, they are in a sense not really dedicated DUAL number morphemes. Given the overall system of number in Marori, we can still maintain the analysis that there is no need to have a DUAL feature in this language. DUAL morphemes such as *-den*, while glossed as DU(AL) for simplicity can be specified as carrying [-SG, -PL] as part of feature bundles [1, -SG, -PL, PRES] features; i.e., meaning a first person dual present tense morpheme. The analysis accounts for the fact that the auxiliary it is affixed to (e.g., *kuyeden*) can enter into subject agreement with *nie* '1NonSG' as in (18) because the subject carries the same feature value [-SG] with which it can unify. The mechanism of unification is further discussed in section 3.3.3 below.

(18) *Nie* *purfam* *Jayapura* *di* *kuye-den.*
1NonSG person Jayapura soon be.at-1DU.PRES
'we (two) are in Jayapura soon.'

To conclude, there is good evidence to support the analysis that the three-way NUM system in Maori has two basic NUM features, SG and PL, with binary (+/-) values. The feature bundles in Marori are represented in figure 8.

SINGULAR	DUAL	PLURAL	(NUMBER CATEGORIES)
[SG +] [PL-]	[SG -] [PL -]	[PL +] [SG -]	(FEATURE BUNDLES)
SG	NSG&NPL	PL	(MORPHOLOGICAL CODING)

FIGURE 8. NUM system in Marori

3.3.3. Distributed NUM exponence across morphology and syntax. Having discussed the feature structure, we are now ready to discuss the issue of distributed NUM exponence further, in a principled and precise way. This can be straightforwardly done within the unification-based model of grammar as described in 3.1. In what follows we discuss and exemplify how DUAL is arrived at in morphology and syntax.

Consider example (19a) where the Actor 'you' is DUAL. Its NonSG exponents come from syntax (the free pronoun *kie* '2NonSG') and morphology (the affixes *n-* and *-m* in the verb). These morphemes carry the NUM feature with compatible values which then unify to form DUAL. The unification is shown in (19b).

(19) a. *Kie* *tamba* *Maria-na* *bosik* *eyew* *nadam.*
2NonSG PERF M-for pig search Ø-n-nda-m
3-2NonSG-AUX-2/3.NonPL.PST
'You (DU) searched for a pig for Maria.'

b. [NUM [SG -]] U [NUM [SG -]] U [NUM [PL -]] = [NUM [SG - / PL -]]
kie *n-* *-m*

The formation of DUAL can take place in the lexicon and syntax. In (19), it is formed by the unification of *n-* and *-m* when the verb *nadam* is created. When the verb *nadam* combines in syntax with the free pronoun *kie*, the NUM information from these units further unifies. The verb *nadam* is not acceptable if the actor is singular *ka* '2SG', for which the verb *ndam* must be used as seen in (19c). The unification fails because *ka* carries [NUM [SG +]] feature which is incompatible with that carried by *n- -m* (19d).

c. Ka Maria=na bosik eyew ndam / * nadam.
2SG Maria=for pig see Ø-nda-m
3-AUX-2/3.NonPL.PST
'You (SG) hunted a pig for Maria.'

d. * [NUM [SG +]] U [NUM [SG -]] U [NUM [PL -]]= [NUM [SG - / PL -]]
ka *n-* *-m*

In (20), DUAL is formed in syntax, not in morphology. Unlike in *nadam* (19), there

is no NonSG affix in the verb morphology of *ndam* to make DUAL. The verb *ndam* emerges from the lexicon searching out its Actor and Undergoer arguments, and the NUM information ([PL -]) from the actor suffix *-m* unifies with the NUM information from the free pronoun *emnde*, giving rise to DUAL. The unification is shown in (20b).

(20) a. *Emnde* *na-n* *bosik* *eyew* *ndam.*
3NonSG 1SG-for pig search Ø-nda-m
3-AUX- 2/3.NonPL.PST
'They (two) searched a pig for me.'

b. [NUM [SG -]] U [NUM [PL -]] = $\begin{bmatrix} \text{NUM} & \begin{bmatrix} \text{SG} & - \\ \text{PL} & - \end{bmatrix} \end{bmatrix}$

emnde- [*-m*]$_{\text{VERB}}$

The unification-based feature analysis of Agreement presented above shows the following points. First, it allows us to maintain the traditional distinction of morphology and syntax and at the same time also to capture the projection of morphology to syntax whereby referential information (in this case NUM values) can pass up across the boundary of morphology and syntax. Second, (NUM) agreement is essentially feature value compatibility, which operates on the basis of the same principle irrespective of whether it takes place in a clause or a word level. Third, with features being mapped on the semantic space of NUM, we can also capture the fact that agreement is more than simply compatibility of features. Given the NUM space of figure 7, the agreement is functional because when [SG -] and [PL -] combine, they narrow down to select the NUM space of DUAL.

4. Typological notes. A typological space of NUM is proposed in subsection 3.3.2. PL and SG are the basic NUM features with binary values (+/-). Languages vary with respect to whether one or both of them are activated. It has been argued that the feature structure is hierarchical with +PL being embedded in [-SG] and that DUAL is negatively defined as [-SG, -PL].

The question now is whether the proposed analysis of DUAL in Marori can be extended to account for DUAL in other languages, possibly in those with richer NUM distinctions (e.g. trial or paucal). Discussing these in depth across languages is beyond the scope of the present paper. However, in what follows, I discuss DUAL in three other languages: Nen (Papuan), Hopi (Uto-Aztecan, US) and Larike (Austronesian, Maluku-Indonesia).

4.1. DUAL in Nen. DUAL in Nen (Evans 2009, this volume) is unmarked; Non-DUAL (ND) is marked, e.g. *owab* 'talk (of two) → *owab-ta* 'talk (of one, or three or more').

DUAL in Nen, however, may also be marked, e.g. *aka-w* 'see-DU' vs. *aka-ta* 'see-ND. Unlike in Marori, the DUAL vs. non-DUAL marking is systematic in Nen. Nen arguably activates DUAL as a relevant NUM feature in its grammar. Importantly, there is no specific morphology for PL in Nen: it is expressed by means of a compositional strategy making use of the available (underspecified) NUM markers. This is further discussed below.

Like Marori, Nen also shows distributed exponence for the formation of specific NUM

reference. Absolutive free pronouns show no NUM distinctions, but PERS distinction only. Bound inflectional affixes on the verb show the SG vs. NonSG distinctions. Thus, in the following examples, the specific reference of the first person singular (21a), first person plural (21b), and first person dual (21c) is determined by the combination of the first free pronoun *ynd* (unspecified for NUM) and the agreement morphology on the verb (which supplies NUM information):

(21) a. *tog-am* *ynd* *w-aka-t-e* (Evans 2009)
child-ERG 1ABS 1.SG.U-see-NonDU-3sgA
'The child sees me.'

b. *tog-am* *ynd* *yn-aka-t-e*
child-ERG 1ABS 1.NonSG.U-see-NonDU-3sgA
'The child sees us (3 or more).'

c. *ämbs* *är-äm* *ynd* *yn-akae-w-ng*
one man-ERG 1ABS 1.NonSG.U-see-DU-~1.SG.A>DU.O
'One man sees the two of us.'

On the basis of the available evidence, I propose that the structure of the semantic space of NUM and related features in Nen is shown in figure 9.

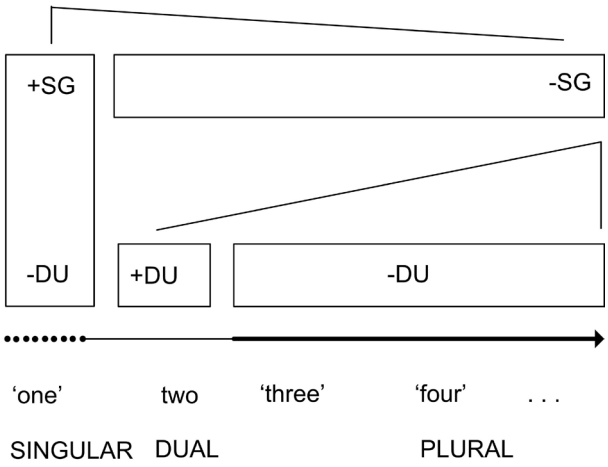


FIGURE 9. The semantic space of NUM in Nen

The analysis as depicted in figure 9 has the following advantages. First, Nen activates the feature DUAL in its NUM system. Crucially, the DUAL feature is structured as part of [-SG]. The space of DUAL 'two' is referable by means of [+DU].

Note that both negative (underspecified) and positive number value number may be associated with overt morphemes in Nen; e.g., the non-dual morpheme *-t* as in *w-aka-t-e* '1sgU-see' (21a). In other words, we adopt an analysis where a negative value of number does not necessarily mean that it is morphologically unmarked. Conversely, a

morphologically simple form may carry a positive value of DUAL number feature lexically by default as is the case with *owab* 'talk (of two).

The analysis of embedding [+DU] under the [- SG] feature in the structured semantic space finds its empirical support from the distributed exponence featuring this language. The presence of NonSG and DUAL morphemes to express DUAL as seen in (22) in the presence of a DUAL morpheme seems to be redundant at first. However, given the distributed exponence (where two exponents are needed to express DUAL), the two exponents are naturally those associated with the DUAL space, namely [+DU] and [-SG]. The space of DUAL is a specific portion of NUM space of [-SG].

(22) *mnḡ* *yä-trom-aran*.
house 3.NonSG.U-be.erected-STAT:DU
'Two houses are standing.'

Second, the space of [-DU] is split. This is consistent with the meaning of Non-DUAL (ND) and the distribution of ND morpheme in this language. For example the ND morpheme is expected to be used for plural reference. This is indeed the case; cf. [-SG] (NonSG) which combines with [-DU] (ND) in example (21b).

Third, in our analysis the spaces of two categories may overlap, e.g. the spaces of [+SG] and [-DU] in Nen. The conception of overlapping spaces is in fact significant for specific NUM reference and coding. Thus, the coding of a SG referent in Nen makes use of the exponents signifying SG and NON-DUAL as seen in (23). This is expected on the proposed structured semantic space of NUM in distributed morphology.

(23) *mnḡ* *y-trom-ngr*.
house 3.SG.U-be.erected-STAT:NonDU
'A house is standing.'

Fourth, the analysis with conception of NUM involving overlapping spaces provides a natural account for what is otherwise a peculiar strategy of coding plural and exhaustive plural/paucal in Nen.

As seen in figure 9, the space associated with Non-DUAL ([-DU]) is split into two; one overlaps with the space of [+SG] and the other with the space of [-SG]. In the latter case, it is equivalent to the space of plural (i.e. 'three or more'). In other words, the space of plural is the portion of the space of [-SG] that is Non-DUAL ([-DU). Since both [-SG] and [-DU] have their respective coding morphemes, it is not surprising that Nen does not need a special marker for plural. Both NonSG and NonDU morphemes are usable to encode plural, as exemplified in (24) below. Their use meets the language-specific requirement of distributive exponents in expressing NUM in this language.

(24) *mnḡ* *yä-trom-ngr*.
house 3.NonSG.U-be.erected-STAT:NonDU
'Three or more house(s) are standing.'

Finally, the expression of what Evans (2009) calls the exhaustive set/universal 'all', which also appears to be unusual at first, but is in fact a natural way of expressing NUM in the proposed analysis. Exhaustive set is expressed by means of singular morphology in combination with dual morphology, as seen in (25). This might be equivalent to 'paucal' in other languages.

The expression of exhaustive set/all must refer to the space that is complementary to the space of the non-exhaustive plural expressed by the combination of non-singular and non-dual morphemes as exemplified in (24). As seen in Figure 9 the space of the non-exhaustive plural is in the right-end of NUM space. The space for the exhaustive plural (or paucal) is logically the one in the left, including that of [+SG].

Again, due to the distributed exponence requirement of NUM expression in Nen, Nen needs no special morpheme to encode the exhaustive/paucal NUM because there are resources already available for this, namely the morphemes signifying [+SG] and [+DU], as seen in (25). However, we have to note the fact that the combination of these morphemes is not compositional: the meaning has been 'lexicalised' as 'exhaustive plural' in contrast to 'unlimited or general plural' expressed by the combination of non-singular and non-dual morphemes in Nen.

(25) *mnḡ* *y-trom-aran.*
house 3.SG.U-be.erected-STAT:DU
'All the houses are standing.'

4.2. DUAL IN HOPI. The proposed analysis for Marori and Nen can be applied to account for Hopi data. In Hopi (Hale 1997, Corbett 2000:169), the combination of SG and PL morphemes give rise to DUAL interpretation as seen in (26c). Corbett explains cases exemplified in (26c) as 'constructed' numbers: dual is constructed from the number on the pronoun and that on the verb.

(26) a. *Pam* *wari.*
that.SG run.PERFV.SG
'He/she ran.'

b. *Puma* *yùutu.*
that.PL run.PERFV.PL
'They (plural) ran.'

c. *Puma* *wari.*
that.PL run.PERFV.SG
'They (two) ran.'

The analysis of Hopi agreement in this paper is in the same spirit as the analysis suggested by Hale (1997). He suggested that DUAL interpretation could be achieved via intersection of the two binary oppositions ([+/- SG], [+/-PL]). However, the precise detail of Hale's analysis as to how the 'intersection' exactly works remains unclear. In this paper we present the analysis as a system of feature mapping onto a structured semantic space of NUM as shown in figure 10. The relevant NUM features are processed in the same way as other grammatical features in the grammar. The unification of NUM features is expected to be constrained and/or functionally motivated by the possible reference to the semantic space of NUM.

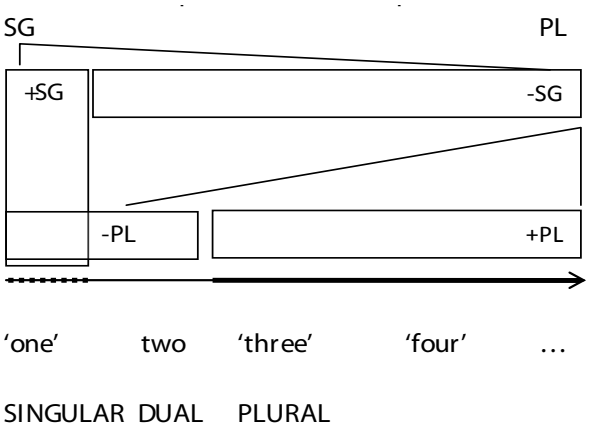


FIGURE 10. The semantic space of NUM in Hopi

Let me be specific about how DUAL in Hopi can be arrived at. It is essentially in the same way as that in Marori, but with some constraints due to the mapping onto the semantic space of NUM. First we have to specify how the NUM feature is carried by the relevant morphemes in Hopi. It should be noted that Hopi is unlike Marori in that it has no dedicated underspecified NonSG/NonPL morphemes, i.e. those carrying [-SG]/[-PL] features. Pronominal or verbal morphemes in Hopi glossed as SG/PL can be analysed as carrying feature bundles with their values as shown in figure 11.

a. PL morpheme b. SG morpheme

$$\left\{\begin{matrix}[PL\ +] \\ [SG\ -]\end{matrix}\right\} \qquad \left\{\begin{matrix}[SG\ +] \\ [PL\ -]\end{matrix}\right\}$$

FIGURE 11

The second point of the analysis is the implication the mapping of the NUM feature onto the semantic space of NUM has in relation to the combinatory property of the grammar.

The mapping can be wide or narrow. Given the structured space in figure 10, the mapping of PLURAL morpheme (figure 11a), for example, may be associated with the space corresponding to [-SG] ('wide', including 'two'). That is, 'plural' means 'more than one'. Alternatively, the PLURAL morpheme means 'more than two'; i.e. referring to the space corresponding to [+PL] ('narrow', excluding 'two'). Likewise, the SINGULAR morpheme (figure 11b) can refer to the exact narrow space of 'one' due to its [+SG] feature, or alternatively to a wider space including 'two' due to its [-PL].

In unification-based grammar, nothing theoretically prevents the unification of [+PL] carried by the plural form and [+SG] carried by the singular form because each carries

different attributes with respective values. However, given the conception that each NUM feature is mapped onto a portion of semantic space of figure 10, the interpretation of the combinatory possibility of morphemes with [+SG] with that of [+PL] is constrained. In one interpretation, when the 'narrow' space is referred to, the two do not refer to a common NUM space. In the other interpretation where the wide spaces are referred to (i.e. both features [-PL] and [-SG] carried by singular and plural forms refer to the common space that includes 'two'), then the DUAL interpretation is arrived at. The condition of the unification to arrive at DUAL interpretation in Hopi can be shown in figure 12.[17]

Condition: Given the NUM space of Hopi in figure 10, the space of [+SG] is mutually exclusive with that of [+PL] (*i.e. [+SG]U[+PL])
Hence:

PL		U	SG		DU	
[PL	+]		[SG	+]	= [SG	-]
[SG	-]		[PL	-]	[PL	-]

FIGURE 12

In short, underspecified NUM features carried by SG/PL morphemes allow for wide NUM space referents. These serve as resources for combinatory purposes to refer to a specific NUM referent such as DUAL. Thus, languages such as Hopi do not need to a have a dedicated morpheme for DUAL, as SG/PL forms are usable for this.

4.3. DUAL IN LARIKE. Larike, an Austronesian language of Maluku (Laidig & Laidig 1990), is reported to have a four-way NUM system (SG, DU, TRIAL and PL). The full sets of the four-way NUM distinction are only encountered in first, second and third person human pronominal forms. The inflection for the third person non-human is defective. The subject and object sets are shown in table 2 and table 3 respectively. (Non-pronominal forms are not inflected for NUM.)

		SG	DUAL	TRIAL	PLURAL
1	EX	*au-*	*aruai-*	*aridu-*	*ami-*
	INC		*ituai-*	*itidu-*	*ite-*
2		*ai-*	*iruai-*	*iridu-*	*imi-*
3	HUM	*mei*	*matuai-*	*matidu-*	*mati*
	NHUM	*i-*	-	-	*iri-*

TABLE 2. Subject prefixes

[17] While we have unification of features in figure 12, the interpretation of {[SG-][PL-]} as dual is actually associated with the notion of intersection in the semantic space of number.

		SG	DUAL	TRIAL	PLURAL
1	EX	*-a/u*	*-arua*	*-aridu*	*-ami*
	INC		*-itua*	*-itidu*	*-ite*
2		*-ne*	*-irua*	*-iridu*	*-imi*
3	HUM	*-ma*	*-matua*	*-matidu*	*-mati*
	NHUM	*-a*	-	-	*-ri*

TABLE 3. Object suffixes

The NUM space in Larike can be represented in Figure 13 for the following reasons. First, DUAL and TRIAL are NUM features without [+/-] value, i.e. privative. Unlike in Nen, there is no evidence in Larike that the system makes use of the opposition of DUAL vs. non DUAL. Dual and trial in Larike were (historically) derived from numeral 'two' and 'three' respectively (Laidig & Laidig 1990). They are true dual and trial forms in the sense that they refer to exact quantities of 'two' and 'three', and never used to refer to vague notion of several as is a paucal or limited plural in other languages such as Yimas (Foley 1991, Corbett 2000).

It should be noted that the consequence of the analysis adopted in this paper is that we have a hybrid feature system. As seen from figure 13, the feature system in Larike consists of SG and PL with binary values as well as privative DUAL and TRIAL. This might not be preferable as the analysis shows a proliferation of features. However, it is not clear how any alternative analysis could be offered where DUAL and TRIAL are derived from more basic features or having binary values and where language-specific patterns (further described below) are also accounted for.

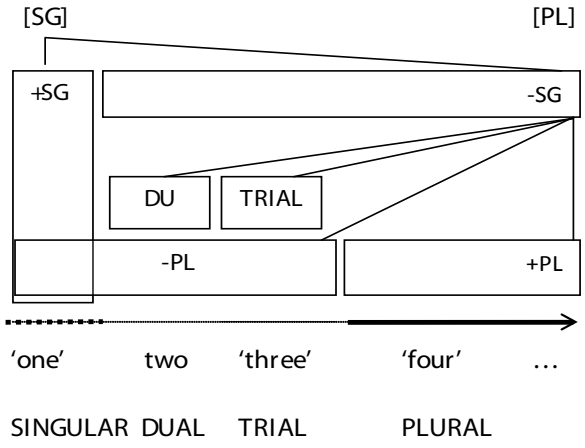


FIGURE 13. The Semantic space of NUM in Larike

Second, the plural forms in Larike may be also used when referring to quantities of two or three (Laidig & Laidig 1990). This is the evidence that the plural forms contain

[-SG, +PL] features, which allow the wide interpretation of PL. That is, its [-SG] feature allows the use of the PL form to cover the wide semantic space that includes 'two', 'three', and 'four or more'. In this case, the exact referent depends on the context. In short, the proposed analysis captures what Corbett (2000) calls the facultative nature of the PL form in Larike.

Finally, as seen from tables 2 and 3, there is a gap in the form for third person non-human. It is reported by Laidig & Laidig (1990) that it is common to use the singular non-human form to refer to a limited plural. This is expected on the analysis that a SG morpheme carries a feature bundle of [+SG] and [-PL] and that each feature can operate independently. Thus, the [-PL] feature of the SG form has its own mapping onto the semantic space of NUM that is complementary to that of [+PL]. This complementary space is in a sense 'plural' because it covers the spaces of 'two' and 'three'. It is however limited as it is contrasted with the space of PL (+PL) which, as indicated in the diagram, has no upper limit point. In short, because of the absence of DUAL/TRIAL form for the third person non-human, the SG form is naturally extended to refer to this limited plural space because the SG form carries [-PL] feature in it.

5. Conclusions. This paper has provided an explicit analysis of how NUM morphology is projected to syntax in Marori. It is proposed that NUM features be established on language-specific structured semantic space of NUM and that there are two basic NUM features, namely SG and PL, each with binary values. Each is possibly independently mapped onto the semantic space. It is argued that DUAL can be unmarked, analysed as [-SG,-PL]. However, DUAL can be also marked, expressed by a dedicated DUAL form. This is encountered in Nen and Larike.

It has also been demonstrated that the proposed analysis treats NUM morphemes as carrying a bundle of features, with each operating independently. This provides a natural explanation for what appears to be unusual NUM agreement or expressions as found in Nen and Hopi. The phenomena of facultative PLURAL as found in Larike can also be accounted for in the proposed analysis.

References

Arka, I Wayan. 2003. *Balinese morphosyntax: a lexical-functional approach.* Canberra: Pacific Linguistics.

Arka, I Wayan. 2005. Speech levels, social predicates, and pragmatic structure in Balinese: A lexical approach. *Pragmatics* 15(2/3). 169-203.

Boelaars, Jan H.M.C. 1950. *The linguistic position of south-western New Guinea*. Leiden: Brill.

Bresnan, Joan. 2001. *Lexical functional syntax*. London: Blackwell.

Bresnan, Joan & S. Mchombo. 1987. Topic, pronoun, and agreement in Chichewa. *Language* 63(4). 741-82.

Chomsky, Noam. 1981. *Lectures on Government and Binding Theory*. Dordrecht: Foris.

Corbett, Greville G. 2000. *Number*. Cambridge: Cambridge University Press.

Dalrymple, Mary. 2001. *Lexical Functional Grammar.* (Syntax and Semantics 34). San Diego: Academic Press.

Drabbe, Peter. 1954. *Talen en Dialecten van Zuid-West Nieuw-Guinea*. (Micro-Bibliotheca Anthropos 11). Freibourg: Anthropos-Institut.
Drabbe, Peter. 1955. *Spraakkunst van het Marind*. (Studia Instituti Anthropos 11). Mödling: Missiehuis St. Gabriël.
Evans, Nicholas. 2009. Complementarity, unification, and non-monotonicity: Bound pronominals, free NPs and argument status in a double-marking language (Nen). Unpublished ms.
Evans, Nicholas. 2010. *Dying words: Endangered languages and what they have to tell us*. Oxford: Wiley-Blackwell.
Falk, Yehuda N. 2001. *Lexical-Functional Grammar* (CSLI lecture notes 126). Stanford: CSLI.
Foley, William A. 1991. *The Yimas language of New Guinea*. Stanford: Stanford University Press.
Gebze, Wilhelmus & Mark Donohue. 1998. Kamus kecil bahasa Moraori. [Morori picture dictionary]. Unpublished ms.
Hale, Ken. 1997. Some observations on the contributions of local languages to linguistic Science. *Lingua* 100. 71-89.
Harley, Heidi & Elizabeth Ritter. 2002. Person and number in pronouns: A feature-geometric analysis. *Language* 73(3). 482-526.
Kathol, Andreas. 1999. Agreement and the syntax-morphology interface in HPSG. In Levine, Robert D. & Georgia M. Green (eds.), *Studies in contemporary phrase structure grammar*, 209-260. Cambridge: Cambridge University Press.
Klamer, Marian. 1996. Kambera has no passive. *NUSA: Linguistic studies of Indonesian and other languages in Indonesia* 39. 12-30.
Laidig, Wyn D & Carol J Laidig. 1990. Larike pronouns: Duals and trials in a Central Moluccan language. *Oceanic Linguistics* 29. 87-109.
Pollard, Carl & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. California: CSLI Publications, Stanford & University of Chicago Press.
Ross, Malcolm D. 2005. Pronouns as a preliminary diagnostic for grouping Papuan languages. In Andrew Pawley, Robert Attenborough, Jack Golson & Robin Hide (eds.), *Papuan pasts: Cultural, linguistic and biological histories of Papuan-speaking peoples*, 15-65. Canberra: Pacific Linguistics.
Sag, Ivan A, Thomas Wasow & Emely M. Bender. 2003. *Syntactic theory: A formal introduction*, 2nd edn. Stanford: CSLI.
Sohn, Myo-Sook, Randy Lebold & Ron Kriens. 2009. Report on the Merauke Subdistrict Survey, Papua, Indonesia (SIL Electronic Survey Report 2009-018). http://www.sil.org/silesr/2009/silesr2009-018.pdf.
Wechsler, Stephen & Larisa Zlatic. 2000. A Theory of agreement and its application to Serbo-Croatian. *Language* 76(4). 799-832.
Wurm, Stephen. 1954. P. Drabbe's study on the languages of south-west New Guinea. *Anthropos* 49. 299–304.

I Wayan Arka
wayan.arka@anu.edu.au

7

Language Documentation & Conservation Special Publication No. 5 (December 2012)
Melanesian Languages on the Edge of Asia: Challenges for the 21st Century,
ed. by Nicholas Evans and Marian Klamer, pp. 174–190
http://nflrc.hawaii.edu/ldc/sp05/
http://hdl.handle.net/10125/4564

'Realis' and 'irrealis' in Wogeo: A valid category?

Mats Exter

Heinrich-Heine-Universität Düsseldorf

Finite verb forms in Wogeo, an Austronesian language of New Guinea, are obligatorily marked with a portmanteau prefix denoting person and number of the subject on the one hand, and a grammatical category that is conventionally glossed in the literature as realis–irrealis, on the other. In similar languages, the latter category is usually described as modal, with a certain range of meanings which is, in many cases, only vaguely defined. A more in-depth investigation of the verbal system of Wogeo and the functional distribution of the respective categories shows, however, that the language is quite different from a postulated prototypical realis–irrealis language. Central attributes of the supposed realis–irrealis semantics are not realized by the obligatory prefixes but by other morphosyntactic means, while the prefixes are restricted to only a small part of the assumed realis–irrealis domain.

1. Introduction.[1] In the linguistic literature, 'realis' and 'irrealis' have most often been discussed under the more general heading of *mood* and *modality*. These in turn are terms which are almost universally used in linguistics (with or without difference in meaning), yet a satisfactory definition is largely a matter of ongoing debate. The problem with many existing definitions is that they are either too vague and leave too much to implicit assumptions, as is often the case in purely descriptive contexts; or, if they attempt to be explicit, they frequently resort to disjunctive characterizations, involving statements like

[1] I wish to thank the speakers of Wogeo, above all my main consultants, Conny Tarere, Michael Ganem and the late Albert Kulbobo, for welcoming me and sharing their knowledge of the language with me. I also thank Astrid Anderson for introducing me to the Wogeo world, and the Research Council of Norway as well as the Institute for Comparative Research in Human Culture, Oslo, Norway, for funding the fieldwork that this paper is based on. Thanks are also due to the participants at the Workshop on the Languages of Papua 2, February 8–12, 2010, Manokwari, Indonesia, for helpful discussions and feedback. Finally, I am very grateful to Johan van der Auwera, Marian Klamer, Daniel Kölligan, and an anonymous reviewer for their valuable comments on an earlier version of this paper.

"A is X *or* Y *or* Z." A full discussion of the terms mood and/or modality is well beyond the scope of this paper; however, a working definition is needed to investigate the issue of realis–irrealis in a meaningful way.

In the following section, therefore, such working definitions are discussed, and the position of realis–irrealis with respect to the category of mood (or modality) is discussed. Then, a brief review of proposed 'realis–irrealis' categories across languages is given and the comparability of those categories is discussed. Finally, an overview of the verbal morphosyntax of Wogeo is given and the usefulness of the *realis–irrealis* terminology is reassessed in the light of the evidence that can be gained from the Wogeo data.

2. Terminological Issues. As a first step, as observed by Cristofaro (2012), it is important to distinguish between the semantic (or conceptual) domain we are dealing with, on the one hand, and any grammatical categories that *realize* that domain, on the other. For the former, the term *modality* is often used, whereas the term *mood* is commonly reserved for the latter. The distinction between semantic domain and grammatical category will be taken as fundamental in the discussion that follows.

Palmer (2001:1) defines modality as being "concerned with the status of the proposition that describes the event." This is an example of what has been referred to above as a vague definition, since it is left implicit what exactly is meant by *concerned with* and, especially, the *status of the proposition – status* in relation to what? Somewhat more explicit is the definition given by Portner (2009:1), who suggests that "modality is the linguistic phenomenon whereby grammar allows one to say things about, or on the basis of, situations which need not be real." As Portner himself points out, it is not immediately obvious how to define the term *real*; yet, the definition is more useful in practice than Palmer's.

Further differences can be found in the ways in which different researchers subdivide the modal semantic domain. Givón (2001), e.g., views the division between *presuppositions* and *assertions* as primary; assertions are then divided into *realis* and *irrealis*; and realis assertions are classified as *positive* or *negative*. Palmer (2001), on the other hand, takes a more traditional position, distinguishing *propositional modality* (subdivided into *epistemic* vs. *evidential*) from *event modality* (subdivided into *deontic* vs. *dynamic*[2]). Finally, Bybee (1998) distinguishes four subdomains: *agent-oriented*, *speaker-oriented*, *epistemic*[3] and *subordinating* modality. The most striking way in which Bybee's approach differs from the former two, though, is that she argues that the supposed subdomains of modality are really four independent semantic domains, the connection between which is mainly diachronic, not synchronic. The subdivisions within the domain of modality that Givón, Palmer and Bybee propose are summarized in table 1.

[2] In Palmer's terminology, dynamic modality subsumes ability and willingness.

[3] Agent-oriented modality (in Bybee's terms) includes, but need not be restricted to: obligation, permission, volition, ability; speaker-oriented: imperative, permissive; epistemic: uncertainty, possibility, probability.

Givón	Palmer	Bybee
Presupposition	Propositional modality:	Agent-oriented modality
Realis assertion:	Epistemic	Speaker-oriented modality
Positive	Evidential	Epistemic modality
Negative	Event modality:	Subordinating modality
Irrealis assertion	Deontic	
	Dynamic	

TABLE 1. Subdivisions of modality according to Givón (2001), Palmer (2001) and Bybee (1998)

A different approach is taken by van der Auwera & Plungian (1998). They choose to restrict the use of the term *modal* to those categories whose functions can be described by reference to the concepts of *possibility* and *necessity*, explicitly excluding categories like volition, evidentiality, etc., from the realm of modality. The classification of van der Auwera & Plungian is summarized in table 2.

<table>
<tr><td colspan="4">Possibility</td></tr>
<tr><td colspan="3">Non-epistemic</td><td rowspan="3">Epistemic</td></tr>
<tr><td rowspan="2">Participant-internal</td><td colspan="2">Participant-external</td></tr>
<tr><td>Non-deontic</td><td>Deontic</td></tr>
<tr><td rowspan="2">Participant-internal</td><td>Non-deontic</td><td>Deontic</td><td rowspan="3">Epistemic</td></tr>
<tr><td colspan="2">Participant-external</td></tr>
<tr><td colspan="3">Non-epistemic</td></tr>
<tr><td colspan="4">Necessity</td></tr>
</table>

TABLE 2. Subdivisions of modality according to van der Auwera & Plungian (1998)

Obviously, Givón, Palmer, Bybee and van der Auwera & Plungian subdivide the semantic domain of modality on the basis of different criteria. These should, therefore, be seen as complementary approaches which can very well be applied independently to arrive at cross-cutting classifications. The question that poses itself is, then, which of the strategies (if any) is (or are) most fruitful in solving the realis–irrealis issue we are currently concerned with. For reasons which will become clear in sections 4 and 5 below, I will adopt the restrictive approach of van der Auwera & Plungian (1998) as a working hypothesis for the domain of modality.

As will become clear in section 3, the semantic domain that a putative realis–irrealis domain has been claimed to subsume overlaps to a large degree with what different authors assume to be within the realm of modality, plus other areas that would not traditionally be

viewed as modal, such as, e.g., future (tense) or habitual (aspect). It is therefore instructive as a starting point to look at different proposals as to what realis–irrealis really is. Mauri & Sansò (2012) provide a very good overview of the current debate. The main positions that are relevant to the present discussion can, according to them, be summarized as follows:

1. Irrealis is a kind of 'mega-modality' subsuming a number of modal subdomains.
2. Realis–irrealis is the same as modality.
3. Realis and irrealis are themselves modal categories.
4. Realis and irrealis are the values of a category 'reality status' which is independent of modality.

If the last position, advocated e.g. by Elliott (2000), is correct, it should be possible to identify the semantic content that is expressed by such a category. Pietrandrea (2012:186) argues in a top-down approach in favor of a category of 'reality status' as distinct from modality. For her, irrealis states of affairs are *non-actualized*, meaning they are "presented as not grounded in perceivable reality."

The task of identifying the meaning expressed by 'reality status' is taken up in a very different way by de Haan (2012). In his bottom-up typological study, he sets out to investigate the claim that there is a prototypical semantic core that can be assigned to those cases that have been analyzed as instances of realis–irrealis. His conclusion, however, is negative: Many alternative core meanings can be found, none of which can convincingly be argued to have priority over the others. Thus, it is completely open what should be the core and what should be the periphery of the category 'reality status'. Therefore, de Haan argues, it cannot at present be shown to be a typologically valid category.

3. Previous Typological Studies. Having been sensitized to the complexity of the issues involving modality and reality status as well as the relationship between the two, we are now in a position to give a concise overview of previous typological studies relating to the issue of the elusive 'realis–irrealis' category in various languages, language families and geographical areas. We will focus on three studies: Bugenhagen (1993), Elliott (2000) and van Gijn & Gipper (2009).

Bugenhagen's (1993) paper is particularly interesting in the present context because it investigates the semantics of what is called 'irrealis' in seven Austronesian languages of New Guinea. The languages in his sample are therefore both genetically and geographically comparable to Wogeo.[4] On the basis of his database, he identifies what can be described as a *prototypical semantic core* for the realis and irrealis categories (for the given language family and area): prototypical realis semantics is associated with positive polarity, non-future tense, perfective aspect and declarative speech acts,[5] while irrealis semantics is associated with future tense, hypothetical conditional clauses, counterfactual conditional

[4] In Bugenhagen (1993), as almost everywhere else (including this paper), *irrealis* is taken to be the category in need of explanation, with *realis* left as the unmarked member of the dichotomy. The relationship between the two terms is thus fundamentally *asymmetrical*.

[5] A slightly different core meaning for realis is assumed by van der Auwera & Devos (2012:172), namely a "main clause affirmative declarative referring to the present time sphere".

clauses, complements of 'want', and negative purpose clauses ('lest'). Bugenhagen's prototypical uses of realis and irrealis are summarized in table 3.

Realis	Irrealis
Positive polarity	Future tense
Non-future tense	Hypothetical conditional clauses
Perfective aspect	Counterfactual conditional clauses
Declarative speech acts	Complements of 'want'
	Negative purpose clauses ('lest')

TABLE 3. Prototypical uses of realis and irrealis in Austronesian languages of New Guinea according to Bugenhagen (1993)

The characterization of the (supposed) irrealis semantic domain by means of a number of notions reminds us of Bybee's (1998) view of the domain as a set of notions linked by partial similarities (family resemblances) as discussed above. This view is augmented by Bugenhagen (again, for his data set only) by explicitly postulating a semantic *focal area* within the broader domain where the languages are largely in agreement, and more peripheral areas where individual languages show specific patterns. (Looking at Bugenhagen's list, one would have to state more precisely that it represents *several interconnected focal areas* rather than one, as proposed by de Haan 2012.) Bugenhagen explicitly points out, however, that despite the relatedness and close proximity of the languages, "no two of them exhibit a completely identical range of uses for their irrealis forms" (1993:35). We shall see below whether Wogeo fits Bugenhagen's generalizations.

Elliott, too, investigates a number of languages with an alleged realis–irrealis distinction, with the aim to "arrive inductively at a typological description of this category" (2000:56). The number of languages included in her database (16) is slightly larger than the number of languages investigated by Bugenhagen, and she uses a different sampling strategy, with languages drawn from widely different families and geographical areas.

Elliott arrives at a result which is completely different from Bugenhagen's (1993): She argues for a grammatical category *reality status* (the term originating in Whorf 1938) with the values *realis* and *irrealis*, and she claims that it is in fact possible to identify a common semantic component in all uses of the category. For Elliott, the common semantic core of irrealis is that "irrealis events or states are perceived as being located in an alternative hypothetical or imagined world, but not the real world" (2000:81). The semantic area thus covered by 'irrealis' is, however, extremely broad and includes potential events, conditionals, events qualified by modality, and commands; additionally, negations, habituals, and interrogatives may also be subsumed by 'irrealis' (2000:70).

I see two problems in Elliott's approach: First, the distinction (if any) between modality on the one hand and her 'reality status' on the other is not defined systematically; and second, the large cross-linguistic differences in the semantics of 'irrealis' are left unexplained.

Van Gijn & Gipper (2009) use a third approach, providing an in-depth analysis of the

realis–irrealis system of a single language (Yurakaré, an unclassified South American language) and comparing it to six other languages from different families and areas. They arrive at the conclusion that the semantic domain underlying the alleged realis–irrealis distinction is best described not in binary terms, but in terms of a *continuum* (from *counterfactual* via *possible* to *factual*) – with the endpoints typically marked by irrealis on the one hand and realis on the other hand, and a 'grey area' in between – which languages divide in specific ways. Particular areas on the continuum are then again subdivided: possible events into events *with* and *without speaker commitment*, and factual events into *temporal* and *atemporal* events. These findings are then expressed in terms of an *implicational hierarchy* (2009:176; SC = 'speaker commitment'; TS = 'temporally specific'):

counterfactual < possible [−SC] < possible [+SC] < factual [−TS] < factual [+TS]

Van Gijn & Gipper thus introduce the idea of an empirically based implicational hierarchy (and subhierarchies) into the discussion. Unfortunately, however, as we will see below, Wogeo constitutes a clear counterexample to the generalization expressed in that hierarchy. It seems likely that the data base that van Gijn & Gipper base their proposal on is much too small to adequately capture a phenomenon as complex as the one under discussion here.

In my view, what van Gijn & Gipper's (2009) approach does not adequately explain is the fundamental asymmetry between the alleged endpoints of the continuum (on the one hand, 'realis' as a cross-linguistically fairly well-defined category covering a rather narrow semantic area; and on the other hand, 'irrealis' as an extremely wide, vague, and fuzzy category with large cross-linguistic variation and no clearly discernible semantic core). Moreover, 'factuality' is usually (if not always) not the *only* semantic component of the relevant grammatical categories; therefore, the supposed continuum may be better described as the result of cross-classification by different independent categories.

4. Realis and Irrealis in Wogeo. We will now turn to Wogeo and the formal and semantic properties of its 'realis–irrealis' morphological category. Wogeo is an Austronesian language spoken by at most (and probably less than) 1600 people on Vokeo and Koil Islands off the north coast of New Guinea. Previous anthropological studies on Wogeo include Hogbin (1970, 1978) and Anderson (2011). Exter (2003) is an analysis of the phonology of the language, and Anderson & Exter (2005) is a collection of traditional Wogeo texts for the speech community as well as a mainly anthropological academic audience. Exter (2012), still work in progress, is intended to be a comprehensive grammatical description. The data presented here are based on my own fieldwork, conducted in 1999 and 2000.

Finite verbs in Wogeo (i.e. all verb forms except verbal nouns / gerunds) are marked with an obligatory portmanteau prefix that denotes the person and number of the subject as well as realis or irrealis.[6] That means that none of the values of the dichotomous realis–irrealis category is formally unmarked in Wogeo. It also means that every sentence with a

[6] Imperative and prohibitive forms are the only exceptions to this generalization (see below). – To facilitate the discussion below, I will continue to use the terms *realis* and *irrealis* for the time being.

verbal predicate in Wogeo is marked either as realis or as irrealis; there are no unmarked sentences (and by the same token, no unmarked events). No other part of the verb in Wogeo (apart from the stem) is formally obligatory. Thus, it is fair to say that in all respects the Wogeo verbal system is organized around the realis–irrealis category.

As can be seen from the template in table 4, slots −6 and −5 (optional) and slot −4 (obligatory) all contain information related to tense, aspect, and/or mood: Slot −6 contains the *counterfactual* prefix; slot −5 contains the *future*, *tentative*, *proximal imperfective* and *distal imperfective* prefixes; and slot −4 contains the *person/number/realis–irrealis* portmanteau prefixes.[7]

CNTF	TAM	**PNM**	INCH	CAUS	IPFV (RDP)	**Stem**	IPFV (RDP)	DIR	APPL	P	N	BEN	P	N
−6	−5	**−4**	−3	−2	−1	**0**	1	2	3	4	5	6	7	8

TABLE 4. Schematic morphological structure of the verb in Wogeo (obligatory slots are bold; slots that show higher internal coherence are shaded grey)

Table 5 gives an overview over the PNM prefixes (slot −4) in Wogeo. As can be seen, there are four number categories (singular, plural, dual, paucal); tildes indicate synonymous forms. Inspection of the paradigm immediately shows that it is quite 'messy': There are many homonymous forms (e.g. 1PL.RLS and 1PL.IRR, 1DU.RLS and 1PAU.RLS) and partly homonymous forms (e.g. 2SG.RLS and 2SG.IRR) without a clearly discernible pattern (although conspicuously, the distinction between realis and irrealis is neutralized in the plural). Not surprisingly, corresponding realis and irrealis forms appear to be diachronically related; synchronically, however, the two categories cannot be reduced to a simpler analysis.

The table only shows the so-called *plain* realis–irrealis paradigm (i.e. with slots −6 and −5 remaining empty). If the complete PNM paradigms of all complex categories are taken into account, an extremely complex picture emerges, which includes multiple complicating factors such as vowel assimilation; idiosyncratic fusions, vowel changes, and vowel deletions; and even more complex patterns of synonymy and homonymy. For the point made in the present paper, therefore, this morphophonological and morphological complexity will not be dealt with further.

[7] Abbreviations used in this paper: A='aspect'; APPL='applicative'; BEN='benefactive'; CAUS='causative'; CNTF='counterfactual'; DIR='directional'; DIST='distal'; DU='dual'; FOC='focus'; FUT='future'; INCH='inchoative'; IPFV='imperfective'; IRR='irrealis'; M='mood'; N='number'; NEG='negative'; NMLZ='nominalizer'; P='person'; PAU='paucal'; PL='plural'; PROH='prohibitive'; PROX='proximal'; RDP='reduplication'; RECP='reciprocal'; RLS='realis'; SG='singular'; T='tense'; TENT='tentative'; TOP='topic'.

Person/number	(Plain) realis	(Plain) irrealis
1SG	*o-lako*	*go-lako*
2SG	*go-lako ~ ko-lako*	*go-lako*
3SG	*e-lako*	*de-lako*
1PL	*ta-lako*	*ta-lako*
2PL	*ka-lako*	*ka-lako*
3PL	*da-lako*	*da-lako*
1DU	*to-lako ~ te-lako*	*tog-lako ~ teg-lako*
2DU	*kad-lako ~ kod-lako*	*kad-lako ~ kod-lako*
3DU	*do-lako ~ de-lako*	*dog-lako ~ deg-lako*
1PAU	*to-lako ~ te-lako*	*tog-lako ~ teg-lako*
2PAU	*koto-lako*	*koto-lako*
3PAU	*doto-lako*	*doto-lako*

TABLE 5. The PNM prefixes in (plain) realis and irrealis forms of Wogeo *lako* 'go'

Slots other than −6, −5, and −4 in table 4 (namely slots −3, −1, and 1) contain TAM-related information, too, but it is argued here that the aforementioned slots (i.e. slots −6, −5, and −4) form a unit of their own. Formally, they are a unit because they display morphological idiosyncrasies between each other, such as fusion, vowel assimilation, and a number of other irregularities. Functionally, they are a unit in showing a number of combinatory interdependences (obligatory, optional, and excluded combinations). The same does not apply to the other slots, where agglutination and a large degree of combinability predominate. The resulting combinations of slots −6, −5 and −4 form complex TAM categories[8] which are given convenient summary labels (which I will call *complex-category* labels) here. Those TAM combinations that are well-formed, along with their complex-category labels, are shown in table 6. Where more than one form is given for any complex category, those forms are synonymous.[9]

[8] 'Complex' should here be taken to mean formally, not semantically, complex.

[9] Note that the so-called *tentative* forms express the meaning 'to try it with X-ing' (or 'to X and see what happens'), not 'to try to X'. – As will become obvious from a closer inspection of table 6, the *tentative* and *counterfactual* markers are homonymous. Two lines of argument are put forward here to justify their analysis as different morphemes: (1) Forms such as *s-o-lako* 'I try it with going' (tentative) and *s-o-lako* 'I would have gone' (counterfactual) show a contrast in meaning that I consider fundamental enough to exclude an analysis with a single polysemous morpheme. (2) The description of the distributional facts is simplified if one assumes that the tentative morpheme is in slot −5 (along with the future morpheme), while the counterfactual morpheme is in slot −6 (cf. table 4): The tentative and future markers (being in the same slot) show identical morphophonological behavior in every detail; the counterfactual marker can then be prefixed to the future + PNM complex. The conspicuous non-combinability of the counterfactual and tentative markers (cf. table 8) might have phonological reasons (haplology leading to a change of **se-s-o-lako tabo* > *s-o-lako tabo*), thus rendering the negative tentative form homonymous to the

Complex category	Example	Range of meanings
(Plain) realis	*o-lako* 1SG.RLS-go	'I go', 'I went'
(Plain) irrealis	*go-lako* 1SG.IRR-go	'I must go', 'I want to go', 'I will go (now)'
Future	*m-o-lako* FUT-1SG.RLS-go	'I will go', 'I can go', 'I may go'
	mo-go-lako FUT-1SG.IRR-go	
Tentative	*s-o-lako* TENT-1SG.RLS-go	'I try it with going'
	so-go-lako TENT-1SG.IRR-go	
Counterfactual	*s-o-lako* CNTF-1SG.RLS-go	'I would have gone'
Proximal imperfective	*k-o-lako* PROX.IPFV-1SG.RLS-go	'I am going (nearby)', 'I was going (nearby)'
Distal imperfective	*o-lako* DIST.IPFV;1SG.RLS-go	'I am going (further away)', 'I was going (further away)'

TABLE 6. Complex TAM categories encoded on Wogeo *lako* 'go'

As mentioned above, imperatives and prohibitives are exceptions to the pattern illustrated in table 6. The imperative is formed by the bare stem without the otherwise obligatory PNM prefixes; the prohibitive is formed by a combination of a verbal noun and a free grammatical morpheme. The formation of imperatives and prohibitives is summarized in table 7.

Complex category	Example	Range of meanings
Imperative	*lako* go	'Go!'
Tentative imperative	*se-lako* TENT-go	'Try it with going!'
Prohibitive	*lako~lako dol* go~NMLZ PROH	'Don't go!'

TABLE 7. Imperative and prohibitive forms of Wogeo *lako* 'go'

negative (plain) realis form.

Negations in Wogeo (with the exception of prohibitives) are formed analytically by a combination of the *counterfactual* prefix, a *realis* PNM prefix, and the negator *tabo*. Table 8 presents the negative forms of the corresponding non-negative forms found in table 6.[10] Several interesting facts can be noted: firstly, the obligatory combination of the counterfactual with the realis is unusual and surprising. Secondly, in the only complex category where realis and irrealis prefixes can be used interchangeably in the non-negative form (namely the future), the presence of the counterfactual plus negator *precludes* the use of the irrealis prefix (the other non-negative category compatible with both realis *and* irrealis prefixes, the tentative, does not have a specific negative form, as explained above.) And thirdly, there is one category (the future) where the counterfactual prefix is optional.

Corresponding complex category	Example	Range of meanings
(Plain) realis	*s-o-lako* *tabo* CNTF-1SG.RLS-go NEG	'I do not go', 'I did not go'
(Plain) irrealis	[No negative form exists]	—
Future	*se-m-o-lako* *tabo* CNTF-FUT-1SG.RLS-go NEG	'I will not go', 'I cannot go', 'I may not go'
	m-o-lako *tabo* FUT-1SG.RLS-go NEG	
	Not possible:	
	se-mo-go**-lako* *tabo* CNTF-FUT-**1SG.IRR**-go NEG	
	mo-go**-lako* *tabo* FUT-**1SG.IRR**-go NEG	
Tentative	[No negative form exists]	—
Counterfactual	*s-o-lako* *tabo* CNTF-1SG.RLS-go NEG	'I would not have gone'
Proximal imperfective	*se-k-o-lako* *tabo* CNTF-PROX.IPFV-1SG.RLS-go NEG	'I am not going (nearby)', 'I was not going (nearby)'
Distal imperfective	[No negative form exists]	—

TABLE 8. Negation of complex TAM categories encoded on Wogeo *lako* 'go'

[10] Three of the categories in table 8 have no specific negative form: *(plain) irrealis*, *tentative*, and *distal imperfective*. To express the meaning of a negative (plain) irrealis, the *prohibitive* is used (cf. table 7), while the meanings of negative tentative and negative distal imperfective are both expressed by the *negative (plain) realis*.

This brief exposition of the verbal morphology of Wogeo shows that in the majority of forms, the language employs a system where the realis and irrealis morphemes co-occur with other grammatical markers in complex categories, forming a *joint* system in Palmer's (2001:145–146) terminology. But while the realis and irrealis morphemes *can* also occur independently in the so-called *(plain) realis* and *(plain) irrealis* categories, the more peripheral markers, such as future etc., are obligatorily bound to the realis and irrealis morphemes and cannot occur without the latter.

To sum up: the so-called 'realis' prefixes are involved in the formation of the following complex morphological categories in Wogeo: *(plain) realis*, *counterfactual*, *proximal imperfective*, *distal imperfective*, *future* and *tentative* (in the latter two, optionally – they are alternatively formed with the 'irrealis' prefixes without change in meaning). The so-called 'irrealis' prefixes, on the other hand, are used in the formation of the following categories: *(plain) irrealis*, *future* and *tentative* (again, in the latter two, their use is optional and alternates with the 'realis' prefixes). Seen from the opposite perspective, the following complex categories are formed *exclusively* with the 'realis' prefixes: *(plain) realis*, *counterfactual*, *proximal imperfective* and *distal imperfective*. It is thus only the *(plain) irrealis* that is formed exclusively and obligatorily with the 'irrealis' prefixes.

Having looked at the formal distribution of the 'realis/irrealis' morphemes in Wogeo, we will now turn to the range of meanings that is associated with each of the respective forms.[11] First, the 'realis' morphemes are associated with the following meanings:

1. General:
    a) Present, past (obligatorily)
    b) Counterfactual; proximal imperfective; distal imperfective (obligatorily, but always in combination with the respective markers)
    c) Future, ability, permission; tentative (optionally; always with the respective markers)

2. Specific syntactic constructions:
    a) Negations (obligatorily)
    b) Protasis and apodosis of simple conditional clauses (obligatorily)
    c) Protasis of counterfactual conditional clauses (obligatorily; always with the counterfactual marker)
    d) Protasis and apodosis of hypothetical conditional clauses, apodosis of counterfactual conditional clauses (optionally; always with the future marker)

[11] 'Associated with' is a deliberately vague term: while the attribution of certain meanings to individual morphemes is straightforward in the case of the (plain) realis and irrealis categories, it is not at all clear what the contribution of the respective morphemes is in the case of the complex categories. In some, the 'realis/irrealis' prefixes may contribute to the resulting grammatical meaning, while in others, they may merely be *compatible* (synchronically) with those meanings. This question is not trivial and beyond the scope of this paper.

The semantic associations of the 'irrealis' morphemes, on the other hand, are as follows:

1. General:
    a) Obligation, volition, immediate future (obligatorily)
    b) Future, ability, permission; tentative (optionally; always with the respective markers)

2. Specific syntactic constructions:
    a) Complements of 'want' (obligatorily)
    b) Protasis and apodosis of hypothetical conditional clauses, apodosis of counterfactual conditional clauses (optionally; always with the future marker)

Some typical examples will serve as illustrations of the kinds of contexts in which the various forms occur. Example (1) shows the use of the (plain) realis form, in this case expressing *past tense*. This is a prototypical example in the sense of Bugenhagen (1993) in that it illustrates the use of a realis form to express positive polarity and non-future tense in a declarative speech act.

(1) (Plain) realis

va,	*ilo-g*	***e-****la-muta~muta-k-iko*
I	inside-1SG	**3SG.RLS-**INCH-be.tired.of~IMPV-APPL-2SG

'Me, I became tired of you.'

Turning to the 'irrealis' prefix, we can observe that in (2), one of the core meanings of (plain) irrealis in Wogeo, *obligation*, is expressed.

(2) (Plain) irrealis

iko	***go-****la-boalé*	*va*	*na*	*o-taval=te*
you	**2SG.IRR-**INCH-tell.3SG	I	FOC	1SG.RLS-die=TOP

'You must tell him that I did die.'

Another typical, construction-specific use of the (plain) irrealis is shown in (3), namely as a complement of 'want'. Like the example given in (2), this use is exclusive to the irrealis.

(3) (Plain) irrealis as complement of 'want'

do-boré	***dog-****va*	*gon-iak,*	*vaine*	*boe*	*ramata*
3DU.RLS-want	**3DU.IRR-**RECP	play-APPL.PL	woman	and	man

du-rú	*ma*
they-DU	FOC

'They wanted to sleep with each other, that woman and man.'

In the examples we have seen so far, there was a biunique relationship between the formal markers and the meanings they expressed. Examples (4) and (5), in contrast, show the indiscriminate use of the 'realis' and 'irrealis' prefixes in combination with the future prefix.

(4) Future (formed from the realis base)

vavá	*iko*	*va*	***m-u-****kila-k-an-iko*	*udemtaregá*
name.3SG	you	I	**FUT-1SG.RLS-**call-APPL.3SG-BEN-2SG	Udemtaregá

'Its name, which I will call it for you, is Udemtadegá.'

(5) Future (formed from the irrealis base)

va	*kat*	*va*	***mo-go-****jale-k*	*oageva*
I	canoe	I	**FUT-1SG.IRR-**go.down-APPL.3SG	Vokeo

'I will bring my canoe down to Vokeo.'

The somewhat unexpected exclusive association of the counterfactual with the 'realis' prefixes is illustrated in (6), where it is used in the protasis of a counterfactual conditional.

(6) Counterfactual

s-e-vá	*iko*	***sa-k-****lako,*	*katé*	*mo-la-moet*
CNTF-3SG.RLS-happen	you	**CNTF-2SG.RLS-**go	thus	FUT.2SG.RLS-INCH-disappear

'If you had gone, you would have been lost.'

Example (7), finally, illustrates what is by far the most common use of the counterfactual category in Wogeo, namely as the negated counterpart of the (plain) realis category (the so-called 'negated realis'). As in (1) and (6) above, this form and function is exclusively associated with the 'realis' prefix.

(7) Counterfactual as negated counterpart of (plain) realis

natú	*e-ot*	*taumdabí,*	*e-ot,*	*e-t-dom~doma,*
child.3SG	3SG.RLS-come	afternoon	3SG.RLS-come	3SG.RLS-INCH-look~IPFV[3PL]

tabo	*tiná*	***s-i-****mia*	*tabo*
but	mother.3SG	**CNTF-3SG.RLS-**stay	NEG

'Her son came in the afternoon, he came, looked around, but his mother was not there.'

Summing up, several observations suggest themselves. What seems to be especially interesting is that van Gijn & Gipper's (2009) implicational hierarchy is not valid for Wogeo, since counterfactuals – crucial to their claim – are always formed from the *realis* base, not the irrealis base. That exclusive association of the counterfactual semantics

with the 'realis' prefix in Wogeo is also one of the two main discrepancies between Bugenhagen's (1993) generalizations and the Wogeo data, the other one being the fact that his list in fact does not include what can be said to constitute the semantic core of the (plain) irrealis morphological category in Wogeo: *obligation* and *volition*. Other than those two (rather substantial) discrepancies, however, the functional range of the 'realis' and 'irrealis' morphemes in Wogeo can be described as largely consistent with Bugenhagen's (1993) results.

To be sure, such a purely negative characterization of the category is not satisfactory. As could be observed in the description of the semantic range covered by forms involving the 'irrealis' prefix in Wogeo (either alone or in combination with other prefixes), that range is largely coextensive with the domain of *non-epistemic necessity* in the sense of van der Auwera & Plungian (1998):[12] irrealis in Wogeo can be said to express non-epistemic necessity. Wogeo is therefore arguably a good example of a *mood-prominent language* in the sense of Bhat (1999).

As we have observed above, Wogeo is not untypical in showing such 'aberrations' from a supposed prototypical realis–irrealis system. On the contrary, judging from the typological studies available, Wogeo seems to represent the rule rather than the exception. What can one do with such a situation? Two basic possibilities readily present themselves, neither of which, in my view, is desirable. One possibility would be to say that if Wogeo does not fit the expected (or predicted) pattern, then it follows that the Wogeo category is not an instance of that pattern in the first place. Such an approach might make sense if one has good a priori reasons to assume that the predicted category is indeed valid and useful. The main problem that I see with that approach, however, is that a common semantic denominator can usually be 'constructed' for any subdomain of modality (in fact, that is what constitutes the semantic basis for the observed pattern of 'family resemblances' within the domain). So, if Wogeo is not a good example of the supposed category – which of the many other observed types of systems should be taken as a better example?

The second possibility would be to make the claim more general. However, that may not be a very helpful suggestion when it comes to characterizing individual grammatical systems. Precisely as Bybee (1998) points out: such a concept is too broad to be of practical descriptive use because it glosses over, and fails to explain, the very large differences that exist between individual languages in this respect.

The solution to the problem that I propose is that, as Bybee (1998) suggests, a language-specific, narrower category might be more helpful here than the wide category *realis–irrealis*; and what applies to Wogeo would likewise apply to other languages, too. Observed differences between languages are then best understood as (diachronic) relations of grammaticalization within the semantic domain of modality, and between that domain and its neighboring domains. The terms *realis* and *irrealis* may still be useful for comparative and historical purposes, where precisely such grammaticalization processes and semantic shifts need to be captured – keeping in mind that in that usage they are no more specific (rather, even less specific) than the terms *modal* and *non-modal* themselves.

[12] Note, however, that *volition* would have to be explicitly included, e.g. as a special case of van der Auwera & Plungian's (1998) *participant-internal necessity*.

As for Wogeo, the language seems to be in the middle of a grammaticalization process, with the original 'realis/irrealis' markers on the way to being semantically bleached, while the partly fused morphs (combinations of slots −6, −5 and −4 in table 4) are on the way to becoming new portmanteau morphs. On the other hand, in the majority of cases, the old 'realis/irrealis' markers are still more or less formally and/or functionally transparent in the formation of parallel sets of what I have called *complex categories* (cf. table 6).[13]

5. Conclusion. In this paper, I have tried to assess the conceptual relevance of the terms realis–irrealis, their relationship with the domain of modality (itself a controversial area), and their appropriateness as descriptive grammatical terms.

It was shown that languages that have been claimed to make use of a realis–irrealis category show extremely large variation in the semantic content of that category; indeed, not even a prototypical core meaning can be identified cross-linguistically. Neither a top-down nor a bottom-up (typological) approach has, in my view, so far been able to provide convincing evidence that there is indeed a need to postulate such a category.

It is of course conceivable that something like non-factuality is a valid concept in the minds of speakers, and that all the partial resemblances and diachronic developments that can be seen in the data are actually grounded in such a concept. However, I see a danger of circularity in the analysis here: it is equally possible that parallel, overlapping and interacting diachronic developments of neighboring (but in principle independent) domains could create the *illusion* of an underlying 'supercategory' like reality status. Does a putative concept of reality status *bring about* the observable facts, or do the observable facts (which really arise through independent developments) *look as though* they instantiate some concept?

Different typological studies were assessed that try to characterize *realis–irrealis* either as a well-defined (yet abstract) category, as a category with a prototypical core and fuzzy boundaries, or as an implicational hierarchy. However, it has been argued in this paper that all those attempts fail to solve the basic problem: namely, that the supposed category is either too vague (so that practically any language may fit in it), too narrow (so that language-specific idiosyncrasies outweigh any generalizations), or too language-specific (so that the category itself becomes arbitrary, and not comparable from a typological point of view). Data from Wogeo was presented to illustrate this point.

Taking into account the theoretical difficulties with the concept *reality status*, the lack of unequivocal linguistic evidence in favor of it, and the facts that can be learned from Wogeo, my view is that it is probably wisest at this point to side with Bybee (1998) and de Haan (2012). I agree with them in saying that, until evidence to the contrary is presented, what we are dealing with is not one large, highly abstract domain but rather many smaller, independent domains. The connection between those smaller domains is mainly *diachronic* via common paths of grammaticalization (van der Auwera & Plungian 1998). *Synchronically*, the domains are characterized mainly by partial resemblances.

As to the nature of the smaller domains that, as a whole, take the place of 'reality

[13] Practically, this creates the problem of glossing morphs, like in Examples (1)–(7), that are arguably in some contexts semantically empty, but not in others, like the 'realis/irrealis' prefixes in the complex morphological categories of Wogeo.

status', it is probably best to stick to fairly well-defined domains, like the (rather reduced) domain of modality as defined by van der Auwera & Plungian (1998) alongside domains like evidentiality, illocutionary force, polarity, etc. It is the language-specific interaction between them that accounts for the type of 'reality status' system characteristic of any given language.

Finally, it was suggested that *realis–irrealis* may nevertheless sometimes be useful as a pair of terms to capture certain formal diachronic processes and relationships within and between languages (e.g. in the historical-comparative study of Austronesian or New Guinea area languages), but that different terms that more accurately capture the semantics of a given language-specific category may be more helpful in many, if not most, descriptive contexts.

References

Anderson, Astrid. 2011. *Landscapes of relations and belonging: Body, place and politics in Wogeo, Papua New Guinea* (Person, Space and Memory in the Contemporary Pacific 3). New York: Berghahn.

Anderson, Astrid & Mats Exter. 2005. *Wogeo texts: Myths, songs and spells from Wogeo Island, Papua New Guinea* (Kon-Tiki Museum Occasional Papers 8). Oslo: Kon-Tiki Museum, Institute for Pacific Archaeology and Cultural History.

Auwera, Johan van der & Maud Devos. 2012. Irrealis in positive imperatives and in prohibitives. *Language Sciences* 34. 171–183.

Auwera, Johan van der & Vladimir A. Plungian. 1998. Modality's semantic map. *Linguistic Typology* 2. 79–124.

Bhat, D. N. S. 1999. *The prominence of tense, aspect and mood* (Studies in Language Companion Series 49). Amsterdam: Benjamins.

Bugenhagen, Robert D. 1993. The semantics of irrealis in Austronesian languages of Papua New Guinea: A cross-linguistic study. In Ger P. Reesink (ed.), *Topics in descriptive Austronesian linguistics* (Semaian 11), 1–39. Leiden: Vakgroep Talen en Culturen van Zuidoost-Azië en Oceanië.

Bybee, Joan L. 1998. 'Irrealis' as a grammatical category. *Anthropological Linguistics* 40. 257–271.

Cristofaro, Sonia. 2012. Descriptive notions vs. grammatical categories: Unrealized states of affairs and 'irrealis'. *Language Sciences* 34. 131–146.

de Haan, Ferdinand (see Haan)

Elliott, Jennifer R. 2000. Realis and irrealis: Forms and concepts of the grammaticalisation of reality. *Linguistic Typology* 4. 55–90.

Exter, Mats. 2003. *Phonetik und Phonologie des Wogeo* (Arbeitspapier, N. F. 46). Cologne: Institut für Sprachwissenschaft, Universität zu Köln.

Exter, Mats. 2012. A grammar of Wogeo. Unpublished ms.

Gijn, Rik van & Sonja Gipper. 2009. Irrealis in Yurakaré and other languages: On the cross-linguistic consistency of an elusive category. In Lotte Hogeweg, Helen de Hoop & Andrej Malchukov (eds.), *Cross-linguistic semantics of tense, aspect, and modality* (Linguistik Aktuell 148), 155–178. Amsterdam: Benjamins.

Givón, T. 2001. *Syntax: An introduction*, 2nd edn. Amsterdam: Benjamins.
Haan, Ferdinand de. 2012. Irrealis: Fact or fiction? *Language Sciences* 34. 107–130.
Hogbin, Ian. 1970. *The island of menstruating men: Religion in Wogeo, New Guinea*. Scranton: Chandler.
Hogbin, Ian. 1978. *The leaders and the led: Social control in Wogeo, New Guinea*. Melbourne: Melbourne University Press.
Mauri, Caterina & Andrea Sansò. 2012. What do languages encode when they encode reality status? *Language Sciences* 34. 99–106.
Palmer, F. R. 2001. *Mood and modality*, 2nd edn. (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.
Pietrandrea, Paola. 2012. The conceptual structure of irreality: A focus on non-exclusion-of-factuality as a conceptual and a linguistic category. *Language Sciences* 34. 184–199.
Portner, Paul. 2009. *Modality* (Oxford Surveys in Semantics and Pragmatics). Oxford: Oxford University Press.
van der Auwera, Johan (see Auwera)
van Gijn, Rik (see Gijn)
Whorf, B. L. 1938. Some verbal categories in Hopi. *Language* 14. 275–286.

Mats Exter
exter@phil.uni-duesseldorf.de

8

Language Documentation & Conservation Special Publication No. 5 (December 2012)
Melanesian Languages on the Edge of Asia: Challenges for the 21st Century,
ed. by Nicholas Evans and Marian Klamer, pp. 191–218
http://nflrc.hawaii.edu/ldc/sp05/
http://hdl.handle.net/10125/4565

From mountain talk to hidden talk: Continuity and change in Awiakay registers

Darja Hoenigman
The Australian National University

When the Awiakay of East Sepik Province in Papua New Guinea left their village or bush camps and went to the mountains, they used a different linguistic register, 'mountain talk', in which several lexical items are replaced by their avoidance terms. In this way the Awiakay would prevent mountain spirits from sending sickness or dense fog in which they would get lost on their journeys. Over the last decade people's trips to the mountain have become more frequent due to the eaglewood business. However, Christianity caused a decline in the use of 'mountain talk'. Yet a linguistic register similar in its form and function has sprung up in a different setting: *kay menda,* 'different talk', or what people sometimes call 'hidden talk', is used when the Awiakay go to the town to sell eaglewood and buy goods.

Like other cultural phenomena, linguistic registers are historical formations, which change in form and value over time. This paper aims to show how although in a different social setting, with an expanded repertoire and a slightly different function, *kay menda* is in a way a continuity of the 'mountain talk'.

1. Introduction. This paper will look at two linguistic registers practiced by the Awiakay people.[1] One of these, which we can refer to as 'mountain talk', was originally used when travelling to the mountains, but is now more or less obsolete. The other, newly developed register, which we can name 'hidden talk', is used when Awiakay people travel to town. Both are referred to as kay menda 'different language' or kay momba 'different talk' by the Awiakay. I will explore the ways in which 'hidden talk' can be viewed as a continuation of 'mountain talk'.

It is not uncommon for languages of the New Guinea Highlands to have special linguistic registers characterised by lexical substitutions and used in particular social contexts. In Kewa, for example, the use of a special speech variety is associated with notions of high mountains being inhabited by wild dogs and spirits from whom one must protect oneself. Similarly, Huli use a special vocabulary when travelling through country inhabited by demons (Franklin 1972). Other 'hidden languages' are used in ritually restricted contexts: while hunting (Telefol trapping rats; ibid.), or on pandanus harvesting expeditions when cooking and eating cassowary (Pawley 1992), etc. However, some of these registers have declined (Franklin & Stefaniw 1992).

While several authors have looked into lexical substitution registers, few have attempted to trace the diachronic changes. This paper will show how the use of a register is adapted to new socio-economic circumstances. The example of 'hidden talk' provides us with the rare opportunity of analysing this process while it unfolds.

2. The Awiakay and their Language. Awiakay is a Papuan language spoken by 300 people living in Kanjimei village in the East Sepik Province of Papua New Guinea (see map 1).[2] The village itself is located on the Konmei River, which is a tributary of the Karawari, while the major part of the Awiakay land stretches south into the mountains.

---

[1] This paper was originally presented at Workshop on the Languages of Papua 2: *Melanesian Languages on the Edge of Asia: Past Present and Future* in Manokwari, Indonesia, 8-12 February 2010. I would like to thank Nick Evans for inviting me to participate at the conference and for suggestions on how to improve the paper. For valuable comments on earlier drafts I wish to thank Christian Döhler, Bethwyn Evans, Andrew Pawley, Alan Rumsey, Lila San Roque and Borut Telban. The accompanying films would be but mere cuts without the expertise and artistic eye of Gary Kildea who was generous with his time and patience in helping me edit the footage and sub-title the edited segments. *Tenkyu tumas, Masta G*! I am also grateful to the two referees, Rupert Stasch and Lourens DeVries, for their detailed reviews of the paper and helpful suggestions for further work on this subject. My greatest debt, however, lies with the Awiakay people for sharing their lifeworld with me.

[2] Awiakay is not only how the speakers refer to their language, but it s also used by the inhabitants of Kanjimei to refer to themselves.

Map 1. Kanjimei - Wewak route

The Awiakay economy remains largely a subsistence one. People supplement their sago diet by hunting, fishing and gathering. Gardening is of minor importance.[3]

Figure 1. Tikinjao washing sago

Nowadays, all Awiakay adults are bilingual in Tok Pisin and Awiakay. Multilingualism in other local languages is less common, but it occurs in the few families where one spouse is from the neighbouring Asangamut village. Among adult Awiakay, the use of Tok Pisin is confined mainly to situations where it functions as a language of authority. Code-switching between Tok Pisin and Awiakay occurs in public speeches, in quarrels and in other situations where a speaker (of any gender and age) wants to take an authoritative position in the communicative act. All children are fluent in Awiakay, but acquire Tok Pisin at a very early stage. They are addressed primarily in Awiakay, while Tok Pisin is used for scolding.

Words from Tok Pisin – particularly ones denoting items and concepts which have entered the village from the outside – do enter Awiakay and are used in everyday speech. Many of them are nativised, that is, adapted to the rules of Awiakay phonology and morphology. Moreover, Tok Pisin verbs which are borrowed into Awiakay acquire a special suffix, *-bapo-*, which is attached to the borrowing and precedes the normal Awiakay verb ending (cf. Hoenigman 2007: 209). For example, Tok Pisin verb 'buy' gets adapted by

[3] There is both linguistic and cultural evidence that gardening has been adopted relatively recently (Hoenigman 2007: 102-4).

adding the 'Awiakayser' *-bapo-*, as well as Awiakay tense, number and person endings.[4]

(1) *baim* → *baim-bapo-pali-k*
(TP) buy (TP) buy-LA-PRES-1SG

As we shall see, Awiakay words have also been coined for many of these borrowings, but are only used in specific situations.

3. Awiakay 'mountain talk'. The Awiakay employ four basic terms to describe their landscape: *andaŋ* 'swamp', *mip* 'flood plain', *palakay* 'flat ground', and *pondoŋ*, which denotes land of significantly higher elevation than its surroundings and can be translated as 'mountain'. The Awiakay consider their mountainous land to begin above ~70m above sea level, with the highest mountain on Awiakay land, Injaiŋ, being 1,331m above sea level.

Figure 2. Injaiŋ

[4] Nonstandard abbreviations to be found in this paper: (TP) for Tok Pisin; LA for loan adaptation.

FIGURE 3. Drawing Awiakay mountains and creeks

Before the commercial eaglewood[5] trade, which started in the Awiakay region just before 2000, the Awiakay went to the mountains mainly in search of *kanuŋ isa* 'blackpalm', which they used for making bows, or for short hunting trips. Such trips were restricted to some mountains only, as others were perceived to be heavily populated by both *endembaŋ* 'mountain spirits' and *taŋgia* 'spirits of the dead', the latter being particularly malevolent. On these trips people used to employ a different linguistic register, which I refer to as 'mountain talk', in which certain lexical items are replaced by avoidance terms. There is a myth about a man who became lost in the mountains and met a female mountain spirit. This spirit hid him and taught him the avoidance terms for animals, plants and foods which people should use while in the mountains. Some of these lexical prohibitions and their replacements are mentioned in the myth. Today Awiakay people remember no more than some 20 avoidance terms.

[5] Eaglewood (*Gyrinops ledermannii*) or Tok Pisin *garu*, from Indonesian *gaharu*, is known for its fragrant resin, which Awiakay call *is-kamia* (literally 'tree-meat'/'wooden meat'). It is produced as the tree's response to an injury and is thus found only in a small percentage of eaglewood trees. This black resinous wood is highly sought after by traders because of its commercial value, and is sold to Middle Eastern countries and Japan "for religious, medical, ceremonial and domestic activities by Asian Buddhists and Moslems" (Gunn et al. 2004:1).

Ordinary Awiakay	Mountain Awiakay
aisia 'eel'	no term should be used at all
ayŋgwaŋ 'flying fox'	*apuria* 'type of bee'
kamdok 'cloud'	*kandukya* 'white'
kawin 'mountain bird – spirit of a dead man'	*tiñe pawiakay* 'red bird'
kayma 'cassowary'	*tumanjiŋgoy* 'the hairy one' OR *kondamin panba* 'two legs'
koŋgonoŋ 'a tall type of ginger (*Alpinia sp.*)' (= name of a mountain spirit)	*is kaŋga* 'tree leaf'
momok (*tawa*) 'spine of a type of cane which the Awiakay use in circular roof building' (= name of a mountain spirit)	*injam kanja* 'cane tooth'
muŋguma 'termites' nest'	*nam tapuka* 'old woman'
tao 'sago spines'	*andaŋgamgoy kolokot* 'something belonging to swamp' OR no term should be used at all
tay 'sago'	*kandukya kolokotay* 'white food'
umbuŋ 'slit drum' OR 'garamut tree (*Vitex confossus*)' from which slit drums are made	no term should be used at all
yaki 'tobacco'	*emwi kolokolay* 'the smoking thing'
yambuka 'leaves of a type of ginger' (= name of a mountain spirit)	*is kaŋga* 'tree leaf' OR no term should be used at all
name of any fish found in the upper parts of Awiakay creeks	no term should be used at all

Table 1. Avoidance terms in mountain Awiakay

By using this linguistic register, the Awiakay would satisfy the demands of mountain spirits and prevent them from carrying out malevolent acts, such as sending sickness or

dense fog in which they would get lost.

Over the last decade, people's trips to the mountains have become more frequent due to the commercial eaglewood trade. This wood grows mainly at altitudes between 70 and 850m. On Awiakay land this is the region south of Kanjimei, where the land starts rising into the mountains (see map 1).

As people spent more time in the mountains, one might expect that 'mountain talk' would thrive (at least I did). However, a Catholic charismatic movement, which the Awiakay accepted in 1995, demanded of people that they radically cut their traditions and break their relationships with the spirits (Telban 2008a, b, 2009) – and therefore with their land.[6] Sickness (and even death) caused by not thoroughly implementing the expected practices would now no longer be inflicted by the spirits for *not* following their demands, but rather by God for following them. In order to protect themselves from God's anger, people were now forced to abandon the very practices which used to protect them from spirits, which has meant a gradual decline in the use of 'mountain talk'.

FIGURE 4. Searching for eaglewood at Umbim

4. AWIAKAY 'HIDDEN TALK'. While 'mountain talk' has declined, a new linguistic register rather similar in its form and function has sprung up in a different setting. *Kay menda*

[6] The Awiakay gradually accepted Christianity in the 1960s. Catholic missionaries who occasionally visited the village put most effort into uprooting initiation rites, while many of the customary practices connected with spirits continued to coexist with the nominal Christianity.

'different language', or what I will refer to as 'hidden talk' is used when the Awiakay go to the town to sell the eaglewood which they have harvested, and to buy goods.

Wewak is the provincial capital and has in recent years become frequently visited by people from remote areas who come to sell eaglewood and small quantities of gold. In the early years of the eaglewood trade Indonesian buyers would themselves travel around the province to buy the aromatic wood. However, this became dangerous, as it was known that they were carrying huge amounts of money, and they were often robbed (reportedly two of them were even killed in early 2004 in an attack on the Sepik River). When the initial boom in the eaglewood trade declined, these foreign buyers did not earn enough to be willing to take the risks and so they gave up their field trips. On the other hand, people whose land is rich in eaglewood earned enough to buy outboard motors and started travelling to Wewak themselves, to sell their eaglewood and also to buy goods which can only be obtained in town.[7] This increase in visitors with money who are not used to town has coincided with an increase in crime. It is not uncommon for visitors to town to be robbed of all their possessions.

Being aware of these dangers, the Awiakay people try to be extremely cautious when in Wewak, and have (among other things, such as carrying cassowary bone daggers) started practising 'hidden talk'. Hidden talk is a register of lexical substitution, in which all Tok Pisin borrowings, which are used in everyday Awiakay in the village – and may therefore be understood by outsiders – are replaced by newly-coined Awiakay terms. Coining new words for newly introduced items and concepts is a common practice in many languages of New Guinea. But what makes this phenomenon different from similar processes in other languages is the special function that the Awiakay attribute to these newly-coined expressions in their vernacular (namely concealing the meaning of commonly used Tok Pisin borrowings), and the special social setting in which this is done (not in the village, but when going to the town). In other words, it is important to note that Tok Pisin terms – particularly ones denoting items and concepts which have entered the village from the outside – *do* enter everyday Awiakay as it is used in the village. Even though many of these borrowings have been nativised by adapting them to the rules of Awiakay phonology and morphology, they are nonetheless parts of the Awiakay language (which is not spoken by people from other villages and even less so by anyone in the town) that could be understood by other people. Awiakay words have therefore been coined for many of these loans, but they are primarily used in situations requiring *kay menda*, while Tok Pisin terms continue to be used in Kanjimei.[8]

To illustrate how *kay menda* works, consider how people refer to a 44-gallon drum. In Tok Pisin this is *fotifo*. A traditional item with the most similar function to a drum was a bucket made from a large bamboo, used for carrying water. In Awiakay it is called *yomoy*,

[7] Out of six outboard motors in Kanjimei four were bought with people's earnings from selling eaglewood.

[8] In certain situations, when the Awiakay people want to conceal their talk from visitors from other villages, they will resort to kay menda even at home in Kanjimei, but typically this register is used whenever they go to town.

and people adopted this term to replace *fotifo*. However, while at home (on their land), the Awiakay would continually use *fotifo*, but leaving their land, particularly when going to town, they would call it *yomoy* when speaking Awiakay to each other.

So far I have collected about 120 Awiakay creations used in *kay menda* in place of Tok Pisin terms (see Appendix). We can divide these terms into five groups according to the way in which they were created:

1 – terms which denote objects with similar functions
- wallet: (TP) hanpaus → *kundambi* 'coconut fibre for storing tobacco'

2 – terms which denote objects similar in form (they look similar)
- petrol: (TP) petrol → *yom* 'water'
- balloon: (TP) balun → *mumba* 'bladder'
- gold: (TP) gol → *kiŋim* 'sand'

3 – descriptive terms
- store: (TP) stoa → *kolokot yawa* 'things house'
- bra: (TP) susu kalabus → *isik ulakaplakay* '(something that) covers breasts'

4 – lexical calques
- toilet: (TP) haus pekpek → *eneŋ yawa* 'shit house'

5 – absurdly incongruous terms (a word denoting something that people find disgusting is used for something they find delicious on the basis of physical resemblance)
- noodles: (TP) nudols → *kundam enga* 'earthworm shit'
- chocolate cream: (TP) soklit krim → *eneŋ mola* 'diarrhoea/rotting shit'
- tinned (mushroom) sauce: (TP) (?) → *meŋgwak* 'vomit'

Although coining new words for newly introduced items and concepts is a common practice in many languages of New Guinea, it is the special function that the Awiakay have attributed to using these newly-coined expressions in their vernacular (i.e. concealing the meaning of commonly used Tok Pisin borrowings) and the special social setting in which this is done (i.e. not in the village, but when going to the town) which makes this phenomenon different from similar processes in other languages.

Some of the examples of how *kay menda* is used can be drawn from a video recorded eaglewood selling trip to Wewak in September 2009.

In spite of a fortnight without any rain in the mountains, which would fill up the creeks that would send water to the Karawari river, the Kangrimei passage was still navigable, which saved Desmon Asuk, Dicson Tumak, Sailus Kaim, Justin Pupi and me a whole afternoon's journey down the Karawari River. Apart from shortening the long journey, using this shortcut also means that the Awiakay can save about five gallons of petrol (and a bit more on the way back when going upriver), trade for food with the people from

Karawari-speaking Kaiwaria and Masandanai villages along the channel and overnight in one of their camps. Kangrimei was very low though, so we had to turn off the motor so as not to hit the branches and tree trunks lying at the bottom. While paddling, Asuk, who is more experienced in travelling to town, started a conversation in which he repeated for the younger boys and me how we should behave when we come to Wewak.[9]

KAY MENDA – FILM 1 transcript

Asuk: *Noŋ omgusanda aŋ kak pekeŋgoy enduŋ opiangombemgoy olukunja taŋan aka paŋgumbem* ***iskamia*** *salimbapoŋgapekeŋbop.*
When you meet somebody in the town, don't tell them that we came down to sell tree meat [eaglewood].
Aunda yameŋga pekeŋ. Taŋan paŋgombem.
We just came down for a trip. Tell them that.
Ya noŋ kele koŋ kakanua: "Aka aunda, aunda yameŋga pekua."
And they will say: "True, they just came for a trip."

Sailus: *Yo. Aŋ opepaluŋ.*
Yes. We know [what to do].

Asuk: *Taŋan ponua.*
That's what they'll think.
Ya elak kele emepanda ulakapep pakayamenaŋ aŋgumgoy kolokot kele.
That's how we will be able to hide what it is we are carrying.

Sailus: *Emepanda tok.*
That's good.

Asuk: *Mawia tok kele.*
It's great.

Sailus: *Mawia.*
Great.

[9] In the transcript lexical substitutes are bolded and underlined. In translation, original Awiakay meanings are underlined, while their 'hidden', *kay menda* meanings follow in square brackets, e.g. ***iskamia*** (lexical substitute) is translated tree meat [eaglewood]. Tok Pisin expressions are *blue and underlined like this*. In the transcripts the translations are more faithful to the original Awiakay text, while they had to be slightly modified (shortened) for subtitling purposes. The transcripts come first so that the reader can become familiar with the meaning of avoidance terms before watching the films, which are an integral part of this paper. In the subtitles lexical substitutes are in yellow. In order to illustrate how *kay menda* works, I have glossed Awiakay terms using their original Awiakay meanings, rather than their *kay menda* meanings, e.g. I have glossed *ikakapan* as 'carving' (ordinary Awiakay) instead of 'writing' (*kay menda*).

Asuk: *Koŋgotmay an anda aka yañaŋgunay elaŋ an. An aka yañaŋgunay.*
Koŋgotmay [Darja] will not tell them either. She won't tell.
Pupi: *An opepon.*
She knows.
Darja: *Niŋ ... niŋ anda opepalik.*
I ... I know.
Asuk: *M-m. Anda opepon. Aka yañaŋgunay.*
M-hm [agrees]. She knows. She won't tell.
Darja: *Andoposa opepalik.*
I know that very well.
Asuk: *Elak tok aŋgumgoy kunja kolokota elakay **paypmaŋga** epaluŋgoy tok.*
This is the only thing we get stones [money] for.
Darja: *Yo.*
Yes.
Asuk: *Akanja olukunjan mokongunuam epop **emayn**, **emay kuŋanjan**.*
Bad people can mug us. Sorcerers [rascals] or sorcerer children [pickpockets].
Darja: ***Emay** wakon. Kumbi akanja Wewak.*
There are many sorcerers [rascals] there. Wewak is a bad place.
Asuk: *Elaŋ anduŋ ... Elaŋ anduŋ koŋ aka kakapaluŋ.*
That stuff [of ours]... let's not talk about it at all.
Aŋ anda tui mambipep, pakambaluŋña, s-salimbapopaluŋña, ya koŋ wambopaluŋ.
We'll just hide it, bring it there, sell it and come back upriver.

The reader is now invited to watch a film excerpt from our trip to the town, which is available at http://youtu.be/tLzLCpwz6Aw [1:22].

The Awiakay are afraid of being held by rascals and robbed, so eaglewood selling trips are always permeated with secrecy. No one ever discusses their business with people whom they meet on the river or in the camps where they overnight, let alone with anyone in the town. Wewak is perceived to be a dangerous place, yet one where the Awiakay can get all the goods they desire. Young boys already learn that by listening to the conversations of the more experienced men in the village, but Asuk repeats it in order to make sure that it is clear to all of us. As I had travelled to town to sell eaglewood with other men before, all the boys knew that I had learned how important it was to keep our business secret (line 11). In line 16 Asuk explains that selling eaglewood is the only way in which they can get money. This, however, attracts robbers and pickpockets, for whom he uses *kay menda* terms, *emay,* (TP *sanguma*) '(assault-)sorcerer', and *emay kuŋanja,* '(assault-) sorcerer children' in the

meaning of 'pickpockets' (line 18). Calling names of dangerous entities is often avoided,[10] so although the five of us are alone in the canoe on the Kangrimei, and there is no danger of anyone else overhearing our conversation, he chooses to use *kay menda* terms for rascals.

As it started getting dark we decided to spend the night at Kambatiman, a Masandanai camp in the middle of the Kangrimei passage, about halfway to Angoram (see map 1). We were not alone there – a family from a nearby Kaiwaria village stayed in another shelter. That is why Tumak and I used *kay menda* to replace Tok Pisin terms which could reveal Tumak's plans in town.

KAY MENDA – FILM 2 transcript

Darja:	*Tumak, o!*	1
	Hey, Tumak.	
	*Amba... amba im... amba momba **ikakapan**?*	
	What... what are you do... what are you carving [writing]?	
Tumak:	*Amba ...? Ey!*	
	What ...? Oh [looks up in surprise]!	
	Paypmanda *...* ***paypmanda*** *George sakay mamgoy* ***ḅaḷ emba*** *tike* ***mimbia ikakapalik****.*	
	Stones [money] ... stones ... I am carving [writing] the name of the stones [amount of money] George's [wife] gave me to take [buy] a ball.	
Darja:	*Yo. M-m. Kaykay olukunja* ***giyim paypmanda****. Yo.*	5
	Yes. M-hm. Many people gave stones [money] to you. Yes.	
Tumak:	*Ponde* ***ḅaḷ epep*** *pakinakoy Tanday sakay.*	
	Tomorrow I will take [buy] a ball for Tanday [George's son] and take it back upriver [to the village].	

Now please watch http://youtu.be/D4I3SijASw8 [0:37].

When asked what he was doing, Tumak was taken by surprise when he saw me with a video camera. As he was aware of the presence of people from another village (a Kaiwaria woman was sitting in a nearby wind house and approached when we started speaking), and possibly reminded by myself using a 'hidden' term for writing, he knew that he had to replace the expressions like 'money' and 'buy', with their *kay menda* terms. However, we could hear him hesitating, carefully thinking how to formulate his sentences.

Arriving in Angoram in the early afternoon of the following day and storing the canoe with the Imanmeri people, we managed to find an early 'backload' truck that was going to

[10] Even in the village people would often avoid using a word for e.g. a harmful spirit. Instead of saying *nuŋgum* 'gigantic python' when describing a picture where a speaker believed this creature was threatening a man depicted in the drawing, he would say *kalak ambam* 'this what' or *kolokolay* 'this thing'.

take us to Wewak. The major part of the journey was over and although the ride along the dirt road to Wewak is a rough one, the men who were in charge of the canoe could now take some rest. There were only a couple of other people sitting on the truck and waiting with us, and the leisurely conversation that took place did not involve anything that would demand secrecy, but the closer to Wewak the Awiakay get, the more urge they feel to speak among themselves in a way that other people do not understand them.[11]

KAY MENDA – FILM 3 transcript

Asuk:	*Mawia tok.*	1
	Great.	
Sailus:	*Pekepiaŋ, pekepiaŋ, ya ambuŋ.*	
	We came downriver and we are going now.	
	Kumap mandaŋ *koloŋ. Tom...*	
	We're in the coconut shell. Later...	
Asuk:	*Aka* ***kumap manda****.* ***Yomgoŋ manda!***	
[interrupts]	Not coconut shell. Turtle shell [car/truck]!	
Sailus:	***Yomgoŋ manda.*** ***Yomgoŋ mandaŋ*** *koloŋ ya. Ambopaluŋ.*	5
	Turtle shell. We're in a turtle shell [car/truck]. We're going.	
	[we're on our way.]	
Asuk:	*Ambopaluŋ ya...*	
	Yep, we're going. [we're on our way.]	
Tumak:	*Ambembapopaluŋ ya, ṭaụṇụŋ.*	
	Now we are going to town.	
Asuk:	***Yomgoŋ mandaŋ*** *kolopep oŋga kolopaluŋ ya.*	
	We're sitting together in our turtle shell [car/truck] now.	
	Unja tok kele Wapiak yomonaŋ.	
	Now [tonight] we will sleep in Wewak.	
	Mawia.	10
	Great.	

Now please watch http://youtu.be/rvZaC41ZPKc [0:48].

Individuals first become acquainted with *kay menda* in the village, but only put it into practice when travelling to town. Every trip to the town is therefore a training for the boys who are not yet fully competent in this register. They are taught *kay menda* by the more experienced men. These also correct the boys when they make mistakes. Film 3 shows how Asuk corrects Sailus, who calls car 'coconut shell' instead of 'turtle shell'.

[11] After one of the internal village fights Pupi's brother Namay said: "*Ɠụtpẹlą ṭaịṇ buŋ aka ambaluŋ. Ṭąụṇụŋ aninangoy tok pukuninaŋ.*" 'We never get together at good times. When we go to the town, we think of each other.' [In the village we tend to quarrel and fight. But when we go to the town we stick together as one and take care of each other.]

When they are in Wewak, the Awiakay normally overnight with people in Masandanai camp at Kria (Kreer market), a settlement of the Karawari-speaking communities. Although they are on friendly terms with Karawari people (albeit not their wantoks), they find it very important to conceal their business and plans from them. The boy we see sitting and writing in film 4 after we arrive in Wewak is a Masandanai boy who goes to school in town, the others are Awiakay, discussing their plans for the following day.

KAY MENDA – FILM 4 transcript

Asuk: *Noŋ amba kolokota mae **enamin** nan?* 1
What will you take [buy].
Sailus: *Niŋ aninakoy ... amba ... eŋganinak ...*
I will go ... what... go and take [buy] ...
Asuk: *Pisikanda, pisikanda kakaym.*
Quickly [come on] tell me.
Sailus: *Ya, amba oŋga **eŋaninak**?*
Yes, what is it I'm taking [buying]?
Amba endeplakay. 5
What ... they strain [stuff] with it.
Pupi: *Tay muŋga ...*
Sago starch ...
Sailus: *Tay muŋga endeplakay.*
They strain sago starch [with it].
Tumak: *Streina.*
A strainer.
Asuk: *Aka pukupan.*
You don't remember.
Sailus: *Iss! Elak an aka koñim. Numbinman!* 10
Iss, don't call its name. You fucker!
Tumak: [laughs]
Asuk: *Aunda endaŋ aka tapuka yañinak.*
No other way of telling him.
Kak.
Tell us.
Sailus: *Kak mom agalon ...*
Nothing more to tell ...
Asuk: [*Niŋ ponde anakoy* + coughing in the background]. ***Kamboy*** *kondamin* 15
enakoy, *eŋa kunjakanta, taŋan enak.*
[Tomorrow I will go] and take [buy] two stone axes [axes] and a bushknife.

Sailus: *Amba tok **yom** **eŋambongoy**.*
And what else, she [Darja] will take [buy] water [petrol].
Ya tok wakonduŋ okokoaninaŋ.
We'll all go with her.
Asuk: *Elak tok. Wakonduŋ anaŋ.*
That's it. We'll all go [together].
Tumak: ***Yom** **epep** embepenaŋ...*
We'll go and take [buy] the water [petrol] ...
Asuk: ***Yom** omgusanda eŋambopep, kaŋ embepenaŋ.*
We will all go to get the water [petrol] and bring it here.
Tumak: *... kaŋ embepenaŋ.*
... we'll put it here.
Sailus: *Mae anamgoy ambaŋ anayke **tasia yawaŋ**. Amba pondanayke*
First she'll go to her what... spirit house. To take out ... what?
Asuk: ***Paypmaŋga** eŋanambop.*
She will go and take stones [money].
Sailus: ***Paypmanda** enayke ...*
When she takes the stones [money] from ...
Asuk: ... ***paypmaŋga yawa** ...*
... the house of stones [bank] ...
Sailus: *... anamgoy kolokot **enayke**, pakapukundinaŋ mae.*
... and takes [buys] her things, we'll load them [onto the truck].

Now please watch http://youtu.be/INnBhGApTxc [1:52].

Asuk asks Sailus what he is going to buy, and Sailus hesitates with his answer, not knowing what to call 'the thing for straining sago flour' in *kay menda*. He avoids using the Tok Pisin term by calling it 'what for straining' (line 5) ,'what' standing for 'that thing' (cf. fn. 3). Both Pupi and Sailus are searching for the right term (lines 6 and 7) when Tumak gives up and calls a Tok Pisin word *streina* 'strainer' (line 8) at the same time when Asuk says that they cannot remember. Tumak is instantly reprimanded by Sailus who calls him 'fucker' (line 10), which makes them all laugh, but Asuk defends him by saying that there was no other way of telling this (line 12), as he himself, as the most competent speaker of *kay menda* and the leader of this trip to Wewak, cannot think of a suitable avoidance term. The conversation continues by Asuk telling what he will buy the next day and turns to how they are all going to go with me to buy petrol and bring it to Masandanai camp. Sailus is stuck again when he wants to say that I first need to go to the bank to take out my money. By calling the bank a 'spirit house' (line 23) he is confused again, and uses *amba* 'what' even when he wants to refer to money. Asuk helps him out by reminding him of both terms, *paypmaŋga* 'stones' for money and *paypmaŋga yawa* 'stone-house' for bank (lines 24 and 26). Sailus corrects himself by using an alternative term for money, *paypmanda*, saying that when I take out my money and buy all the goods, we will load it all onto a truck (for Angoram).

In situations like this *kay menda* becomes a kind of a mind game which all participants enjoy, even though its primary purpose is to make the Awiakay feel safer while in town. The next day we went shopping. As Asuk wanted to put some of the money he earned with eaglewood in the bank, he and Pupi went to do this business, while Sailus, Tumak and I went to the shops.

KAY MENDA – FILM 5 transcript

Tumak: *Wakon. Ṣkuluŋ pakayamenakpokoy ḅag kalakiay* ***enapok***.
So many. If I could take [buy] this bag I could carry it to school.

Sailus: *Aka anda. Ṣkuluŋ pakayamenakpokoy ḅag kalakiay* ***enapok***.
True. If I could take [buy] this bag I could carry it to school.

Tumak: *Aka kolokot.* ***Paypmanda*** *tonaypeke wakon aka kiay* ***enapok***.
What a thing! If I had lots of stones [money], I would take [buy] many.

Sailus: *Kandikak.* ***Andaŋguŋ yaka yamblakay****. Andaŋguŋ.*
Here. [Something for] wandering around in swamps [gum boots]. For swamps.
Andaŋguŋ yaka yamblakay ...
[Something for] wandering around in swamps [gum boots] ...

Tumak: *Ḅag, o!*
Oh, bag!
Mawiakay kalak.
This is a great one.

Sailus: ***Emay*** *kalak yambongoy, poka pukulakana pokoy anda kaykay wakakanaype.*
If this sorcerer [rascal] keeps tailing us, I will bash his face till he screams.

Sailus: *Amba pia kandikakay?*
Is this a piece of something?

Tumak: *Kolokot muŋayambla.*
[See], they are wandering around and looking at things.
Apiaŋ sakay amba pisipmgoy, ṭawel pisip.
This is like Apiaŋ's what ... like a towel.

Sailus: *Amgam? Wakon.*
How much? A lot.

Tumak: ***Pokonuŋ pasiplakay***.
Something to clean your face with [towel].
Kuŋanja amgoy taŋanim?
Is it for children?

Sailus: *Tom kele elokiay opoŋanak.*
I'll look at that later.

Tumak: *Aŋge ya.*
Let's go.
Sailus: *Aŋge.*
Let's go.
Tumak (to me): ***Emay*** *nanday okokaim yambon.*
A sorcerer [rascal] is following you.
Emay *nanday okokaim yambon, yo kon yambon.*
A sorcerer [rascal] is following you, he's walking just there.
Niŋ mae mae aniŋ.
Let me go first.
Sailus: *Aka ... aka mokoinay. Tawa pokombakanak.*
He won't ... he won't touch you. I'll break his bones.

20

Now please watch http://youtu.be/14rblIjE1EA [2:41].

While wandering around the store and looking at articles such as gum boots – which he does not know what to call in *kay menda*, and therefore uses a descriptive term 'something for wandering around in swamps' (lines 1 and 2) – Sailus got a feeling that the men behind him were not just eye-shopping. In line 8 he boasts how he will bash the rascal's face, which is at the same time a warning for Tumak and me to be careful. Finding a small towel he wonders what it is, while Tumak's attention is with the alleged pickpockets. He then answers Sailus, attempting to remember an avoidance term for towel, but in the end uses the Tok Pisin word *tawel*. Later he corrects himself, using a descriptive term, 'something to clean your face with' (line 13). If people do not know or do not remember an already established *kay menda* term, they often try to create one on the spot, and in such cases they would frequently resort to description. However, Tumak is alert and anxious because of the alleged rascals and he suggests that we leave. He calmly warns me that a (potential) rascal is following me and suggests that he goes ahead (line 20). Being nervous himself, Sailus boasts again, assuring me that he can protect me if somebody wanted to rob me (line 21). Having experienced some troubles themselves, and hearing stories about people being attacked and robbed, the Awiakay are always tense when in town. Many of them, particularly young boys, release this uneasiness by boasting how mean they will turn if anyone dares attack them. While this can be a meaningless, even jocular, everyday practice in the village (though also employed during fights), it becomes a means of reassuring one another when in town.

We were just about to leave the store when Asuk and Pupi, who had finished their business and were already looking for us, came in. As a group of five we were a less attractive target for the robbers or pickpockets, so we stayed there to take a look at tapes with popular music, torches and knives. Tumak and I were looking for a lamp for his *pap* 'maternal uncle' (and my 'father') Aymakan.

KAY MENDA – FILM 6 transcript

shop assistant: *Bilong disla bateri save stap insait. Narapela kain em i stap long hapsait long narapela glas.*
These ones have batteries inside. The other ones are over there.
Yu miniim wanemplakain?
What kind do you want?

Tumak: *Narapela. Glas tasol em i go daun olsem… Disla em nogat?*
Another kind where the lamp folds down. You don't have them?

Darja: *Em olsem bikpela lait liklik. Em nogat?* 5
With the slightly bigger lamp. You don't have it?

shop assistant: *Nogat.*
No.

Tumak: *Kay **yawaŋ** wakanjiŋ aninaŋ kolokot kaykoy salimbapoplaka.*
We'll look for it in another house [store], they sell different ones [there].
Elaŋ tok ton kak agalon.
Here they don't have it.
Aunda wakanjiŋ.
We'll keep looking.

Darja: *Kay kolokot yawa.* 10
Another house of things [store].

Tumak: *Mm. Kay kolokot yawa.*
M-hm [agrees]. Another house of things [store].
Kalak kay kon tola.
There are different things here.

Darja: ***Paypmaŋga kandeŋge**?*
Big stones [is it expensive]?

Tumak: ***Paypmaŋga wamoŋan.***
The stones have gone upriver [the price has gone up].

Darja: ***A, wamoŋan?*** 15
Ah? Gone upriver [gone up]?

Tumak: ***Wamoŋan.***
Gone upriver [gone up].

Darja: *Yo.*
Yes.

Tumak: *K 39.90 kak. Akanja. **Wamoŋan**.*
This is 39.90 Kina. Crap. It's gone upriver [the price has gone up].

Darja: ***Kondamin isapasa****?*
By two sticks [By 20 Kina]?
Tumak: ***Kondamin isapasa*** ... 20
By two sticks [By 20 Kina] ...
Kay... kay ***yawaŋ*** *wakanjiŋ aninaŋ.*
Another... we'll go and search for it in another house [store].
OK, kay ***ya****... Angoram wakanjinaŋ.*
OK, we'll go and search for it in another h[ouse]... in Angoram.

Now please watch http://youtu.be/tNIxwOv9z_0 [1:20].

As it turns out that they do not have the kind of lamps that Aymakan asked for, Tumak suggests that we search in another store, for which he uses a shortened version of *kay menda* term, *yawa* 'house', instead of *kolokot yawa* 'house of things'(line 7). People tend to shorten words in ordinary Awiakay all the time, and this practice is sometimes applied to *kay menda* as well.

Tumak then looks at other lamps and torches they sell in this store and says that the prices went up since he was last in the town a few months ago. For the price going up he uses the verb *wam-*, which originally means 'go/come up' in the sense 'in the upriver direction' or 'up to the house' (but not 'up to the mountain'). As Aymakan expected that the lamp he wanted would cost around K20, I ask whether the price was doubled, and Tumak confirms that it went up by 'two sticks', one tree stick equalling K10. This term comes from the colonial days in PNG when the first money was introduced to the Awiakay and they devised their own naming system for the coins and notes.

Most registers are not "sociologically homogeneous formations" (Agha 2004: 38), which means that not everyone is equally competent in them, and Awiakay 'hidden talk' is no exception. While every Awiakay person can speak at least a little bit of *kay menda*, the most competent speakers are the men who travel most frequently to town. However, we could see that even in this group the level of fluency varies and depends on several factors, not excluding an individual speaker's skills such as cunning, which is an essential part of 'hidden talk'.[12]

At the moment *kay menda* is still in the making and we can witness its on-going development. Women, who normally stay in the village, do not have many chances of using *kay menda* in practice; however, many of them take an active part in creating it. With a huge influx of material goods from Indonesia, shops in town are full of items previously unknown to the Awiakay, which means that they borrow terms for them from Tok Pisin. When such an item is brought to the village, its form and function is eagerly studied and discussed, and sooner or later somebody comes up with an Awiakay term for it. It takes some time before the speakers adopt such a term or create a new one, which they find more appropriate. The usage of a number of terms varies, and one can either (a) use the same avoidance term for several different Tok Pisin expressions, e.g. (TP) *kemera* 'camera', (TP)

[12] Having the ability to skilfully deceive other people is highly valued by the Awiakay.

skrin 'television screen', (TP) *gras* 'mirror' are all referred to as *memek* 'lightning' in *kay menda*, or (b) use different *kay menda* expressions for the same thing, e.g., *map kulamba yomba* 'water from ground hole' or *payp kulamba yomba* 'water from stone shelter' for 'shower'. The latter usually happens when a term has not been adopted by all speakers.

5. Continuity and change in Awiakay linguistic registers. I would argue that *kay menda* as it is spoken today in town is not a completely new register, but a continuation of the *kay menda* which used to be spoken in the mountains. There appear to be many functional and social, as well as some structural/linguistic similarities in their use. There is, however, no overlap in vocabulary, as 'mountain talk' used to 'hide' the meaning of Awiakay terms denoting people's immediate environment, while 'hidden talk' creates avoidance terms for Tok Pisin borrowings denoting recently introduced items and concepts.

Parallels between the two varieties of *kay menda*	'mountain talk' past; nowadays obsolete	'hidden talk' present; register in the making
used in unfamiliar territory; far from the village or camps	mountains (inhabited by spirits)	Wewak, all stops on the way there where the Awiakay encounter unfamiliar people
people go there to get something they need	*kanuŋ isa* (wood for bows), hunting (nowadays harvesting eaglewood)	selling eaglewood, buying goods
dangerous entities	*endembaŋ* 'mountain spirits'	*emay,* 'assault sorcerer' = rascals, pickpockets
possible dangers	sickness caused by spirits, getting lost, death	robbery, theft, physical injury
prevention of dangers	possible by implementing the expected practices, i.e. using *kay menda,* nowadays praying	possible by using *kay menda* and praying
persons who engage in relevant social practices (going to mountains/town) and are proficient in this register	men	men
others familiar with the register	women, teenagers	women, teenagers
created by	'mountain spirit' taught people how to protect themselves	all Awiakay; in the making

Table 2. Parallels between the two varieties of *kay menda*

For instance, both varieties of *kay menda* employ descriptive terms for their substitutes, e.g. 'cassowary' becomes *tumanjiŋge,* 'the hairy one' in mountain talk, while 'store' becomes *kolokot yawa* 'things-house' in hidden talk. Both varieties are used when people venture into the 'unknown' territory, far away from the village or camps in order to get

something they need. The mountains, which are not empty, but are – just like the rest of Awiakay land – inhabited by spirits, are a place where men go hunting, get black palm for their bows and nowadays harvest eaglewood, while the town, with all the unfamiliar people they meet, is the place where the Awiakay sell their eaglewood and buy the goods they need. In both settings they may encounter dangerous entities – *endembaŋ* 'mountain spirits', or the rascals and pickpockets in the town – which may damage them or their possessions. In both cases the dangers can be prevented by using *kay menda*, just that due to the changed relationship with spirits 'mountain talk' is nowadays replaced by praying, while in the town prayer is only supplementary to 'hidden talk'. In both contexts it is men who venture to these faraway places and use *kay menda* there, while women, even if they accompany their husbands or brothers, stay behind – either in bush camps, waiting for the men to return from the mountain, or in Angoram, waiting for the men to return from Wewak. In both cases women and teenagers are nevertheless familiar with the register. While 'mountain talk' is seen as a gift a spirit gave to the people to protect themselves, it must originally have been a fairly conscious creation, in which people chose to modify certain elements of ordinary Awiakay in order to arrive at a different code (cf. Pawley 1992: 315 on Kalam 'pandanus language'). 'Hidden talk', however, is being continually and actively created by all Awiakay.

In some socio-linguistic contexts the introduction of new commercial trades leads to increased exposure to and use of regional languages and a decline of local languages. However, in this instance it has also created circumstances in which the local language has developed a new dimension. The eaglewood trade seems to have re-strengthened people's relationship with their land, which had otherwise been weakened.

FIGURE 5. Continuity and change in the varieties of *kay menda*

By interpreting their environment through the same cosmology, with their actions being strongly influenced by the Catholic charismatic movement (Telban 2008b), and by the rules and changes it brought, the Awiakay have transferred the same practice (namely a lexical substitution register) with a similar function (hiding the meaning in order to protect themselves from being harmed) from their mountains to a different social setting of town.

References

Agha, Asif. 2001. Register. In Alessandro Duranti (ed.), *Key terms in language and culture*, 212-215. Oxford: Blackwell.

Agha, Asif. 2004. Registers of Language. In Alessandro Duranti (ed.), *A companion to linguistic anthropology*, 23-45. Cambridge: Cambridge University Press.

Agha, Asif. 2007. *Language and social relations.* Cambridge & New York: Cambridge University Press.

Briggs, Charles L. & Richard Baumann. 1992. Genre, intertextuality and social power. *Journal of Linguistic Anthropology* 2(2). 131-172.

Feld, Steven & Bambi B. Schieffelin. 1979. Modes across Codes and Codes within Modes: A Sociolinguistic-Musical Analysis of Conversation, Sung-Texted-Weeping, and Stories in Bosavi, Papua New Guinea. Paper prepared for [the session: Communication in ritual and everyday life: The interpretation of ways of speaking. American Anthropological Association, Annual Meeting, 1979, Cincinnati.

Feld, Steven & Bambi B. Schieffelin. 1982. Hard words: A functional basis for Kaluli discourse. In Deborah Tannen (ed.), *Analyzing discourse: Text and talk* (Georgetown University Roundtable on Languages and Linguistics [GURT] 1981), 350-370. Washington DC: Georgetown University Press .

Franklin, Karl J. 1972. A ritual pandanus language of New Guinea. *Oceania* 43. 61-76.

Franklin, Karl J. & Roman Stefaniw. 1992. The 'pandanus languages' of the Southern Highlands Province, Papua New Guinea: a further report. In Tom Dutton (ed.), *Culture change, language change: Case studies from Melanesia* (Pacific Linguistics C 120), 1-6. Canberra: Pacific Linguistics.

Gunn, Brian et al. 2004. *Eaglewood in Papua New Guinea*. RMAP Working Paper No. 51.

Hoenigman, Darja. 2007. Language and Myth in Kanjimei, East Sepik Province, Papua New Guinea. Ljubljana: Institutum Studiorum Humanitatis, Ljubljana Graduate School of the Humanities MA thesis.

Laycock, Donald Clarence. 1977. Special languages in parts of the New Guinea area. In Stephen.A. Wurm (ed.), *New Guinea area languages and language study*, vol. 3, *Language, culture, society and the modern world.* Canberra: Pacific Linguistics C-40.

Laycock, Donald Clarence & Peter Mühlhäusler. 1990. Language engineering: special languages. In N.E. Collinge (ed.), *An encyclopaedia of language*, 843-853. London & New York: Routledge.

Pawley, Andrew. 1992. Kalam Pandanus Language: An old New Guinea experiment in language engineering. In Tom Dutton, Malcolm Ross & Darrel Tryon (eds) *The language game: papers in memory of Donald C. Laycock* (Pacific Linguistics C 110), 313-334. Canberra: Pacific Linguistics.

Stasch, Rupert. 2002. Joking avoidance: a Korowai pragmatics of being two. *American Ethnologist* 29 (2). 335-365.

Stasch, Rupert. 2008. Referent-wrecking in Korowai: A New Guinea abuse register as ethnosemiotic protest. *Language in Society* 37(1). 1-25.

Telban, Borut. 2008a. The Poetics of the Crocodile: Changing Cultural Perspectives in Ambonwari. *Oceania* 78 (2). 217-235.

Telban, Borut. 2008b. Modification of perception in a Sepik community. Paper presented at the 7th Conference of the European Society for Oceanists, Verona 10-12 July 2008.

Telban, Borut. 2009. A struggle with spirits: Hierarchy, rituals and charismatic movement in a Sepik community. Pamela J. Stewart & Andrew Strathern (eds), *Religious and ritual change: Cosmologies and histories* (Ritual Studies Monograph Series), 133-158. Durham, NC: Carolina Academic Press.

Telban, Borut & Daniela Vávrová. 2010. Places and spirits in a Sepik society. *The Asia Pacific Journal of Anthropology*. 11(1). 17-33.

Darja Hoenigman
darja.hoenigman@anu.edu.au

Appendix: GLOSSARY OF *KAY MENDA* TERMS

English	Tok Pisin loan	*kay menda*	gloss
44 gallon drum	fotifo	*yomoy*	bucket made of a big bamboo
airplane	balus	*naim tandoŋga*	eagle-canoe
amount of money	hamas moni	*paypmaŋga mimbia*	name of money
axe	tamiok	*mundum*	stone axe
bag	bek	*yambam*	grass basket
ball	bal	*papukay maŋga*	orange tree fruit
ball	bal	*yupim*	wild pandanus ball
balloon	balun	*mumba*	bladder
bank	benk	*paypmaŋga yawa*	house of stones
basin	bikpela dis	*yakaopay*	earthen dish (large)
beer	bia	*o yomba*	water mixed with bark ashes (traditionally made salt)
big dish, boat	dis	*mondaŋ*	dish made of the soft part of the Arecoid palm (*Rhopaloblaste sp.*) petiole
big sturdy bag	renbo bag	*yambam*	basket
book, paper, anything for reading	buk, niuspepa	*kasaŋga*	dry banana leaf
bra	susu kalabus/bra	*isik ulakaplakay*	[something that] covers breasts
bullet	bulit	*tasia tamanda*	spirit arrow
bush knife, machete	busnaip	*malay eŋgaya*	sago machete
buy	baim	*e-*	take
camera television screen mirror	kemera skrin gras	*memek*	lightning
candle	kendol	*yandom endia*	tree sap
cap	kep	*koponuŋ tia*	head skin
chewing gum	big boy / P.K.	*kamba endia*	breadfruit sap
chocolate cream	soklit krim	*eneŋ mola*	diarrhoea /rotting shit
cigarette (bought in town)	Spia, Pal Mal	*kandukya yakia*	white [man] cigarette
clothes things	klos ol samtink	*kolokot*	things
computer	bikpela (save) man	*kandeŋ olukunja*	big man / person

English	Tok Pisin loan	*kay menda*	gloss
cup, mug	kap	*palendem*	coconut shell
cup, mug	kap	*wauna*	carnivorous plant (*Nepenthes ampullaria sp.*)
eaglewood	garu	*is kamia*	tree meat
Eucharist	yukarist	*pamben*	a kind of a nut
firelighter	masis	*pat*	stick for making fire
fishing hook	huk	*tao*	sago thorn
fishing net	net	*ewey*	net made of bark rope
frying pan		*epay*	earthen 'frying pan' or flat stone used for cooking sago
gaol	kalabus	*wanday yawa*	chickens' house
glasses, sunglasses, diving goggles	(ai) glas, gogols	*nokomgunuŋ tia*	eyelids
gold	gol	*kiŋim*	sand
guitar	gita	*tasia punjimba*	spirit hand drum
gum boots	gam but	*andaŋguŋ yaka yamblakay*	[something with which to] walk in the swamps
gun	gan	*tasia kanuŋga*	spirit bow
gun	gan	*yambuŋ kunda*	tree species having buttress roots (buttress roots can be kicked with the heel or struck with an ax or other tool to make a gun-like booming sound)
hard biscuits	biskit	*tasia taya*	spirit sago
house with a tin roof / town house	haus kapa	*tasia yawa*	spirit-house
instant noodles	nudols	*kundam en(eŋ)ga*	earthworm shit
iron post	ain	*makam*	main post in a house
K10	ten kina	*isapasa*	stick
knife	naip	*yombay (kapaya)*	bamboo (small knife)
lamp	lem	*yambat*	sago stem torch
learned man	saveman/ savemeri	*nokomga pawi*	red-eyed
lighter / torch / lamp	masis / tos / lem	*tasia yamba*	spirit fire
loudspeaker	spika	*tepuŋ*	bamboo / wooden 'loudspeaker'
marble	marbol	*imaŋ maŋga*	tree nut
medicines	marasin	*tasia pamyamba*	spirit ginger

English	Tok Pisin loan	*kay menda*	gloss
money	mani	*payp maŋga / payp manda*	stone
mosquito net	taunam	*aiŋ*	basket for sleeping
necklace	neklis	*tokombonoŋ tia*	neck skin
oat/nut/dried fruit bar	??	*koña taya*	honey(comb)-sago
oil / gear oil	wel / giawel	*tomba / tasia tombaya*	oil of native tree *Campnosperma brevipetiolata* (TP wel diwai)
outboard motor	moto	*tasia monaŋga*	spirit-paddle
outboard motor	moto	*wao ayma*	beetle family *Rhinchophoridae* (sago beetle)
paint (for grass/ sago)	pen	*kunakumbuŋ*	leaves for producing paint for *kuna*
pencil	pensil	*kaway tiñiplakay*	paint drawing
pencil / biro	pensil / bairo	*yambao*	ember
petrol / kerosene / beer / soft drinks	petrol / kerosin / bia / sop drink	*yom*	water
pillow	pilo	*tasia kumunda*	spirit wooden pillow
plate	pleit	*tane*	earthen plate
policeman	polis	*tam*	dog
pot		*aŋgas*	earthen pot
powder milk	Sunshine	*isik (yomba)*	breast (milk)
price went up	prais em i go antap	*paypmaŋga wamoŋan*	stones went upriver
radio	redio	*emuŋ kunda*	buttress roots
radio car	redio kar	*yomgoŋ manda*	turtle shell
rascal, bandit	sanguma	*emay*	assault sorcerer
rice	rais	*kauŋwa waya*	seeds of Arecoid palm *Rhopaloblaste sp.*
rope nail	rop nil	*awam*	vine/ thorns of a vine
rubber gloves	glav	*kolonoŋ tia*	hand skin
salt	sol	*tasia oua*	spirit 'salt'
serving spoon	kumu spun	*ipikapa*	halved coconut shell for pressing sago flour
serving tongs	??	*kula*	bamboo tongs (for holding hot items or ritual use)

English	Tok Pisin loan	*kay menda*	gloss
ship	sip	*mondaŋ kandeŋge*	big dish dish made of the soft part of the Arecoid palm (*Rhopaloblaste sp.*) petiole
shoes	su	*panben tia*	leg skin
shovel	sawel	*siŋgayaŋ*	??
shower	sawa	*map kulamba yomba* / *payp kulamba yomba*	water from ground hole / ~ stone shelter
soap body spray	sop bodi sprei	*tomba*	tree (oil)
soap	sop	*yom eŋay*	water spit
soap	sop	*yom karay*	water foam
spoon	spun	*kap*	spoon made of coconut shell
store	stoa	*kolokot yawa*	house of things
string	string	*pipisimba*	pandanus string
sugar lollies, candies	suga loli	*imat*/*tasia imata*	sugar cane/spirit sugar cane
sunglesses	sanglas	*tem nokomga*	sun-eye
tabernacle	tabernakol	*yao*	house
telephone	telipon	*tasia umbuŋga*	spirit *garamut* (TP for 'slit-drum')
tin roof	kapa	*wakŋga*	sago thatch shingles
tinned (mushroom) sauce	??	*meŋgwak*	vomit
toilet	toilet	*eneŋ yawa*	shit house
trousers	trausis	*kumbayn tia*	tree bark skin
trousers	trausis	*wasipi tia* [> *wai pia* 'part of a torn string bag']	string bag skin
T-shirt	singlis	*omunuŋ tia*	body skin
umbrella	ambrela	*ayŋgwaŋ tia*	skin of a flying fox
umbrella	ambrela	*embum*	grass hood
wallet	hanpaus	*kundambi*	coconut fibre for storing tobacco
watch	hanwas	*tem maŋga*	sun/time 'fruit'
write	raitim	*ikak-*	carve

9

Language Documentation & Conservation Special Publication No. 5 (December 2012)
Melanesian languages on the edge of Asia: Challenges for the 21st century,
ed. by Nicholas Evans and Marian Klamer, pp. 219–238
http://nflrc.hawaii.edu/ldc/sp05/
http://hdl.handle.net/10125/4566

Cross-cultural differences in representations and routines for exact number

Michael C. Frank

Stanford University

The relationship between language and thought has been a focus of persistent interest and controversy in cognitive science. Although debates about this issue have occurred in many domains, number is an ideal case study of this relationship because the details (and even the existence) of exact numeral systems vary widely across languages and cultures. In this article I describe how cross-linguistic and cross-cultural diversity—in Amazonia, Melanesia, and around the world—gives us insight into how systems for representing exact quantities affect speakers' numerical cognition. This body of evidence supports the perspective that numerals provide representations for storing and manipulating quantity information. In addition, the differing structure of quantity representations across cultures can lead to the invention of widely varied routines for numerical tasks like enumeration and arithmetic.

1. Introduction.[1] The relationship between language and thought is one of the most fascinating—and the most controversial—topics in cognitive science. Posed by Whorf (1956), the question of whether cross-linguistic differences lead to differences in cognition has been studied extensively across a wide range of domains. Recent work on this question has come from color perception (Kay, Berlin, Maffi & Merrifield 2003, Winawer et al. 2007, Roberson & Henley 2007), navigation and spatial language (Hermer & Spelke 1994, Levinson, Kita, Haun & Rasch 2002), theory of mind (Pyers & Senghas 2009), gender

[1] Thanks to Mark Donohue for encouraging me to visit Manokwari, Indonesia. I gratefully acknowledge all of my collaborators in the work reported here, including Ted Gibson, Evelina Fedorenko, Rebecca Saxe, Dan Everett, and David Barner. Thanks also to Susan Carey and Lera Boroditsky for valuable discussion of the theoretical ideas presented here. Finally, thanks to David Barner, Nick Evans, Ev Fedorenko, Ted Gibson, and two anonymous reviewers for giving comments on a previous version of this manuscript.

(Boroditsky, Schmidt & Phillips 2003), event perception (Papafragou, Hulbert & Trueswell 2008, Fausey & Boroditsky 2011), object individuation (Lucy 1992, Barner, Li & Snedeker 2010), categorization (Lupyan, Rakison & McClelland 2007), and many others. Yet despite considerable empirical progress, the general form of the relationship between language and thought remains hotly contested (Davidoff, Davies & Roberson 1999, Gentner & Goldin-Meadow 2003, Gumperz & Levinson 1996, Levinson et al. 2002, Li & Gleitman 2002, Pinker 1994).

Numerical cognition—and specifically, the use of language to represent large, exact quantities—is an exciting case study of this relationship in a domain that is both cognitively central and at the core of many technical achievements. Although there has been considerable discussion of the role of grammatical number marking as a case study of language and thought (e.g. Barner et al. 2010), the ability to represent arbitrarily large, exact numbers may have somewhat larger cultural and technical consequences. Hence, this review will cover only conventionalized representations that are suitable for representing large quantities—numbers like "seven" or "thirty-four"—and the routines that allow us to use them.[2]

The goal of the review is to give a sketch of some cross-cultural evidence on the relationship between numerical representations and routines. Rather than attempting to perform a comprehensive review of ethnographic evidence, I will instead focus primarily on recent psychological work that uses experimental methods in the field. Although there is tremendous value in linguistic and ethnographic work on number—and I discuss some in the final sections—my hope is to highlight how cross-cultural experiments can sharpen hypotheses about the relationship between language and thought by providing measurements of behavior in situations where numerical representations vary.

The outline of the review is as follows. I begin by describing background on representations and routines for number. I then present studies on numerical cognition in the absence of linguistic representations of numbers (evidence from Amazonian languages) and cases where language for number is culturally available but either not available to individual speakers (in Nicaraguan signers and home-signers) or not available online (in the moment in which a task is being performed). This body of evidence supports the idea that storing and manipulating exact quantity information depends on having both a representation of quantity and a routine for the appropriate task available in the moment when they are needed. I finish by surveying some examples of how number representations can vary due to cultural demands (examples from Melanesia) and how routines can vary depending on the structure of the representations they operate over (focusing on mental abacus users in India).

Taken together, the evidence supports a view that my collaborators and I have referred to as the "cognitive technology" view (Frank, Everett, Fedorenko & Gibson 2008, Frank,

[2] The term "number" is generally ambiguous between grammatical markings like singular/plural and numerals that describe the exact cardinality of sets. Here I will avoid the cumbersome language necessary to disambiguate in every instance and use the terms "numbers" and "numerical cognition" under the assumption that these terms refer to numerals representing the exact cardinalities of large sets and the broad range of cognitive operations that are carried out with such sets, respectively.

Fedorenko, Lai, Saxe & Gibson 2012): that numerical representations are cultural artifacts that are used for the online encoding of quantity information. The form of a linguistic or cultural representation of number and the efficiency of the routines for manipulating this representation each affect what computations are possible using this representation; the online availability of this representation (in the moment a computation is desired) is a prerequisite for performing the computation. One version of this view was first articulated by Kay and Kempton (1984) and it and its variants are currently experiencing a resurgence in cognitive science (Dessalegn & Landau 2008, Gentner 2003, Wiese 2007); see e.g. Frank et al. (2012) for more detailed discussion.

A secondary goal of this review is to argue for an approach whereby fieldworkers supplement standard elicitation techniques with psychological experimentation that tests the cognitive consequences of different numerical representations and routines. Because of the immense linguistic and cultural diversity in regions like Amazonia and Melanesia and the relative isolation of these populations, investigation of numerical systems in these regions' indigenous cultures provides especially rich evidence regarding the range of variation in number systems. Melanesia, in particular, is likely to harbour the greatest diversity of number systems in the world (Lean 1992). Ethnographic observation and psychological observation can play complementary roles in characterizing this diversity, providing both naturalistic observations and precise and generalizable measurements. And given the rapid decreases in linguistic diversity in these regions (Evans 2009a), it is especially important to document not only the facts of languages in Amazonia and Melanesia, but also the psychological consequences of these languages for their speakers.

2. Representations and routines for number. The past twenty years have seen an explosion of interest in representations of exact number as an example of an important, uniquely human concept, yet one that is built out of primitive components that can each be observed in infants and members of other species (Dehaene 1997, Carey 2009). On the one hand, numbers are a key part of every modern society: they facilitate a huge set of human behaviors, from complex feats of engineering to economic exchanges using currency. On the other, representations of quantity information can be observed in infants, monkeys, fish, and a host of other creatures (Gallistel 1993, Xu & Spelke 2000, Hauser et al. 2003). Thus, in the domain of number, cognitive scientists can ask how basic cognitive abilities can be combined into a sophisticated conceptual system and, in particular, what role language plays in this combination.

The basic cognitive systems that provide non-verbal representations of quantity are now well established (Feigenson, Dehaene & Spelke 2004). The first is a system that can track the location and identity of up to three or four objects at a time, likely based in visual attention or tracking. The second is the approximate number system (ANS), which can represent the approximate magnitude of sets of objects but not the identities of individuals within these sets. Despite the presence of both of these systems in prelinguistic infants, learning how to use linguistic numerals is a protracted process. In typically-developing English-speaking children, the time period from learning the meaning of "one" to mastering the use of number words up to "ten" can last a year or more (Wynn 1990).

Despite consensus about the basic facts, the role of language is contested in both this developmental progression and its end result. On the "bootstrapping" account, learning the meanings of numerals in the count list is a result of first mapping number words from

"one" up to "three" or "four" onto small number representations, and then performing an inductive step that recognizes the parallel between the sequential relationship between the words in the count list and the sequential relationship inherent in their definitions. The specifics of language—both in the structure of the count list and in the use of number names as placeholders for concepts—play an essential role in this account (Carey 2009, Piantadosi, Tenenbaum & Goodman 2012). In contrast, the "mapping" view suggests that words like "four" or "seven" are defined in terms of innate number concepts, and identified either noisily, using the ANS, or precisely, using a count routine. On this kind of account, language plays a peripheral role: it does not help to create new concepts, it simply helps to name and recognize pre-existing concepts by using enumeration routines like counting (Gelman & Gallistel 1978).

One broad area of agreement between these views, however, is the distinction between numerical representations and numerical routines, and the importance of their interaction in allowing their users to store and manipulate exact quantities (Gelman & Butterworth 2005, Carey 2009). By numerical representation, I mean here a set of symbols used for the task of representing exact quantities. The choice of a representation of number includes the medium of representation (linguistic, like a count list; externalized, like a counting stick; or even supported by visual imagery, like a mental abacus representation) and the internal structure of these representations (e.g. that English speakers say "ninety-nine" = 90 + 9 to mean 99, while French speakers say "quatre-vingt-dix-neuf" = 4 * 20 + 10 + 9). By numerical routine, I mean an algorithm that is commonly used to leverage such a representation in a particular numerical task. Examples of routines range from simple enumeration to the complex sets of steps that schoolchildren are taught to follow in order to perform addition or division of large quantities.

3. Numerical abilities without representations of exact number. What is numerical cognition like in the absence of linguistic numerals in a language?[3] Are there any routines for manipulation of exact quantity that are possible in the absence of exact numerical representations? This section reviews recent work with the Mundurukú and Pirahã, two indigenous groups in Brazil, that explores the cognitive consequences of speaking a language with limited or no vocabulary for exact quantities.

3.1. Measuring number vocabulary. Gordon (2004) claimed that Pirahã had a counting system consisting of words for the quantities 1 (*hói*) and 2 (*hoí*) as well as a word for "many" (*aibaagi*).[4] He reported data from only a single elicitation (in which a speaker

[3] The question of what it means to have exact numerals in a language is ambiguous: an individual speaker can in principle have access to a particular, idiosyncratic mapping between symbols and quantities; or a mapping can be conventionalized and available to many or all speakers of a language. Although there are cases of idiosyncratic or heterogeneous number systems (for preliminary data on this issue, see e.g. Frank & Honeyman 2011), the examples discussed here all show relatively broad consensus across speakers, shown via experimental procedures used with a sample of individuals from the community.

[4] Here and throughout the article I will use the Arabic numerals as a shorthand for the expression "the quantity N" regardless of whether the quantity is large or small, rather than following standard typographical conventions ("one" vs. 11) depending on quantity. I will quote numbers like "seven" to refer to a word for a quantity.

used the "two" word *hoí* to refer to the quantities 3 and 4). These data were broadly in accordance with a description of Pirahã as a "one, two, many" language, a type found in other non-industrialized societies (Menninger 1969, Hammarström 2010).

In their work on Mundurukú, Pica, Lemer, Izard, and Dehaene (2004) performed a structured elicitation experiment. They presented sets of 1–15 dots in random order to adults and children and asked how many dots were present in each set. Mundurukú participants responded consistently with a set of conventionalized terms for the quantities 1–3. These terms were used by participants in nearly all cases. For 4, participants used a conventional term almost as often, but occasionally used the same term to refer to 5 and 6. For 5, 25% of participants used a term meaning "one hand," while 35% of others used a vaguer term that Pica and colleagues translated as "some, not many" and that was used for other quantities 5–15 as well. Above 5, only this latter term and a term meaning "many" were used with any frequency. This experiment gives evidence that Mundurukú does have some exact numerals, but lacks a recursive number naming system and exact number vocabulary for large quantities.[5]

Following on Gordon (2004), our own work revealed a different view of Pirahã quantity vocabulary, using a structured elicitation task like Pica et al. (2004). We showed participants sets of objects and asked "how much/many are there?", increasing the cardinalities of the set from 1–10 and then decreasing from 10–1 (or vice versa). We found that the quantities for which our participants used particular words changed depending on the context of the elicitation (increasing vs. decreasing). In particular, although participants used *hói* only for 1 in the increasing context, they used it for up to 6 objects in the decreasing elicitation. This context effect strongly suggests that *hói* is not a word for 1. On our view, the most likely conclusion from these data is that it is a relative term like "few," "fewer," or even "small." Another possible position, however, is that *hói* is polysemous between "one" and "a few"; this view is of course logically possible, but provides no account of why or under what conditions an exact meaning would be available. The three words documented by Gordon are confirmed by several non-native Pirahã speakers to be the only words for quantities, leading us to conclude that Pirahã seems to have no (unambiguous) words for exact numbers: not even a word for 1.

The Amazonian findings suggest that representations of exact quantities are not a linguistic universal. In addition, they raise the intriguing question of whether any other languages without numerals have been misclassified as "one, two, many" languages due to the absence of experimental data.[6] In order to determine the semantics of possible numerals, single-participant elicitations should be replaced with structured elicitations and numeral comprehension tasks (Wynn 1990). Even data for a handful of participants in

[5] Note that for developmental researchers, the gold standard for children having acquired the meaning of a numeral for 7 is success in comprehension-based tasks like "give a number" (Wynn 1990, Le Corre et al. 2006, Condry & Spelke 2008). In the "give a number" task, participants are simply asked to "give me N objects" and the cardinality of the set they give is reported. Neither the Mundurukú nor the Pirahã have been tested on such a task, so more work remains to be done to probe the meanings of the attested vocabulary items.

[6] Hammarström (2010) gives a list of other languages that have such systems and notes this possibility, though Pirahã may be the only one of these that lacks any singular-plural marking as well.

such tasks can be informative and can provide an inexpensive supplement to current field methods.

3.2 Consequences of limited number vocabulary. In contrast with linguistic representations of number, which vary across societies, a large body of evidence shows that an approximate number sense (ANS) is available to all human beings as well as members of other species. This approximate sense leads us to be able to make estimates of a set's quantities without using an enumeration routine.

The ANS has been characterized extensively in human and non-human animals (for review see Feigenson et al. 2004, Gallistel 1993). Estimates of quantity made by the ANS follow Weber's law (e.g. Whalen, Gallistel & Gelman 1999, Xu 2002), which states that the probability of a correct response in a discrimination task is related to the magnitude of the stimulus being discriminated. Weber's law leads to the prediction of the relation $\sigma/\mu = c$ in participants' data, where μ and σ are the mean and standard deviation of the magnitude estimates (across trials or participants) and c is a constant holding across a range of magnitudes. The term c is often referred to as the *coefficient of variation* or COV. A constant COV implies that the larger the quantity being estimated, the larger the average error, in turn signaling that the ANS is being used.

In Pica et al.'s study, Mundurukú participants and French controls performed comparison, addition, and subtraction tasks. When participants were asked to choose the larger of two large sets of dots (and were not given enough time to count), both groups performed similarly, showing a constant COV, consistent with Weber's law. However when participants were asked to give the resulting quantity in a subtraction paradigm where objects were first added to and then subtracted from an opaque container, French participants performed nearly perfectly, while the Mundurukú made errors that were again consistent with the operation of the ANS. Crucially, the design of this task required only responses in the range where the Mundurukú could have responded verbally (quantities 0–2), ruling out the explanation that they could not indicate the correct response even though they knew it.

Like the Mundurukú, the Pirahã also relied on the ANS to perform numerical tasks. Gordon (2004) performed a range of matching tasks designed to probe the ability of participants to store and manipulate exact quantities. In the simplest task, participants were asked to produce a 1–1 match between two sets by selecting the correct quantity of objects to align with a target set. In more difficult tasks, the target set was presented in a cluster or was presented only briefly, and participants were again asked to produce a target set of the same cardinality. Participants made errors in all tasks, even the 1–1 match task, although their errors were larger in those tasks where the target set was presented for a short period of time. When Gordon consolidated data across all tasks, the pattern of responses again showed a constant COV. Like the Mundurukú results, these findings suggest that analog estimation using the ANS is the default strategy in situations where no count list is available.

Both sets of results left open an important question, however: did Mundurukú and Pirahã participants understand that large quantities *could be* exact, even if they did not know how to express or manipulate them? For example, Gordon's 1–1 matching task was the simplest task in either assessment, yet Pirahã still made errors. Were these errors due to confusion about what was being asked or difficulties in completing the task, or were

they instead due to a more fundamental conceptual difference? On the first interpretation, the Pirahã made errors in matching up larger quantities of objects either because they did not understand that an exact response was called for (even though they could have produced such a response) or because they made manual errors in alignment even though they understood what was being asked of them. On the second interpretation, however, the Pirahã did not understand that a correct response required matching exactly, because they did not even have available a concept of exact equivalence.

The actual computational demands for success in the 1–1 matching task are quite low. In order to succeed, it is only necessary to match individuals until there are no more left to match. This task can be accomplished without ever representing the total quantity, so success in the task does not demonstrate the existence of exact quantity representations. A 1–1 match of exactly 7 items can be performed without ever mentally representing 7. On the other hand, a true failure in the task—an inability to select the 1–1 matching algorithm, even with appropriate training and unlimited time—would suggest that the Pirahã truly did not think in terms of exact equivalence or exact matches.

On a recent visit to the Pirahã, my collaborators and I replicated a number of Gordon's tasks with a larger sample of participants (N=14, as opposed to N=5 in the previous study). In order to ensure task understanding, we included a systematic training phase in which we demonstrated what the correct response would be for one trial with a small quantity and then gave corrective feedback on another set of small-quantity trials until participants were performing consistently (Frank et al. 2008). In the more difficult matching tasks, we found precisely the pattern of ANS usage that Gordon documented, with errors increasing along with the quantity of objects being estimated (see figure 1 for an example of the testing environment). Our results differed from Gordon's in the 1–1 matching task, however. There, only one participant made any errors and the rest performed perfectly, suggesting that this task was qualitatively different from the others. Despite not having linguistic representations of exact quantities available to them, this group of Pirahã understood that an exact response was required. This result shows that our participants made the appropriate generalization from a few training examples with small numbers: that every target item should be matched with *exactly* one item, not that the two sets should match approximately. That they made this generalization consistently across individuals strongly suggests that the notion of an exact, rather than approximate, 1–1 match was available to them (though again, not the representation of a particular exact quantity like 7).

One final dataset bears on this question, however. Everett and Madora (2012) conducted a replication of our previous work with another group of Pirahã from a different village. Although they again replicated the pattern of ANS usage on more complex matching tasks, they found results congruent with Gordon's: their participants made systematic errors on the 1–1 matching tasks. Everett and Madora argued that the success of the particular participants in our 2008 experiments was due to exposure that members of this village had to innovated number words and numerical procedures. Apparently, Madora had conducted numerical training sessions with the members of this village; nevertheless, our elicitation tasks showed no evidence for knowledge of innovated number words. This claim brings up an interesting possibility: could it be that exposure to some representations of exact number— even without the long-term adoption of these representations—facilitates the construction of a 1–1 match strategy? Although the current data do not provide enough

information to evaluate this claim, perhaps it can be assessed via future developmental or cross-cultural work.

FIGURE 1. A Pirahã participant in Frank et al. (2008), in the orthogonal match condition. The experimenters have placed 10 spools of thread, and the participant has matched them with 9 balloons.

To summarize, evidence from the Pirahã and Mundurukú demonstrates that in cultures without representations of large exact quantities, individuals are not able to remember or manipulate such quantities exactly, suggesting a connection between linguistic representations and the ability to create routines for manipulating exact number. Instead of remembering exact quantities, both groups used an estimation strategy which allowed for approximately correct responses even in relatively difficult tasks. Nevertheless, evidence from the Pirahã suggests that it is possible to create and use a routine for exact, 1–1 match even without an unambiguous linguistic representation of 1.

4. DISTINGUISHING COGNITION FROM CULTURAL EXPOSURE IN NUMBER REPRESENTATION. The evidence above suggests that routines for storing and manipulating exact quantities correlate with the cultural presence of linguistic representations of number, but the precise nature of this correlation is unknown. One possibility is that language for number could simply co-occur with cultural routines for number, rather than being a causal factor in the cognition of individual speakers. On this kind of account, language for number would develop alongside a set of (possibly non-verbal) routines for manipulating exact quantities, springing from the same basic cultural needs. Speakers would learn number words, but they would also learn algorithms for doing matching tasks, for chunking large quantities

into sets of smaller quantities, and for tallying to keep track of quantities over time. For example, the use of an abacus would constitute a parallel, non-linguistic routine that could support numerical calculation (see below for more details). On the other hand, another possibility is that language for number could be necessary in the moment for the precise manipulation of exact quantities: that is, language could be a necessary constituent in these routines (like in the case of verbal arithmetic, but unlike in the case of an abacus).

Recent studies have begun to differentiate between these two accounts. First, work with signers in Nicaragua has investigated the numerical abilities of individuals in a highly numerate culture who nonetheless have limited representations of exact number and limited routines for manipulating these representations. Second, psychophysical experimentation with verbal interference tasks has begun to manipulate the online availability of linguistic representations of exact number in highly numerate, educated adults. These two sets of studies are reviewed below.

4.1. Cultural exposure alone does not scaffold exact number. Nicaraguan Sign Language (NSL) is a new sign language created over the last 30 years as specialized schools have brought together the community of deaf individuals in Nicaragua (Senghas, Kita & Ozyurek 2004). As the Nicaraguan deaf community has grown and the age at which children are exposed has become younger, NSL has evolved into a fully-featured, highly grammaticized language that includes number words, complex spatial language (Senghas & Coppola 2001) and sophisticated constructions for reporting the thoughts of others (Pyers & Senghas 2009).

Since NSL speakers live in a numerate community, playing gambling games and using money, they have ample opportunities to acquire numerical routines. Nevertheless, number signs in NSL underwent rapid standardization in the early 1990s, transforming from iconic finger signs—with a number of fingers corresponding to the quantity being indicated—to a set of simpler, one-handed signs that are less iconic. This change has created a population of speakers with a range of experience with numbers signs: there are older adults who did not learn either system as children; younger adults who learned the iconic system but have since learned the second system; and adolescents who learned the second system during childhood (Flaherty & Senghas 2011). By keeping cultural exposure relatively constant but varying linguistic representation, the case of NSL thus presents an opportunity to test whether cultural exposure to numerical routines is sufficient for accurate performance of numerical tasks or whether it is necessary to have linguistic representations in order to acquire or carry out numerical routines.

Flaherty and Senghas (2011) tested NSL speakers across the full range of ages on a set of tasks that included matching tasks like those used by Gordon (2004) as well as tasks requiring tapping out quantities, counting and selecting sets using number words, and translating between monetary notes and coins. Across all tasks, the group that made far and away the most errors were the older adults that had not fully mastered even the iconic count lists. Individuals who had mastered either count list made small but systematic errors—indicating that they were not perfectly accurate in using their count routine in challenging situations— but the performance of older adults who could not count differed significantly from even that of the older adults who had been able to master the iconic count routine.

In addition, as with the Pirahã, all NSL participants—even the non-counters—succeeded in grasping the simplest 1–1 matching tasks. When matching tasks became more

complex and the stimuli being matched were presented ephemerally (via tapping, or via putting items one by one into an opaque cup), accuracy was considerably lower for the non-counters. The non-counters knew that there was something they did not know, however—they expressed uncertainty about larger quantities, and had developed heuristic strategies for making change in the monetary tasks. They knew that an exact answer was required, but did not know how to calculate that answer. Thus, like the Pirahã, NSL speakers without a count routine were able to select an exact quantity matching strategy, even in the absence of a reliable method for mentally representing individual quantities.

Although many deaf children in Nicaragua are now given opportunities to learn NSL, there are still some individuals who have not had access to the broader deaf community and have instead built up more idiosyncratic sign systems for communicating with their families and more immediate community. "Homesign" systems of this sort and their relationship to conventional language have been studied extensively, in the US and around the world (Goldin-Meadow & Mylander 1984). Recent work by Spaepen, Coppola, Spelke, Carey and Goldin-Meadow (2011) investigates numerical cognition in Nicaraguan homesigners. Congruent with the work with NSL speakers, Spaepen and colleagues found that homesigners, who could not produce a consistent count list or perform matching tasks, were still able to compare monetary denominations with high accuracy.

In addition, although they could not produce a correct ordering of number signs, the homesigners did still know words for exact quantities. This knowledge allowed Spaepen and colleagues to perform an important exact numerosity recognition task. In this task, the homesigners were told that some exact number of objects were in a box, and then the array in the box was transformed (either via a change in the number of objects or not). When the transformation did not change the quantity in the box, the homesigners almost always used the same gesture as the experimenter; when the transformation did change the quantity, they never used the same gesture. Ruling out a pragmatic explanation for this behavior (e.g., applying the principle of contrast; Clark, 1988), nearly all participants used gestures that matched the direction of the transformation, for example signaling a larger number than the original gesture when an object had been added to the set. This task gives clear evidence that the homesigners understood that each set had an exact numerical value, even if they did not have an errorless routine for finding that value.

Although both NSL users and homesigners grew up in a highly numerate culture, this fact alone did not create the concepts and routines necessary to succeed in complex exact number tasks. In addition, supporting the Pirahã 1–1 matching results, the Nicaraguan data suggest that neither number words nor a count routine are necessary to understand the idea that a set has an exact quantity, even if that quantity cannot be named or stored in memory.

While the Nicaraguan data implicate linguistic representations (rather than cultural exposure to routines) as playing a causal role in the ability to manipulate exact quantities, it is a separate question whether this role is *online*. In other words, for an individual with a lifetime of practice representing exact quantities, does representing a quantity like 7 require the use of language in the moment such that if linguistic resources were not available at that moment, this task would become much more difficult or impossible? To answer this question, we turn to psychophysical tasks performed with numerate English speakers.

4.2. Number words must be available online for enumeration. Verbal interference methods have been used widely for testing the online dependence of various tasks on language (Newton & de Villiers 2007, Winawer et al. 2007, Hermer-Vazquez, Spelke & Katsnelson 1999). Verbal interference refers to a class of experimental paradigms in which participants are asked to perform a task while simultaneously occupying their verbal system by performing a separate verbal task, such as repeating a word like "the," repeating strings of numbers, or "shadowing" (repeating words after immediately after hearing them spoken on a recording). As a control for the generalized dual-task cost of performing two tasks at once (Pashler 1994), performance in the target task under verbal interference is often compared to performance in the target task paired with a non-verbal task like shadowing a clapped pattern.

A handful of studies have used verbal interference to measure numerical behavior. However, most have done so using number tasks that were themselves verbal. For example, Logie and Baddeley (1987) found that rapid repetition of "the" caused more errors in counting than either listening to speech or tapping a finger, suggesting that active speech production interfered with use of the same system to count. A more recent study by Cordes, Gelman, Gallistel and Whalen (2001) showed an Arabic numeral and asked participants to press a key that number of times while either repeating "the" or counting very quickly. They found that participants under verbal suppression showed a constant coefficient of variation—indicating use of the ANS—while those who were counting showed a decreasing COV (perhaps caused by the binomial errors implied by skipping numbers in the count list). These two studies give evidence that language interference does cause participants to make errors when aspects of the task are linguistic, but leaves open the possibility of better performance in purely non-linguistic tasks.

In order to test this possibility, my colleagues and I conducted a series of experiments where we replicated the matching tasks used with the Pirahã, performing these tasks with a group of English speakers who were simultaneously shadowing complex texts (Frank et al. 2012). This paradigm had the benefit of using a purely nonverbal measure of number knowledge and of providing data that could be compared directly to those collected during fieldwork with the Pirahã. Our results suggested strong parallels between the performance of the English speakers—who did not have number language available in the moment—and that of the Pirahã—who had never known words for numbers. Like the Pirahã (and Nicaraguan populations), the English speakers under verbal interference were able to do the 1–1 matching task with relatively few errors. In addition, the English speakers, like the other populations, showed evidence of relying on the ANS in the hardest matching tasks. Followup experiments using matched verbal and spatial memory interference tasks showed that this pattern was specific to language interference.

However, the English speakers also showed some differences from the Pirahã. In the medium-difficulty matching tasks where there were visual cues (e.g., matching the quantities of two orthogonal lines), they made errors but their overall performance did not show the signature of the ANS (a constant relationship between the quantities being matched and the magnitudes of the errors in estimation). Instead, the magnitude of the errors increased with respect to the quantity being matched. We posited that their errors resulted from the use of ad hoc matching routines like making correspondences between sub-groups of objects. This same pattern of increasing errors was observed in the Nicaraguan signers

who did know a count list, indicating that this pattern of data may generally result from the application of fallible routines.

More broadly, the picture that emerges from the evidence so far suggests an online, causal role for language in the representation of number information. Evidence gathered through psycholinguistic fieldwork, in combination with laboratory control tasks, suggests that representing 7 requires having some internal symbol like "seven" available in the moment. This pattern of evidence should not suggest that there is no role for cultural needs in the creation of numerical routines, however. The next section gives several ethnographic examples of interactions between culture and numerical representations and routines.

5. Numerical representations and routines can be shaped by culture. A common perspective on English numerals—even from sophisticated, numerate adults—is that they are transparent linguistic tools that do not reflect an idiosyncratic evolutionary process driven by specific cultural needs. Yet a closer look at the diversity of count systems in the world's languages falsifies this view. While the examples discussed below only provide an existence proof for cultural effects on representational systems, it is a goal for future research to understand both the prevalence of such effects and the mechanisms by which cultural demands can lead to representational innovations. For example, Wiese (2007) gives an account of how number concepts and numerals evolve in concert; her ideas leave open several places where specific cultural demands could lead to particular representational idiosyncrasies over the evolution of a count list. Thus, my hope is that discussing examples of possible links between culture and number representation can give some insight into how this relationship could function. I give three examples below.

First, the ways that numerals are named can change in response to the needs of individuals in a culture. For example, in Mangarevan, a language spoken on an island in French Polynesia, tools, breadfruit, and octopus are each counted with different sequences (Beller & Bender 2008). The Mangarevan language includes an abstract counting system that extends to high numbers, but it also includes three different systems for applying this list to different kinds of objects. These different systems rename the basic count unit to be groups of 2, 4, or 8 of an object, allowing for much more efficient grouping and counting of large numbers of objects. Beller and Bender argue that this division reflects a case in which a single number system has fragmented into a number of task-specific systems. Although each system incorporates properties of the more abstract count list, the need for greater efficiency and accuracy in specific situations led to the move away from a single, abstract system.

Second, the entire structure of a count system can be determined by a sufficiently important cultural practice. The vast majority of the world's count lists are structured around bases that are 5, 10, or 20 (Hammarström 2010), presumably because human beings have five digits on our hands and feet (and 20 digits overall). Base-5, base-10, and base-20 systems interact with and are supported by finger- and toe-counting routines. The languages of the Morehead-Maro region of Papua New Guinea have received considerable recent attention, however, because they are base-6, an extremely rare pattern (Donohue 2008, Hammarström 2009, Evans 2009b). Many of them include lexical items for relatively high exponents, e.g. up to 65 or 66 in Keraakie. Evans (2009b) and Hammarström (2009) give a compelling account of the origins of this system: it is specialized for the counting of yams, which can be arranged for storage in a petal-like configuration. In addition, in an interesting

twist, the base-6 representation leads to a reinterpretation of finger-counting routines: these routines become base-6 as well, using the wrist as a final location and re-construing the finger count as a count of "attachment points" (finger joints and wrist joint) (Evans 2009b). In this way, finger-counting is reinterpreted with respect to the base-6 representation that has evolved (or been invented) to support an important cultural routine.

Third, changes to a numerical representation can also be motivated directly through changes in cultural routines. A specific example of this kind of comes from Saxe (1982). He documented that speakers of Oksapmin, a language spoken in the West Sepik province of Papua New Guinea, used a body-count system (a common type in the region (Lean 1992)). This count system was base-27, extending from one hand along the arm, over the head, and through the other arm to the other hand. However, when users of this count system had limited experience with manipulating money, they made systematic errors in simple addition problems (e.g. 8 + 6). Saxe found that although they could count out the first addend (8), these inexperienced users had not developed a correspondence strategy so that they could keep track of the number of body parts in the second addend (6).

Users more experienced with money manipulation had developed a number of strategies to circumvent this problem, however, including counting both the second addend and the sum of the addends in parallel, and splitting the body in two and using the second arm to track the second addend. The body-splitting strategy was most successful; it was used by the participants that were most experienced with money manipulation, but also required the most adaptation of the existing representation. To use it, Saxe's participants had to reverse the count list so that it could be initiated from either arm. Saxe's study beautifully demonstrates how cultural pressures can lead to the creation of new routines for arithmetic and can in turn lead to changes in the base representation.

Body count systems also suggest how the choice of a base—or more generally the design of a number representation—can interfere with the development of efficient routines. In the case of Oksapmin, the base was so high that the enumeration routine required both hands and hence could not be easily used to create two separate buffers for addition. This example is minor, however, compared with counting systems like one reported to be used by some speakers of One. This system, described by Donohue (2008), is in principle recursive and infinite, but in practice so cumbersome that it is rarely used to count quantities larger than a handful. One has individual lexical items for 1 and 2, but allows specific, conventionalized combinations of these words up to 6. Their count list admits the following combinations 1, 2, 2+1, 2+2, 2+2+1, and (2+1)+(2+1), but not (for example) 2+2+2. Although this system could be used to express 7, 10, or even 20, it quickly becomes impractical for larger quantities. This system may even be a recent innovation and hence indicative of a community whose use of numbers is in flux (Crowther 2001).

These examples give a flavor for the ways in which the vast range of number representations and routines in the world's languages can be shaped by their cultural context. Nevertheless, understanding the specific cognitive consequences of this variation will require significant experimental fieldwork, and the form of the relationship between particular numerical representations and the routines they support is mostly unknown. The last section of this review gives some evidence on this question by exploring a case study of a number representation that licenses a very different set of routines—in a different medium—from the others we have reviewed: mental abacus.

6. Non-linguistic representations: Effects of representation structure on routine. This review began by asking about the relationship between language and thought and explored this relationship through the diversity in number representations across the world's languages. But in fact there is a wide variety of non-linguistic representations used across different cultures. Aside from finger-count systems, Menninger (1969) describes the near-universal use in the ancient world of tally sticks and knot-based systems to keep track of large quantities. In many cases, tally-based systems evolved into the use of counting boards—devices that allowed for the grouping of tally objects like pebbles. Unlike most tally systems, counting boards incorporated the use of place value (in which a particular position in a notation stands for the order of magnitude of symbols in that position, e.g. $1 = 10^1$, $10 = 10^2$, $100 = 10^3$), a major innovation that allowed them to be used flexibly for a wide variety of record keeping.

The modern Soroban abacus (primarily used in Japan and China) likely evolved from Roman counting boards. Like these counting boards, the Soroban abacus uses a base-10 representation with place values mapped to individual abacus columns. The abacus (and some of the most sophisticated counting boards) incorporates a subsidiary base-5 as well, however: on a standard Soroban each column has a single bead on top that represents 5 units in that place value, and four beads on the bottom that each represent 1 in that place value. Combinations of these beads allow the quantities 0–9 to be represented using a maximum of five beads.

Like tally sticks and other enumeration devices, counting boards and abacuses are external devices that allow their users to enumerate large exact quantities and retain them precisely over long periods of time. However, the abacus allows users to go beyond the simple enumeration task by allowing the development and use of efficient routines for arithmetic computation. Using the base-5 within-column representation and base-10 place value system, large computations can be broken into many small steps consisting of the addition of numbers below 5 and a corresponding set of "carry" operations (in which the parts of a result greater than 9 are transferred to the next highest place value). With practice, abacus calculations can become routinized and highly accurate. In a head-to-head competition in post-war Japan, a skilled abacus operator out-computed a calculator user (Kojima 1954). Crucially, abacus addition operates via a routine using set of memorized operations that are different from the commonly used base-10 addition operations.

Although abacus is an external computation aid, experienced abacus users can learn to internalize the abacus representation and make computations by manipulating beads on a mental image of an abacus. This technique, known as mental abacus (MA), is widely taught in Japan and has been the focus of recent interest in math supplementary education programs in Malaysia, India, China, and a number of other countries in Asia and the Middle East. Studies of MA have suggested that users do truly represent a mental abacus using visual imagery (Hatano 1977, Hatano & Osawa 1983, Stigler 1984). For example, they make off-by-5 errors far more than would be expected in standard linguistic calculation, indicating that they are inadvertently "dropping" the 5 bead from their mental representation. MA users also seem to be able to compute while performing linguistic distractor tasks (Hatano 1977) and neuroimaging studies confirm that MA activity induces activity in cortical areas related to visuo-spatial working memory (Tanaka et al. 2002, Chen et al. 2006). MA is also

highly effective as an arithmetic method: a MA user took top honors in the 2010 World Cup of Mental Computation.

Our own recent work investigated how it is that MA representations are possible, given the attested limits on numerical representation in the visual system (Frank & Barner 2011). Neither of the two systems traditionally implicated in visual number processing (object processing for small numbers up to 4, approximate representations above that) would be able to represent a number like 49 on the abacus, since this would require representing the exact positions of 9 beads. Despite this, MA users are able to do impressive computations with far larger numbers. For a visual comparison of abacus computation and MA, see figure 2.

FIGURE 2. (left) A child performing a physical abacus computation. (right) The same child performing a mental abacus computation.

We tested a large group of children (ages 7–16 years) in Gujarat, India, who were enrolled in MA afterschool programs. We asked these children to do two standard MA tasks—addition of quantities and translation of abacus configurations into Arabic numerals—while we varied the difficulty of the tasks. In both tasks, we found that the limitations on performance came from the number of columns on the abacus representation—in other words, the maximum place value—rather than on other features of a given task. For example, many participants were able to add between 7 and 9 two-digit addends together in under 10 seconds, but very few were able to add 2 four-digit addends in the same time period. In contrast, the number of beads necessary to make a representation (e.g., whether a column showed a number like 0, with 0 beads, or 9, with five beads) in these problems did not seem to affect performance, once the number of columns was controlled.

To test the dependence of MA computations on language, we asked MA experts to perform verbal interference tasks as they did mental computations. While their performance was impaired slightly by verbal shadowing, they were if anything more impaired by simply tapping their fingers during the computation (presumably due to the reliance of the computation on the accompanying gestures, see figure 2). In contrast, a group of American college students—who used linguistic calculation strategies to do mental arithmetic—were highly impaired by verbal interference but experienced no interference from tapping their fingers.

The interference data suggest that MA is a fundamentally visual representation, while performance in the addition experiments described above suggest that MA representations are column-based. We hypothesize that each column in the mental abacus is mapped to a separate object representation in visual working memory, though the substructure of how each column is represented is still unknown. Consistent with this hypothesis, we found that novice abacus users showed some of the same signatures of column-based organization, suggesting that this non-linguistic format for number was adapted to the general visual capacities of its users, rather than being the result of extensive practice. Taken together, these data paint a picture of MA as a visual alternative to linguistic number representations that relies on the distinct structure of visual working memory, rather than phonological working memory (as in language-based techniques for mental arithmetic).

The example of MA goes beyond external physical representations of number like counting boards and gives strong evidence that the mental representation of exact quantity is possible in mediums other than language. Although some authors have speculated that language and exact number rely on the same computational substrate (Hauser et al. 2003), the facility and flexibility in computation shown by MA users suggests a different view. Representations of exact number can be constructed using a variety of different resources—linguistic or visual. In addition, the specific organization of the abacus/MA representation is tailored to allow computations to be decomposed into many simple operations that can be practiced independently. Although the MA addition routine requires more steps than the most common verbal algorithm, it is also more accurate because it never requires storing partial sums.

7. Conclusions. Beller and Bender (2008) write that "there may be no other domain in the field of cognitive sciences where it is so obvious that language (i.e., the verbal numeration system) affects cognition (i.e., mental arithmetic)." The data reviewed here are consistent with this contention: how a language represents large exact quantities dramatically influences how its speakers are able to store and manipulate them. For this reason, number representation presents an important case to go beyond the first order questions of the Whorfian debate—"does language influence thought"—and ask detailed questions about how language participates in constructing representations of exact number and routines for manipulating quantities. Investigations of the richness of cross-cultural variation in number systems suggest that there are major behavioral consequences that correspond to what number words a language has and how those words are structured into a count list. More generally, the form of a numerical representation (linguistic or not) structures the kinds of routines for enumeration and arithmetic that can be performed.

The data that lead to this conclusion could not have been gathered by the standard methods of cognitive psychology, nor by the standard methods of field linguistics. Many of the results cited here come from carefully controlled studies performed in the field with populations that possess culturally, linguistically, or cognitively interesting numerical representations. This generalization suggests the benefits of psycholinguistic fieldwork that combines experimental design with cross-cultural or cross-linguistic populations. Such fieldwork is especially important in the study of the diverse languages of Melanesia, since opportunities to study these languages are quickly disappearing. Future fieldwork—on number and in other domains—should take advantage of these techniques to present a fuller picture of the relations between language, culture, and cognition.

References

Barner, D., P. Li & J. Snedeker. 2010. Words as windows to thought: The case of object representation. *Current Directions in Psychological Science* 19(3). 195–200.

Beller, S. & A. Bender. 2008. The limits of counting: Numerical cognition between evolution and culture. *Science* 319. 213–215.

Boroditsky, L., L. Schmidt & W. Phillips. 2003. Sex, syntax, and semantics. In D. Gentner & S. Goldin-Meadow (eds.), *Language in mind: Advances in the study of language and thought*, 61–79. Cambridge, MA: MIT Press.

Carey, S. 2009. *The origin of concepts*. New York, NY: Oxford University Press.

Chen, F., Z. Hu, X. Zhao, R. Wang, Z. Yang, X. Wang & X. Tang. 2006. Neural correlates of serial abacus mental calculation in children: A functional MRI study. *Neuroscience Letters* 403(1–2). 46–51.

Clark, E. 1988. On the logic of contrast. *Journal of Child Language* 15. 317–335.

Condry, K. & E. Spelke. 2008. The development of language and abstract concepts: The case of natural number. *Journal of Experimental Psychology General* 137(1). 22.

Cordes, S., R. Gelman, C.R. Gallistel & J. Whalen. 2001. Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin and Review* 8. 698–707.

Crowther, M. 2001. All the one language(s): Comparing linguistic and ethnographic definitions of language in new guinea. Sydney: University of Sydney Honours thesis.

Davidoff, J., I. Davies & D. Roberson. 1999. Colour categories in a stone-age tribe. *Nature* 398. 203–4.

Dehaene, S. 1997. *The number sense: How the mind creates mathematics.* New York: Oxford University Press.

Dessalegn, B. & B. Landau. 2008. More than meets the eye: The role of language in binding visual properties. *Psychological Science* 19(2). 189–195.

Donohue, M. 2008. Complexities with restricted numeral systems. *Linguistic Typology* 12(3). 423–429.

Evans, N. 2009a. *Dying words: Endangered languages and what they have to tell us.* Oxford: Blackwell.

Evans, N. 2009b. Two *pus* one makes thirteen: Senary numerals in the Morehead-Maro region. *Linguistic Typology* 13(2). 321–335.

Everett, C. & K. Madora. 2012. Quantity recognition among speakers of an anumeric language. *Cognitive Science* 36(1). 130–141.

Fausey, C. & L. Boroditsky. 2011. Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin & Review* 18(1). 150–157.

Feigenson, L., S. Dehaene & E. Spelke. 2004. Core systems of number. *Trends in Cognitive Sciences* 8. 307–314.

Flaherty, M. & A. Senghas. 2011. Numerosity and number signs in deaf Nicaraguan adults. *Cognition* 121(3). 427–436.

Frank, M. C. & D. Barner. 2011. Constructing exact visual representations of number. *Journal of Experimental Psychology: General* 141. 134–149.

Frank, M. C., D. L. Everett, E. Fedorenko & E. Gibson. 2008. Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition* 108. 819–824.
Frank, M. C., E. Fedorenko, P. Lai, R. Saxe & E. Gibson. 2012. Verbal interference suppresses exact numerical representation. *Cognitive Psychology* 64. 74–92.
Frank, M. C. & T. Honeyman. 2011. Number knowledge in a finite counting system. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.
Gallistel, C. 1993. *The organization of learning*. Cambridge, MA: MIT Press.
Gelman, R. & B. Butterworth. 2005. Number and language: How are they related? *Trends in Cognitive Sciences* 9(1)., 6–10.
Gelman, R. & C.R. Gallistel. 1978. *The child's understanding of number*. Cambridge, MA: Harvard University Press.
Gentner, D. 2003. Why we're so smart. In Gentner & Goldin-Meadow, *Language in mind*, 195–235.
Gentner, D. & S. Goldin-Meadow. 2003. Whither whorf. In Gentner & Goldin- Meadow, *Language in mind,* 3–14.
Gentner, D. & S. Goldin-Meadow (eds.). 2003. *Language in mind: Advances in the study of language and cognition.* Cambridge, MA: MIT Press.
Goldin-Meadow, S. & C. Mylander. 1984. Gestural communication in deaf children: The effects and noneffects of parental input on early language development. *Monographs of the Society for Research in Child Development* 49(3/4). 1–151.
Gordon, P. 2004. Numerical cognition without words: Evidence from Amazonia. *Science* 306. 496–499.
Gumperz, J. J. & S. C. Levinson. 1996. *Rethinking linguistic relativity*. Cambridge, UK: Cambridge University Press.
Hammarström, H. 2009. Whence the Kanum base–6 numeral system? *Linguistic Typology* 13. 305–319.
Hammarström, H. 2010. Rarities in numeral systems. In Jan Wohlgemuth & Michael Cysouw (eds.), *Rethinking universals: How rarities affect linguistic theory* (Empirical Approaches to Language Typology), 11–60. Berlin: De Gruyter.
Hatano, G. 1977. Performance of expert abacus operators. *Cognition* 5(1). 47–55.
Hatano, G. & K. Osawa. 1983. Digit memory of grand experts in abacus-derived mental calculation. *Cognition* 15(1–3). 95–110.
Hauser, M., F. Tsao, P. Garcia & E. Spelke. 2003. Evolutionary foundations of number: spontaneous representation of numerical magnitudes by cotton-top tamarins. *Proceedings of the Royal Society of London* (Series B: Biological Sciences) 270(1523). 1441.
Hermer, L. & E. Spelke. 1994. A geometric process for spatial reorientation in young children. *Nature* 370. 57–59.
Hermer-Vazquez, L., E.S. Spelke & A.S. Katsnelson. 1999. Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology* 39. 3–36.
Kay, P., B. Berlin, L. Maffi & W.R. Merrifield. 2003. *The world color survey*. Palo Alto, CA: CSLI Press.
Kay, P. & W. Kempton. 1984. What is the Sapir-Whorf hypothesis? *American Anthropologist* 86(1). 65–79.

Kojima, T. 1954. *The Japanese abacus: Its use and theory*. Tokyo: Charles E. Tuttle Company.
Lean, G. 1992. Counting systems of Papua New Guinea and Oceania. Lae: Papua New Guinea University of Technology doctoral dissertation.
Le Corre, M., G. Van de Walle, E.M. Brannon & S. Carey. 2006. Re-visiting the competence/performance debate in the acquisition of the counting principles. *Cognitive Psychology* 52. 130–169.
Levinson, S. C., S. Kita, D. B. M. Haun & B. H. Rasch. 2002. Returning the tables: Language affects spatial reasoning. *Cognition* 84. 155–188.
Li, P. & L. Gleitman. 2002. Turning the tables: Language and spatial reasoning. *Cognition* 83. 265–294.
Logie, R. & A. Baddeley. 1987. Cognitive processes in counting. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13. 310–326.
Lucy, J. A. 1992. *Language diversity and thought: A reformulation of the linguistic relativity hypothesis.* Cambridge: Cambridge University Press.
Lupyan, G., D. H. Rakison & J. L. McClelland. 2007. Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science* 18. 1077–1083.
Menninger, K. 1969. *Number words and number symbols: A cultural history of numbers*. Cambridge, MA: MIT Press.
Newton, A. M. & J.G. de Villiers. 2007. Thinking while talking: Adults fail nonverbal false-belief reasoning. *Psychological Science* 18. 574–579.
Papafragou, A., J. Hulbert & J. Trueswell. 2008. Does language guide event perception? evidence from eye movements. *Cognition* 108(1). 155–184 .
Pashler, H. 1994. Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin* 116. 220–220.
Piantadosi, S., J. Tenenbaum & N. Goodman. 2012. Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition* 123(2). 199–217.
Pica, P., C. Lemer, V. Izard & S. Dehaene. 2004. Exact and approximate arithmetic in an Amazonian indigene group. *Science* 306. 499–503.
Pinker, S. 1994. *The language instinct*. New York, NY: Penguin Books.
Pyers, J. & A. Senghas. 2009. Language promotes false-belief understanding. *Psychological Science* 20(7). 805.
Roberson, D. & J. R. Henley. 2007. Color vision: Color categories vary with language after all. *Current Biology* 17. R605–R607.
Saxe, G. 1982. Developing forms of arithmetical thought among the Oksapmin of Papua New Guinea. *Developmental Psychology* 18(4). 583–594.
Senghas, A. & M. Coppola. 2001. Children creating language: How Nicaraguan sign language acquired a spatial grammar. *Psychological Science* 12(4). 323–328.
Senghas, A., S. Kita & A. Ozyurek. 2004. Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science* 305(5691). 1779.
Spaepen, L., M. Coppola, E. S. Spelke, S. Carey & S. Goldin-Meadow. 2011. Number without a language model. *Proceedings of the National Academy of Science*s 108(8). 3163–3168.

Stigler, J. W. 1984. Mental abacus: the effect of abacus training on Chinese children's mental calculation. *Cognitive Psychology* 16(2). 145–176.
Tanaka, S., C. Michimata, T. Kaminaga, M. Honda & N. Sadato. 2002. Superior digit memory of abacus experts: an event-related functional MRI study. *Neuroreport* 13(17). 2187.
Whalen, J., C.R. Gallistel & R. Gelman. 1999. Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science* 10. 130–137.
Whorf, B. L. 1956. *Language, thought, and reality*. Cambridge, MA: MIT Press.
Wiese, H. 2007. The co-evolution of number concepts and counting words. *Lingua* 117(5). 758–772.
Winawer, J., N. Witthoft, M. C. Frank, L. Wu, A. R. Wade & L. Boroditsky. 2007. Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences* 104(19). 7780–7785.
Wynn, K. 1990. Children's understanding of counting. *Cognition* 36. 155–93.
Xu, F. 2002. The role of language in acquiring object concepts in infancy. *Cognition* 85. 223–250.
Xu, F. & E. Spelke. 2000. Large number discrimination in 6-month-old infants. *Cognition* 74(1). B1–B11.

Michael C. Frank
mcfrank@stanford.edu

10

Language Documentation & Conservation Special Publication No. 5 (December 2012)
Melanesian Languages on the Edge of Asia: Challenges for the 21st Century,
ed. by Nicholas Evans and Marian Klamer, pp. 239–253
http://nflrc.hawaii.edu/ldc/sp05/
http://hdl.handle.net/10125/4567

Keeping records of language diversity in Melanesia: The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)

Nicholas Thieberger

University of Melbourne

Linda Barwick

University of Sydney

At the turn of this century, a group of Australian linguistic and musicological researchers recognised that a number of small collections of unique and often irreplaceable field recordings mainly from the Melanesian and broader Pacific regions were not being properly housed and that there was no institution in the region with the capacity to take responsibility for them. The recordings were not held in appropriate conditions and so were deteriorating and in need of digitisation. Further, there was no catalog of their contents or their location so their existence was only known to a few people, typically colleagues of the collector. These practitioners designed the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), a digital archive based on internationally accepted standards (Dublin Core/Open Archives Initiative metadata, International Asociation of Sound Archives audio standards and so on) and obtained funding to build an audio digitisation suite in 2003. This is a new conception of a data repository, built into workflows and research methods of particular disciplines, respecting domain-specific ethical concerns and research priorities, but recognising the need to adhere to broader international standards. This paper outlines the way in which researchers involved in documenting languages of Melanesia can use PARADISEC to make valuable recordings available both to the research community and to the source communities.

1. Introduction. At the turn of this century, a group of Australian linguistic and musicological researchers recognised that a number of small collections of unique and

often irreplaceable field recordings mainly from the Melanesian and Pacific regions were not being properly housed and that there was no institution in Australia which would take responsibility for them. The recordings were not held in appropriate conditions and so were deteriorating and in need of digitisation. Further, there was no catalog of their contents or their location so their existence was only known to a few people, typically colleagues of the collector. These researchers designed the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), a digital archive based on internationally accepted standards and obtained Australian Research Council Infrastructure funding to develop an audio digitisation suite in 2003. This is a new conception of a data repository, built into workflows and research methods of particular disciplines, respecting domain-specific ethical concerns and research priorities, but recognising the need to adhere to broader international standards.

2. Background. Researchers (in particular linguists, musicologists and anthropologists) working with speakers of small languages (those with few speakers) typically conduct fieldwork to learn how aspects of these societies function, how the languages are structured, or how musicological knowledge is constituted, in addition to recording life stories, ethnobiological and other information. Typically these are minority endangered languages for which no prior documentation exists. This is vitally important work which often records language structures and knowledge of the culture and physical environment that would otherwise be lost (see e.g., Evans 2009, Maffi 2001, Harrison 2007). While it is typical for the interpretation and analysis of this data to be published eventually, the raw data is rarely made available. The data—tapes, field notes, photographs, and video—are often not properly described, catalogued, or made accessible, especially in the absence of a dedicated repository. This means that enormous amounts of data, often the only information we have on disappearing languages, remain inaccessible both to the language community itself, and to ongoing linguistic research.

The data that we create as part of our research endeavour should be reusable, both by ourselves and by others, and, in particular by the speakers and the general community with an interest in the nature of linguistic diversity in Melanesia. Beside the imperative to ensure there are good records of these languages this is also because any claims that we make based on that data must themselves be replicable and testable by others, and because the effort of creating the data should not be duplicated later by others, and provide a foundation that can be built on. In order to be made accessible, the data recorded by researchers must be properly collated and indexed for public presentation and archiving (see Himmelmann 1998, Woodbury 1998, 2003). However, until recently there has been no simple means for doing this and access to physical analog records can be difficult, if not virtually impossible, when they are stored in a single location.

This issue is being faced by scholars in many disciplines, and is being addressed under the rubric of cyberinfrastructure (National Science Foundation (U.S.), 2003) or ehumanities—how to build on existing knowledge and how to add new data that is being created in the course of various research projects so that the broader research community can benefit from it. This is all the more important when a linguist makes the only recordings for an endangered language–one that may no longer be spoken in the near future. Australia and its immediate neighbours are home to a third of the world's languages, most of which may never be recorded. Many of these languages could include completely novel structures

or ways of viewing the world, but each of them reflects the history of their speakers and is worthy of detailed recording. Melanesia in particular is among the most linguistically diverse regions (see Hammarström & Nordhoff this isssue), with Vanuatu having the highest density of languages per person of any country.

Significant resources are now being devoted to recording endangered languages in Europe (the Documentation of Endangered Languages project administered by the Max Planck Institute, Nijmegen) and the UK (Endangered Languages Documentation Project) and in the USA (the joint NSF/NEH program titled Documenting Endangered Languages). Furthermore, there are many local initiatives for recording oral tradition, like the fieldworker programme at the Vanuatu Kaljoral Senta or the collections being made by the Agence de Développement de la Culture Kanak (ADCK) or the Academy for Kanak Languages in New Caledonia. If the data arising from all of this effort is not properly safeguarded in our region it will represent a loss of cultural information, not to mention an enormous waste of effort and money. Many recordings are not described sufficiently to allow their contents to be discovered, and often there is little thought given to the methods involved in managing large multimedia datasets, which are especially vulnerable because they are in digital formats that are at risk (either due to lack of suitable digital data preservation and management infrastructure, or because of format obsolescence in a fast-changing digital media environment). Too much data is stored in ways that make it hard to access for the research community, let alone the broader community. Some research groups develop their own computational solutions which, admittedly, serve their needs well but which renders the group and their data isolated from the rest of the scientific community. The development of a new methodology, which includes the adaptation or development of new tools, must be grounded in application of that methodology to real data (Bird and Simons 2003). There are too many examples of 'proofs of concept' which set out directions for further work but which are not immediately applicable to any real-world problem.

3. Technology gap (the digital divide) and multimedia. It is a concern to some that we use increasingly technological methods for recording traditional practices, while the cultures in which they are embedded and the people who practise them have little access to the benefits offered by these technologies. How appropriate is it to use high technology, such as digital multimedia, with languages from villages that have no electricity? Of course, there is nothing new about the gap between the resources available to the researcher and those available to the researched, this is the colonial essence of any research project run by a first-world linguist. Suggesting that a video recorder is more colonial than handwritten notes (see for example Aikhenvald 2007) ignores the extractive nature of both forms of recording, and, more importantly, ignores the need for researchers to make the richest possible record for reuse by the speaker community. We should think in terms of what technology is appropriate for the task, and, in the case of recording oral tradition as the basis for both linguistic research and for heritage purposes, it is clear that we must use methods based in digital technologies (Bowden and Hajek 2006), because analog recording formats and equipment are all but obsolete (Schüller 2004).

The realisation that we can use multimedia data to enrich our understanding of performance is not something recent, and indeed goes back to the days of phonograph recordings, as this quote from Malinowski about his fieldwork in the Trobriand Islands illustrates:

> If I could, by a good phonographic record, counterfeit the living voice of Tokulubakiki: […] I should certainly be better able to translate the text in the sense of imparting to it its full cultural flavour and significance. Again, if by cinematographic picture I could reproduce the facial expression, the bodily attitude, the significant gestures, this would add another contextual dimension. (Malinowski 1935: 26)

While the technology to record and play back performances has been available since the late 1800s, it was rarely used by linguists until the second half of the twentieth century, and even then, analog recordings were difficult to create in the field, and later, and to access. It is only with the advent of digital media that we see the development of instant access to time points within large media corpora and the associated (but still painfully slow) realization among linguists that they can create reusable corpora in which their analysis can be embedded (Thieberger 2009). It is critical that a distinction is clearly made between archival forms of the media (held in high resolution files, such as 24-bit 96 kilohertz uncompressed audio, which are described in a catalog, and given persistent location and naming) and delivery or access forms of multimedia (which will be of lower resolution and often compressed for delivery via appropriate formats, such as the web or mobile phones). Multimedia presentations are seductive in their ability to relate parts of collections, linking texts to media or images and media to dictionaries. We have, however, seen enough examples of multimedia packages that are costly, contain relatively small amounts of information and become unplayable after a few years.

4. Access to digital data in the region. Williams (2002:15), in a report on the status of digital community services in the Pacific, noted that:

> [i]nformation on hardware resources […] shows that while all libraries, archives and museums that responded have access to at least a computer, the situation is bleak. Except for libraries in the Republic of Palau (and presumably in the Micronesian region) and university libraries and centres in the University of the South Pacific network, Fiji Institute of Technology, Fiji School of Medicine, National University of Samoa and University of Papua New Guinea, the computers are used by staff for work operations. In the Library Service of Fiji, there is no computer for public use, with only one computer in the library. The Suva Public Library is in a better situation. The Niue Public School Library, Tuvalu Culture Office and the Samoa National Archives also do not have computer access for students or members of the community.

It is clear from reports such as this (and from our own observation) that there is still a long way to go in the provision of digital information in small Pacific Island communities. Nevertheless, in the decade since Williams's report there have been unexpected advances in access to digital resources in even quite remote areas of the Pacific. Mobile phone technology has been taken up with enthusiasm, and has coverage in many previously unconnected locations, allowing remote use of both telephony and the 'mobile web' (See

Picture 1). The World Bank 'Rural Communication Project' (World Bank 2010) in PNG aims to significantly increase the number of internet users there, from the estimated current 50,000 mostly based in Port Moresby, and to increase coverage in rural district centers.

We can expect to see mobile phones taking over functions of portable computers in remote locations and so should also plan on building access to cultural collections using these technologies. The development of mobile phone dictionaries of small languages based on common formats of lexical databases (see, for example, the PARADISEC project Wunderkammer) can already provide online or local access to electronic dictionaries with sound and images. Similarly, new methods of streaming digital media allow for efficient delivery of ethnographic recordings over low bandwidth, including mobile phones. The PARADISEC project EOPAS streams audio or video recordings of stories over the internet together with text (see the discussion below) using HTML5 and open-source media. HTML5 is an emerging web standard that allows streaming of multimedia within the stamdard web page, thus obviating the need for users to install additional software or plugins (Pfeiffer 2010). All of this indicates that creating proper forms of recordings, images and so on that conform to accepted archival standards will allow them to be transformed into delivery formats appropriate to the context in which they are to be used.

Figure 1. Publicity billboard for internet access via mobile phones (Port Vila, June 2011). Photo by Nick Thieberger

5. Ethics of information provision. In addition to the question of equitable access to the kind of cultural information that is now becoming commonplace on the internet, there is the more complex issue of the sensitivity of archival records being reintroduced in new contexts. Recordings made in the 1950s may take on a considerably different meaning when used today, especially if there are land disputes that otherwise rely on oral accounts remembered by the current generation. The archival record can assume an authority (whether justified or not) that may be advantageous to some in the present dispute, but detrimental to others. While those running an archive can be aware that such problems may arise, it is impossible for them to know such details for all of the locations from which the archive stores material.

In most societies there is some kind of protocol in place for access to certain kinds of information. Not everyone can read the records of company meetings, for example, or of secret government business. In smaller societies, such protocols may include access to songs or stories that relate to the first creation of the land or to the travels of ancestral beings: see for example Lindstrom (1990) on what he terms 'the economy of knowledge' in Tanna, southern Vanuatu. The provision of such information from an archive may subvert the very power structures that promote the ongoing use of traditional languages and clearly this is a potentially difficult situation for an archivist to find themselves in. The Endangered Languages Archive at SOAS has been working on a system for allowing more fine grained access conditions to be specified, including, for example, the ability for people other than depositors to determine who can access the recordings of themselves speaking. However, our present focus has been on preservation of the records we have located and we consider it more important that the material be stored for later reuse than that the safer option (that there be no archival record) be adopted.

6. Implementation of PARADISEC. In the initial phase of the PARADISEC project (2003) we established a steering committee with representatives of each of the partner universities (initially Sydney University, the University of Melbourne, ANU, and later UNE). The director of the project is Linda Barwick at the University of Sydney.

With invaluable technical support from both the National Library of Australia and the National Film and Sound Archive and with funds from the Australian Research Council we bought a Quadriga digitisation suite and employed an audio engineer and administrative assistant, based at the University of Sydney. We also built a vacuum chamber and low-temperature oven to allow us to treat mouldy tapes that required special care before being playable. Tapes stored at the ANU were identified and located and then permission was sought from the collectors or their agents to digitise and accession them into the collection.

In the first year of funding we had to come up with outcomes that would justify further funding grants and we aimed for 500 hours of digitized tapes in that first year (we achieved this goal in ten months). We wrote a catalog database in Filemaker Pro, aware that it would provide us with an immediately usable tool that would ultimately have to be converted to an online database. This database allowed us to refine data entry forms and controlled vocabularies without relying on a programmer. This first catalog worked well and exported to the XML files required for inclusion as headers in Broadcast Wave Format (BWF) files, and also exported to a static repository for Open Archives Initiative harvesting via the Open Language Archives Community harvester.

Files generated by this system (at 96khz/24 bit) are large, around 1.5 Gb per 45-minute

side of a cassette, and so require dedicated storage facilities. We established a tape backup system which ran periodically to copy files from the hard disk to storage tapes, but were fortunate when the Australian Partnership for Advanced Computing (APAC) designated PARADISEC a 'Project of National Significance', allowing us to use their mass data storage system, with considerable storage space provided to support our work. They further provided programming support by writing specialized software (called 'Babble') which provides weekly, monthly and quarterly reports on the state of the collection, as well as nightly querying the server in Sydney and copying files that are ready for archiving.

Data is organized by collector, but also by the internal logic of the collections (the same collector working on two different languages will have two collections, or a collection of video may be distinct from a collection of still images). The collection-level also speeds up a user's typing into the catalog as common fields from the collection level can be inherited down to the item level. Our naming convention is rather simple ('CollectionID'-'ItemID'-'FileID'.'extension') and it also provides the hierarchical file structure into which files are placed and stored on the server (with directories corresponding to the collection level and subdirectories corresponding to the item level).

Subsequently and with funds from the Australian Research Council Linkage Infrastructure Equipment and Facilities (LIEF) programme, we built digitisation suites in Melbourne and Canberra, allowing us to preserve important heritage tape collections such as those shown in table 2, by no means an exhaustive list. Without a dedicated infrastructure to describe, manage and store this material it would simply be lost.

Mark Durie (Acehnese, Indonesia)	Cindy Schneider (Apma, Vanuatu)
Barry Alpher (Cape York, Australia)	Sébastien Lacrampe (Lelepa, Vanuatu)
Sander Adelaar (Selako, Indonesia)	Stephen Morey (Assam, India)
Sebastian Fedden (Mianmin, PNG)	Robyn Loughnane (Oksapmin, PNG),
Amanda Brotchie (Tirax, Vanuatu)	Nick Thieberger (South Efate, Vanuatu)

TABLE 1. Examples of collections from Australia and its the region that have either been digitised by PARADISEC or accessioned as digital data by PARADISEC

Now that many researchers are recording directly to digital formats, we provide advice and guidance on suitable formats and workflows to facilitate ingestion into the repository. On return from fieldwork, depositing in PARADISEC provides a means of secure backup of researchers' otherwise vulnerable digital media files. We still have a need for digitization of older analog collections, a much slower process to produce a high quality digital preservation master file for archiving (International Association of Sound and Audiovisual Archives (IASA), 2004).

7. LICENSING USE OF ITEMS IN THE COLLECTION. The primary aim of the project to date has been on preservation of unique cultural records. Including a licence, or information about how each item can be used, is critical to the establishment of a properly curated collection because without it there is no way of providing access. Each depositor must fill

out a deposit form specifying any conditions that may apply to the material. We provide a default set of access conditions which any user must agree to prior to being given access to data, and depositors can choose to allow this set of conditions to govern their collection, or to determine their own conditions. We are presently investigating the use of Creative Commons licences as a less restrictive and more standardised form of agreement (Newman 2007, Seeger 2005).

8. Delivery of archival material, page images and dynamic media. We provide material from the collection to those authorized to receive it, typically in the form of downloadable files, however we have also worked on specific methods for the online delivery of two kinds of material – page images and time-coded media. We made available images of 14,000 pages of fieldnotes (see figure 2) from three deceased researchers using the Heritage Document Management System with a digital camera rig that we took to the home of the estate's executor, or to the office in which the papers were stored. These notes from deceased researchers would otherwise have only been available in a single physical location. As we do not have the resources to keyboard all of these manuscripts the images are stored in the collection with sufficient contextual metadata to make them discoverable on the web. As noted earlier, the archival version of each image is stored separately from the representational version.

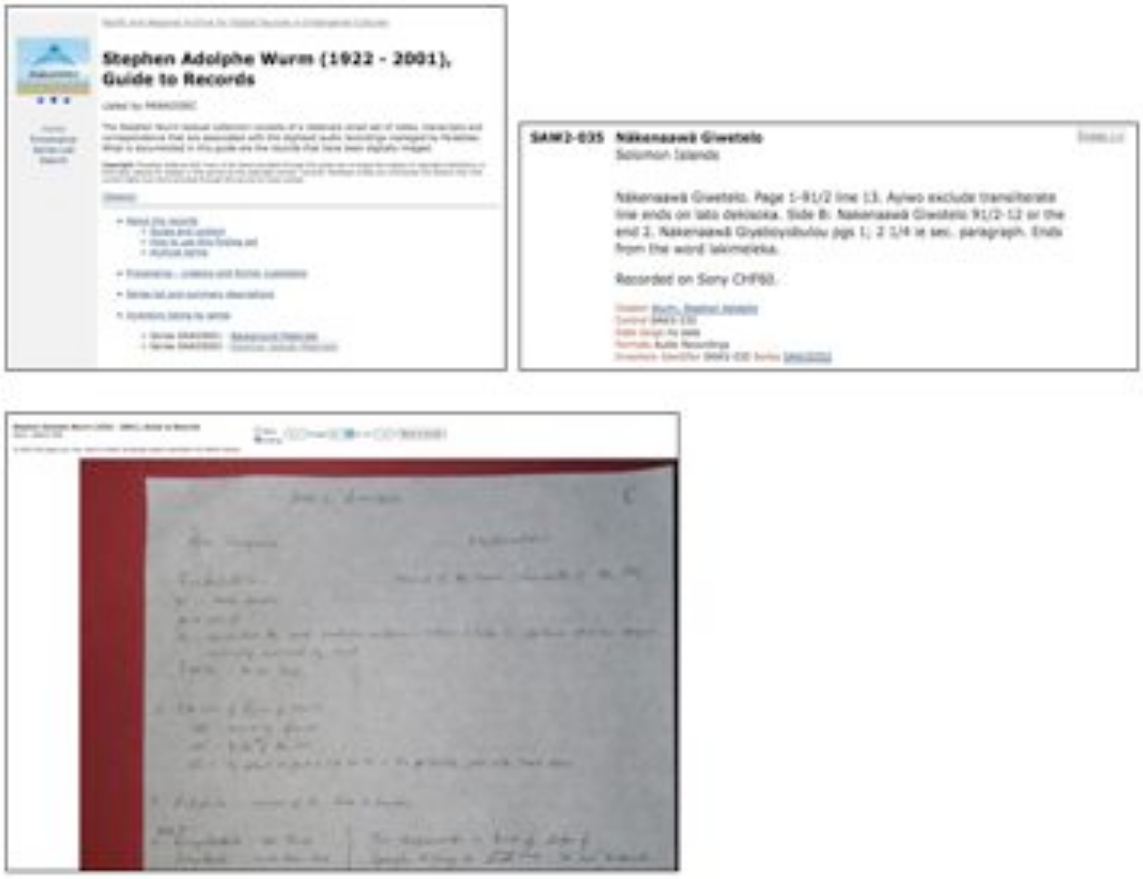


FIGURE 2. Page images from the Wurm collection of online manuscripts showing finding aids from the highest level (top left), to the item level (top right) and finally the image itself (bottom) (http://paradisec.org.au/fieldnotes/SAW2/SAW2.htm)

9. THE ETHNOER ONLINE PRESENTATION AND ANNOTATION SYSTEM (EOPAS). While building a method for working with our own data we consider it important to create generalisable models and structures for others to use, and to engage in discussions and training sessions both in order to refine our methodologies and to impart new ideas. An example of such development is our work on the online presentation of interlinear glossed text together with recorded media (EOPAS), allowing material from any language to be heard in concert with its transcript and translation (Schroeter and Thieberger 2006). A number of tools for annotating language data have been produced recently and it is clear that more are envisaged now that several large projects are engaging with these issues in the USA, UK, Germany and the Netherlands. Annotation is a basic task that is undertaken following recording, and now it is typically carried out with time-alignment, meaning

that the text has references to timepoints within the media file (using software such as Elan or Transcriber) and can take several forms, the most common of which, for linguists, is interlinear text. These texts are analysed and parsed by a glossing tool that produces parallel lines of text, word translation and grammatical information, together with a free translation. These texts are then input into EOPAS, a schema-based XML system for making explicit the relationship between parts of interlinear texts together with links to the source media (see figure 3) which allows searching and concordancing linked directly to the media. EOPAS is portable (the source code is freely available), allowing other initiatives to capitalise on the work and potentially develop it in different directions. The ultimate aim of this approach is to allow new perspectives on the data itself, provided by contextualised access to primary data, and then to allow new research questions to be asked, and richer answers to be provided, all in a fraction of the time that it would have taken with analog data.

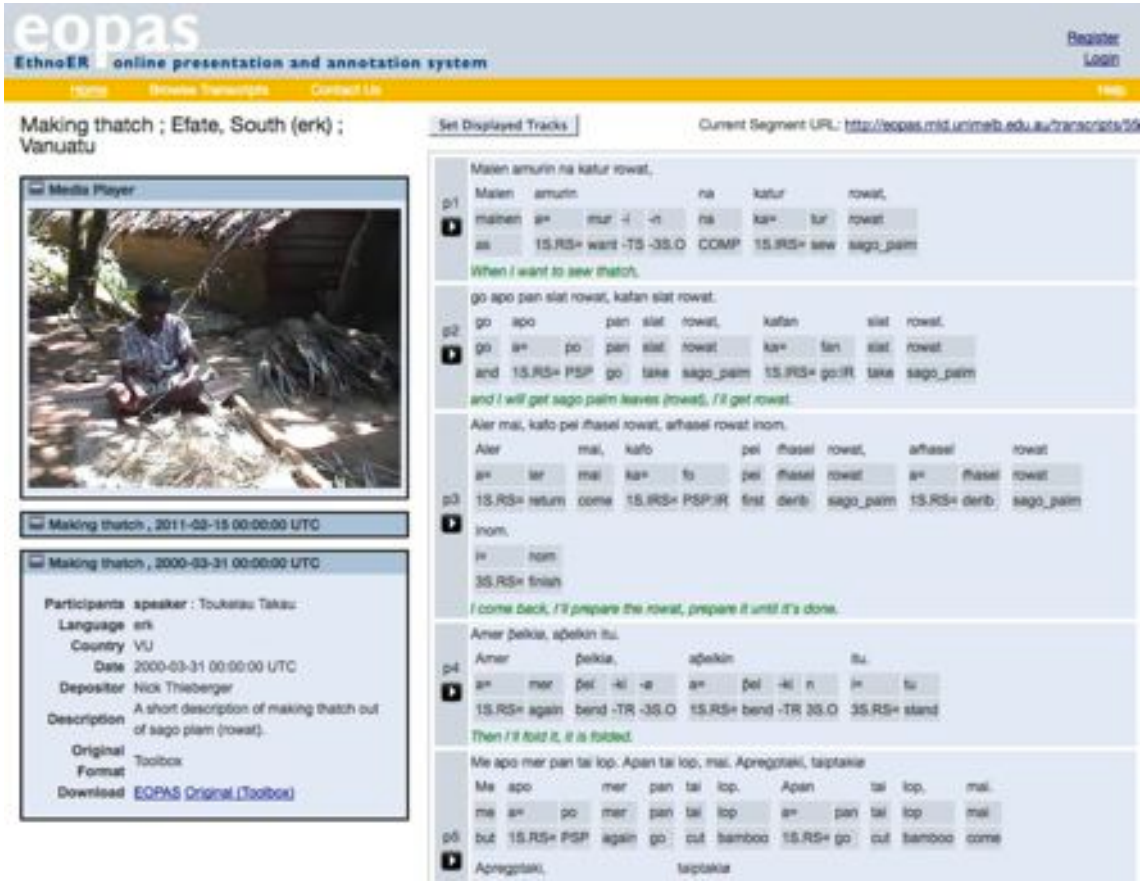


FIGURE 3. Example of a video clip with time-aligned text as presented in EOPAS.

10. Current status of the PARADISEC collection. Currently (late 2011) PARADISEC contains 7,226 items made up of 48,606 files totaling 5.2 TB, with just over 3,046 hours of audio data. Digital video already makes up an increasingly significant part of the collection. We hold data representing 650 languages from 60 countries (see examples of the kinds of collections in table 1) which is organized into 163 collections, some 85 of which represent new fieldworkers who have deposited material on their return from fieldwork (and one during the course of her fieldwork), thus providing a citable form of their data for their own research. This means that in their dissertations and publications they can refer precisely to the relevant linguistic data through citing the timecodes associated with the persistent identifier (web location) of their recordings in the PARADISEC collection. Citation of primary data is a critical step in conducting new research based on that data. The remaining collections are digitised from recordings made since the 1950s. The provision of this service requires ongoing support and negotiation with depositors and we have found that a key to establishing the collection has been the depositors' perception of the benefit accruing to them and to their data in having it well described. In addition, there are collections we know about and would dearly love to digitise but we do not have the resources to do this work. These include large audiotape collections at radio stations around the Pacific, many in local languages, and collections in regional cultural centres that do not have any local equipment to digitize their collections. Further, we are regularly approached by former colonial patrol officers or missionaries who have recordings, notes or photographs that they want to preserve.

Arthur Capell	1950s Pacific and PNG (114 tapes and 30 archive boxes of fieldnotes)
Tom Dutton	1960s onwards, PNG, 295 tapes
William Foley	1970s, PNG, 34 tapes
John Harris	1960s, Kiwai, PNG, 75 tapes
Don Laycock	1960s, PNG, 98 tapes
Al Schütz	1960s onwards, Vanuatu, six tapes
Stephen Wurm	1970s Solomon Islands tapes (~120 tapes and transcripts/ fieldnotes)
Bert Voorhoeve	West Papua, 180 tapes

Table 2. Example collections that have been digitized, described or curated by the PARADISEC project.

We have published on our website a detailed description of our workflow, developed over seven years of operation, that describes the various processes involved in locating tapes and then assessing, accessioning, digitising and describing them, managing the resulting data and metadata, and the return of original tapes. PARADISEC has been cited as an exemplary system for audiovisual archiving using digital mass storage systems by the International Association of Sound and Audiovisual Archives and, in 2008, won the Victorian Eresearch Strategic Initiative prize for humanities eresearch.

Once we built the infrastructure for a research repository, including the catalog, file system and naming conventions, it has been taken up by those researchers who are aware of the need to describe and preserve their research material. Often it is only in the process of depositing with PARADISEC that a collection is first described in a systematic way – one that then allows the description to be searched by Open Archives Initiative search engines (and also google). Every eight hours the PARADISEC catalog is queried by a service run by the Open Language Archives Community (OLAC) and any new or edited catalog entries are copied and made available to their aggregated search mechanism. Similarly, because the catalog complies with relevant standards, the Australian National Data Service (ANDS) has been able to incorporate our 163 collections into its national search mechanism. The quality of the metadata we provide ensures that targeted searches by language name can be resolved without locating similar but irrelevant forms.

11. Regional links and training. While the initial focus for our collection was the region around Australia (as suggested by the name we chose at the outset of the project), it has become clear that we need to accept material that has no other place to be archived. Typically, this means supporting Australian researchers whose research is outside of Australia, with the geographic spread of material we house now extending from India, into China, and across to Rapanui (Easter Island). With limited resources PARADISEC has nevertheless established working relationships with cultural centres in the Pacific region (e.g., the Vanuatu Kaljoral Senta, or the Institute of PNG Studies) which have involved providing CD copies of relevant material and, in the case of the University of New Caledonia, cleaning and digitising old reel-to-reel tapes in Drehu. A serious concern for many such agencies in the region (as observed in Williams' report, above) is the lack of continuity in funding and in staffing, with the potential result that collections established and curated over time may be at risk. We would like to be able to digitize the many hours of tapes held, often in less than ideal conditions, in countries of the region. We have begun an occasional mass backup of significant collections of digital material from the Vanuatu Kaljoral Senta and would like to extend this as a service to other agencies.

We regularly offer training workshops in linguistic research methods, including the use of appropriate tools and recording methods and in data management for ethnographic field material. This is extremely important, as the more informed the research community can become about the need for reuse of primary data, the more likely they are to be creating well-formed data that needs no extra handling by PARADISEC to be accessioned into the collection. Such training has been offered at community Indigenous language centres as well as in academic settings.

We cooperate in two further initiatives for disseminating information. The first is a blog (Endangered Languages and Cultures) and the second a resource website with FAQs and a mailing list (the Resource Network for Linguistic Diversity). Because of the rapid changes in methods for recording, transcribing, and analysing human performance no one can keep completely up to date, so these web-based resources are widely quoted and appreciated by the community of researchers.

12. The future of the collection. As the value of data curation becomes clearer and the use of the collection increases, we will see more theoretical work based on properly curated archival material. We have already seen linguists retrieving what are now historical language records for use in comparison with current usage and for analysis of language change. Serendipitous discoveries in the collections have included the drama specialist Diana Looser finding a performance of Albert Toro's 1977 radio serial, Sugar Cane Days, a historical drama about the 'blackbirding' days of indentured Kanak labour in the Queensland canefields. While discrete sections of Toro's play had been published in local literary anthologies and magazines in the early 1980s, no complete script of the play was available. Tom Dutton had recorded the complete five-part performance taken in Port Moresby in the 1970s, as well as an interview with Toro about the inspiration for, and genesis of, the play. These unique sound files allowed Looser not only to listen to the original radio play in performance, but to create a verbatim transcript from the recording.

PARADISEC is a project ahead of its time and so suffers from a lack of vision among funding agencies. It is truly collaborative, multi-institutional and cross-disciplinary which, despite frequent funding-agency rhetoric to the contrary, weighs against it being supported through normal research funding sources.

We would like to extend the streaming server we have established to allow delivery of any accessible material in the collection. We are also in the process of developing an access system with authentication and authorization of users.

PARADISEC is part of several international networks of similar projects (DELAMAN or OLAC, cited above), but is a leading exponent of linguistic data curation even among that field. Australian government moves to establish a national digital data service (a system of repositories hosting digital data in the way that PARADISEC has done) are still in their early stages, but we are confident that PARADISEC will become part of such a service within the next decade. Our unique collection needs to be safely shepherded through the intervening period, identifying more collections in need of digitisation, accessioning them, and providing the infrastructure for current researchers and postgraduate students to describe and preserve their field recordings. We need to continually provide training and advice for researchers in order that their outputs can be accessioned with minimal extra handling. Research that is conducted without an awareness of appropriate data structures and formats will result in poor outputs that need to be converted, often with considerable effort, to make them archivable. It is unlikely that this arduous conversion effort will be resourced and so we risk losing primary research data.

13. Conclusion. PARADISEC is a practice-based archive, arising from a community of practice who recognised that it was part of our professional responsibility to ensure that the records we create are properly curated into the future. This is a new conception of a data repository, accessioning primary research in the course of fieldwork or shortly after, and building methods and tools to facilitate its deposit and curaton. It is unique in its links on the one hand to fieldworkers and to speakers of Indigenous languages and on the other hand to the cutting-edge technologies of Web 2.0 and HTML5.

PARADISEC has been active in locating records of small languages and making them available for longterm access. We have been particularly aware of the needs of small language communities, especially those in PNG and island Melanesia. In 2012 we

have collaborated with the Solomon Islands Museum and Archives to apply for funding to digitise their audio collections. Similar collections of audio, film and video exist in agencies across the Pacific and are in need of urgent attention. Our new catalog will make streaming media available for viewing on a variety of platforms, including mobile phones, and this should allow delivery of these unique resources to their source communities. PARADISC is keen to attract more funding, so as to locate and digitise more material, and provide training to speakers to create their own records now. We could also increase the representation of languages in our EOPAS system to provide online samples of as many languages of the region as possible. There is much more to be done, but the work done by PARADISEC will allow future work to grow on good foundations.

References

Aikhenvald, Alexandra. 2007. Linguistic fieldwork: setting the scene. *Sprachtypologie und Universalienforschung* 60(1). 3-11.

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3). 557-582.

Bowden, John & John Hajek. 2006. When best practice isn't necessarily the best thing to do: dealing with capacity limits in a developing country. In Barwick, Linda & Nicholas Thieberger (eds.), *Sustainable data from digital fieldwork*, 45-55. Sydney: Sydney University Press.

Evans, Nicholas. 2009. *Dying words: Endangered languages and what they have to tell us*. Maldon, MA: Wiley-Blackwell.

Harrison, K. David. 2007. *When languages die: The extinction of the world's languages and the erosion of human knowledge*. New York: Oxford University Press.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161-195.

International Association of Sound and Audiovisual Archives (IASA). 2004. *Guidelines on the production and preservation of digital audio objects (IASA-TC04)*. Aarhus, Denmark: International Association of Sound and Audiovisual Archives.

Lindstrom, Lamont. 1990. *Knowledge and power in a South Pacific society.* Washington: Smithsonian Institution Press.

Maffi, Luisa (ed.). 2001. *On biocultural diversity: Linking language, knowledge and the environment*. Smithsonian Institution, Washington, D.C.

Malinowski, Bronislaw. 1935. *Coral gardens and their magic*. Vol 2. London: Allen and Unwin.

National Science Foundation (U.S.). 2003. *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Arlington, VA: Office of Cyberinfrastructure, National Science Foundation.

Newman, Paul. 2007. Copyright essentials for linguists. *Language Documentation & Conservation,* 1(1). 28–43.

Pfeiffer, Silvia. 2010. *The definitive guide to HTML5 video*. New York: Apress.

Schroeter, Ronald and Nicholas Thieberger. 2006. EOPAS, the EthnoER online representation of interlinear text. In Barwick, Linda and Nicholas Thieberger (eds.), *Sustainable data from digital fieldwork*, 99-124. Sydney: Sydney University Press.

Schüller, Dietrich. 2004. Safeguarding the documentary heritage of cultural and linguistic diversity. *Language Archive Newsletter* 1(3). 9-10.

Seeger, Anthony. 2005. New technology requires new collaborations: Changing ourselves to better shape the future. *Musicology Australia 27*(2005-6). 94-111.

Thieberger, Nicholas. 2009. Steps toward a grammar embedded in data. In Epps, Patricia & Alexandre Arkhipov (eds.), *New challenges in typology: Transcending the borders and refining the distinctions*, 389-408. Berlin & New York: Mouton de Gruyter. http://repository.unimelb.edu.au/10187/4864.

Watson, Amanda H. A. 2010. Communication and culture: mobile telephony in PNG villages. Paper presented at the Asian Media Information and Communication Centre: Technology and Culture: Communication Connectors and Dividers, 21-23 June 2010, Suntec City, Singapore (AMIC). http://eprints.qut.edu.au/32787/. (28 July, 2010.)

Whimp, Kathy & Mark Busse (eds.). 2000. *Protection of intellectual, biological and cultural property in Papua New Guinea*. Canberra: Asia Pacific Press.

Williams, Esther. 2002. *Digital community services: Pacific libraries and archives.* UNESCO.

Woodbury, Anthony. 1998. Documenting rhetorical, aesthetic, and expressive loss in language shift. In Grenoble, L.A. and L. J. Whaley (eds.), *Endangered languages: language loss and community response*, 234-258. Cambridge: Cambridge University Press.

Woodbury, Anthony. 2003. Defining Documentary Linguistics, address given at the Annual Meeting of the Linguistic Society of America, Atlanta, Georgia, on January 3, 2003.

World Bank. 2010. Remote Rural Communities in Papua New Guinea to Benefit from Improved Access to Telecommunications. http://go.worldbank.org/DHJ6XJO2O0 (28 July, 2010.)

Nicholas Thieberger
thien@unimelb.edu.au

Linda Barwick
linda.barwick@gmail.com