The Wellsprings of Linguistic Diversity (Principal Investigator: Nick Evans, ANU): Project Description

Note: this is the original project proposal. Since the project was almost fully funded, it remains an accurate description of the proposal, except that a delayed start-date (June 30 2014) means that time-lines will be transposed by six months.

Aims

Why are there so many languages in the world? And why do they differ so radically in the way they are organised? Though these are among the first questions linguists are asked, we still have no scientifically-grounded answer to them. The current project will tackle it by carrying out a series of detailed studies of communities at different scales, though with a special focus on small languages. It will investigate the way whole languages diversify out of very small-scale variation, and develop new methods for investigating the interactions of language and social structure in communities chosen for the insights they give into how humans have lived through most of their evolutionary history.

Somewhere between six and seven thousand languages are spoken in the world today. They vary staggeringly in their sound systems, grammatical structures, and the meaning-categories they use to categorise the world. Each represents the outcome of a long process of linguistic evolution during which speakers have developed their languages as complex communicative tools. Strikingly, this process of evolution has largely evolved without conscious design (1,2).

We take these processes for granted, when bathed in our own language. But suppose we encounter an alien linguistic system like Nen (PNG), where the word *ynndandarameng* means '(I command that) you (many people) give the two of us many things, at different occasions/locations in the future'. We can break this up into pieces like the circumfix *yn-....-meng* 'you (many) do it to us two, in the future!', or the reduplicand *-nd-* 'at different occasions/locations'. For each piece of this puzzle, we ultimately require both an account of how it evolved individually, and of why some languages (e.g. Nen) have evolved grammatical devices with these meanings while others (e.g. English) do not. New bits of grammar almost invariably start life as despised 'lazy' shortcuts, like English *gonna talk* for *goin(g) to talk* which is on its way to becoming a future tense comparable to its respectable French future-tense counterpart *parlerai*. Thus do complex structures emerge from a crucible of socially-laden variation.

Each language presents a somewhat different way of being human. Each has been tuned to the culture(s) that evolved it (3). Our minds and our cultures are 'built for diversity' (4,5,6): psychologically, in that we can learn any language and wire our minds to use the concepts it demands, and socioculturally, in that we can evolve such intricate but wildly different systems. *What processes engender this enormous diversity?* The question is central not just for linguistics, but for fields as diverse as cognitive science, human evolutionary biology, anthropology and archaeology.

This project tackles this problem by digging down to the root causes of how diversity arises. It focusses on three of the most linguistically diverse regions on the planet – Papua New Guinea, Arnhem Land, and Vanuatu, where languages have at most a couple of thousand speakers and usually fewer. To examine the impact of demographic scale it will include matched studies from Samoa (as a complex chiefdom, with a quarter of a million speakers for Samoan) and for two major world languages (English and Spanish) where we can examine variability at a number of levels – in small communities (e.g. the small town of Parkes, NSW), at the national level (e.g. Australian English, Andean Spanish) and at world level ('pluricentric' English or Spanish, as spoken in various countries).

The world distribution of linguistic diversity is wildly skewed. Vanuatu counts 105 languages compared to just one in the Korean peninsula, despite having only 0.3% of the population of the two Koreas and 5% of their combined land area. And the island of New Guinea, with just 7.5 million people, boasts a comparable number of language families to the whole of Eurasia with its 4.6 billion people, from Ireland to Japan, from Siberia to Malaysia. Confronted with these facts, the curious often pose two vital questions to linguists.

Firstly, *why are there so many languages in some parts of the world compared to others*? This is particularly relevant to our own region (7), which is home to a third of the world's languages, includes three

of the top five countries by number of languages (#1 PNG, #2 Indonesia, and #5 Australia), and the world's most linguistically diverse country measured by languages per capita (Vanuatu).

Secondly, *why are languages so different from one another in some parts of the world* (New Guinea, the Caucasus) *and much more similar to each other in other areas* (Western Europe, Australia, Vanuatu)? This is an independent question from the first, since 'diversity' – the number of species (in biology) or languages (in linguistics) is different from 'disparity', the degree of difference between them (8). Beetles, for example, are high on 'diversity' with huge numbers of species, but low on 'disparity' because their body plans are relatively similar. Neighbouring languages in Southern New Guinea, like Nen and Idi, differ as much as Spanish and Basque, and are hence high in disparity (9); neighbouring languages in Vanuatu or much of Australia are more like Spanish and Italian, showing numerous parallels of structure – in other words, low in disparity. Thus, we need to examine variation in levels of *linguistic disparity* as well as in *linguistic diversity*.

Background

The mechanisms producing linguistic diversity and disparity, and their uneven global distributions, have not been considered core scientific questions in linguistics. This contrasts with the situation in evolutionary biology, where the mechanisms producing species diversity have been a central concern since the 1970s (10). Sophisticated bioinformatic methods are increasingly used to test models through computational simulation. Linguists' disregard for the question is indicated by the fact that almost all such studies have been carried out by non-linguists (predominantly evolutionary biologists), whose publications on the question have have aroused great interest.

Here are some examples of studies in this vein. There are strong inverse correlations (11, 12) between latitude and number of languages – with the most languages in countries on or near the equator (e.g. PNG, Indonesia, Brazil, Nigeria). Likewise, the range size for both cultures (as measured by languages) and of biological species decreases as one approaches the equator (13, 14), a relation known in biology as Rappaport's Rule. However, formulated in this way, this is at best a *distal cause*. Distance from the equator increases something else (fertility, growing season) which increases resource-availability, which favours the development of small groups (e.g. self-sufficient clans) which mark their identity by differentiating their languages more than groups elsewhere would. Only at this last link in the causal chain can we actually observe differential processes of language diversification taking place, yet it is precisely here that we lack the necessary data. It is at this level of *proximal causation* that the present project is aimed.

Various other correlations have also been put forward in recent years. These include

(a) relative time-depth – languages change through time, so the longer an area has been inhabited the more languages there should be (15),

(b) size of speech community – small speech communities evolve different types of linguistic structure (16, 17), and simulations show they undergo more rapid language change (18, 19),

(c) political complexity – the greater the level of political complexity, the larger number of speakers per language (20).

(d) the wish to exclude other groups from gaining information (21)

Again, each of these hypotheses proposes distal causations across large data-sets, e.g. databases on number of languages, number of speakers etc. Statistical trends sit alongside significant exceptions, which likely reflect the relative weighting of different factors. For example, Vanuatu and Samoa have similar populations, are at comparable latitudes, and have both been settled for a broadly similar timespan, roughly three millennia. Yet Vanuatu has 105 languages to just one in Samoa. Here it is likely that time-depth is trumped by political organisation: stratified hereditary chiefdoms in Samoa, simple non-hereditary chiefdoms in Vanuatu. One challenge of these studies, then, is how to weight a number of factors which work in different directions.

We know from informal observations by fieldworkers in both Australia and PNG that linguistic differences are easily coopted to mark local identity, such as clan membership (22, 23, 24, 25, 26, 27). It has also been pointed out (28, 29) that obscurity (difficult grammatical rules, irregular formations) gets harnessed for group-differentiation processes in small 'eseterogenic' groups in New Britain.

Likewise, an analysis of over 2000 languages (16) found higher levels of morphological complexity in languages spoken by small speech communities. This tallies with recent claims by a number of invesigators

that unusual structures (exhibiting greater 'disparity' with respect to global norms) are commoner in small speaker-populations (17, 30, 31, 32, 33, 34).

One explanation that has been advanced for this is that speakers in such settings can draw on a wider range of mutual knowledge, making communicative shortcuts which facilitate the grammaticalisation of such details as kinship information (3, 35) or complex demonstrative systems (36). Further, widespread multilingualism produces a different semiotic in which sounds, words and grammatical items are positioned in a multilingual choice-set (cf 37 for Amazonia). Establishing one's linguistic identity, in this complex semiotic space, may be achieved by highly unusual changes, such as the exchange of masculine and feminine markers in Buin (38,39), the promotion of the rarest gender to the standard one in Iwaidja (40), the propagation of rare processes like vowel metathesis in Hawu (41), or the blending of elements from two neighbouring languages (42).

In terms of existing work, therefore, we have some broad, quantified studies at macro-level, examining correlations between two variables at the worldwide level over a sample size of thousands, and at micro-level we have intriguing but unsystematic field-based observations.

Innovation

What we lack from existing work is the crucial missing step – the 'smoking gun' of speech communities studied in detail, harnessing data from matched, intensive case-studies. It is at this step that we can see how variation between speakers arises, how it gets invested with social meaning (e.g. as emblematic of one clan, dialect, or emergent language) and of how processes of linguistic fission (comparable to speciation processes in biology) actually occur. The present project innovates by stepping in to fill this gap.

By using matched data and manipulating relevant distal factors (demography, social organisation, time-depth) we can test their impact at the level of the actual speech community, which is the arena where change and diversification ultimately starts out. In order to do this, we need to investigate four sub-questions.

Q1. Can we discover a relationship between macrodiversity and microdiversity?

In other words, can we detect, in progress, the micro-processes that engender these macro-effects, by looking at differential levels of variation within speech communities (cf 43, 44)? This will allow us to put the proximal causes of diversity under the microscope by gathering detailed data on variability in real speech communities.

The formulation of this question follows from the demonstration by Labov (45, 46, 47), and numerous other studies (48, 49, 50) that we can study language change by looking at language variation statistically. Diversity of languages results from the accumulation of thousands of specific 'replacement' changes. And each specific 'replacement' is preceded by a transitional phase where two variants contend. Labov showed we can study change in progress, by focussing on statistical variation among these rival forms. This includes both the precise conditions under which variants are produced (e.g. by streamlining of pronunciation, and regularisation of paradigms), and the way different social evaluations then favour the adoption of one variant over another. The variation we examine may be between different speakers according to their class, ethnicity or gender, by the same speaker in different settings (casual to formal), or in different grammatical or phonological environments. All these impact, for example, on the relative frequency of *-ing* and *-in* (cf *gettin'* vs *getting*) in most varieties of English, from Norwich (51) to Cessnock in Australia (52).

In principle such Labovian methods can measure variability inside any speech community. However, in practice they have largely been confined to large languages (53) – partly because we need good descriptions of the languages in place before we can identify the language elements exhibiting variation, and partly because funding for work on endangered languages over the last two decades has not seen the documentation of variation as a priority. The tension generated by this neglect of small-scale societies is encapsulated in two contradictory quotes from Labov. On the one hand, there is a 'uniformitarian principle' (54), which implies that 'the mechanisms of linguistic change that operate around us today are the precisely the same as those which operated even in the remote past' (17). On the other, we must be 'wary of extrapolating backward in time to neolithic preurban societies' (45).

A truly integrated model of language variation and change needs to build in interactions between social and linguistic processes. Our hypothesis is that we will find higher levels of variability inside small-scale speech communities – that this is a crucial ingredient that has generated the large numbers of

languages found in the world in today. To assess this we need to gather matched data across speech communities of various scales – something that has never been done before.

Q2. If there is a relationship between microdiversity and macrodiversity, is this due to differences in the variability of production, in the variability of evaluation, or in both?

For example, in small communities that are highly multilingual owing to out-marriage with speakers of other languages, speakers may transfer features of one of their languages into the way they speak another. This will generate inter-individual variation. But unless these forms catch on (through prestige, or as identity markers) they will not be propagated socially. In larger communities there may be a relatively evenly spread background of variability, whereas in small clans in Arnhem Land or autonomous villages in Vanuatu each community can select in its own way from the pool of inter-individual variation – there is no centralising norm (as in larger polities) which selects in the same way across a broader population. To understand the full dynamics, we need to know both what people say, and how they evaluate what is said by themselves and others. Getting data on both production and evaluation is therefore a crucial part of our design.

Q3. Are there social factors which engender diversity in some speech communities and retard it in others?

Various authors (28,29,38,39) have suggested that in small-scale, autonomous communities which use language as an index of inner group membership, there are heightened pressures favouring diversity and disparity. Stanford (54,55,56) has recently shown that clan membership is the dominant factor in predicting cross-individual variation in tonal differences in the minority Sui language of Southern China. At the same time, if speech communities are integrated into regional systems it is possible for larger social categories to transcend individual language boundaries in shaping change. An example is the impact of the patrimoiety system in Northeast Arnhem Land where a key phonological variable – vowel-final vs consonant-final – is determined by moiety affiliation, at a level of social aggregation higher than the language (57, 3).

Small-scale speech communities have been the norm for human languages through most of human history, until the advent of the neolithic when aggregation into states, then nations, began to occur. In general, small-scale speech communities are:

- multilingual, creating multiple sets of norms through cross-language transfer, and high levels of metalinguistic awareness (i.e. explicit awareness and articulation of linguistic facts)
- exogamous (i.e. marry out, so parents speak different languages),
- prone to harness linguistic difference to signal intricate social categories
- characterised by high levels of shared knowledge by their members,
- egalitarian, allowing individuals to assert their own distinct norms,
- susceptible to propagating innovations from influential individuals because
- potential adopters are almost all in face-to-face contact with the innovator

What is the impact of these various factors? To answer this, we need case studies that vary crucial social factors such as the above, so we can investigate their impact on micro-variation at the level of the speech community. We also need to gather detailed information on speakers' social backgrounds and linguistic biographies, as well as the social networks they participate in.

Q4. Do situations where structurally disparate languages are in stable, intimate contact produce greater levels of micro-diversification and micro-disparification?

In other words, are the processes of diversification affected not just by the social setting, but by the repertoire of existing language patterns which are fed into processes of learning, using and categorising language in such communities. Our hypothesis here is that the greater the pool of structures which are available, the greater the amount of diversification and disparification we will find at the micro level. To test this we need to vary the degrees of multilingualism found in the communities we are studying, as well as the degree of structural disparity between the languages which are in contact.

To answer the four subquestions above we need to undertake a linked series of case studies that meet a number of requirements. Table 1 details the case studies to be examined in this project, which permute key

variables of demography, culture, degree of language difference between marriage partners if applicable, and overall scale of the language.

A project of this type has not been attempted anywhere in the world, and it needs to be done *at this point in history while a few places remain where this is still possible*. Rapid changes in global language ecology mean that it is increasingly hard to find speech communities, like those chosen here, which are small, thoroughgoingly multilingual, and lack the eroding effects of a large lingua franca. This makes the time-window for studying traditional small-scale speech communities extremely narrow: 'long before a language has reached a point of noticeable moribundity, the sociolinguistic setting of the community has usually been changing, making it difficult to gauge features of language variation and change that may have been present when the language was healthier' (53).

Table 1: Case studies in the project

Site	Languages	Feature
3 small-scale		
Australia: Arnhem Land	Bininj Gun-wok, Dalabon (Gunwinyguan); Mawng, Iwaidja (Iwaidjan)	Small languages in a multilingual mosaic, asymmetric relation between Bininj Gun-wok and Mawng, long in contact but not closely related, though one language closely related to each (Dalabon, Iwaidja) to provide comparative triangulation; traditionally hunter-gatherers
PNG: Western Province	Nen, Nambu (Morehead-Maro Family); Idi, Ende (Pahoturi River Family)	Two closely-related pairs each belonging to very different unrelated families; traditional swidden cultivators with intermarriage between language groups followed by change of main residence, usually for woman but sometimes for men
Vanuatu: South Pentecost	Sa (Oceanic; Austronesian)	One language with great dialectal variation, within a zone of many closely-related languages; village agriculturalists
1 mid-scale		
Samoa	Samoan (Oceanic; Austronesian)	Sole traditional language of a unified, monolingual, and stratified polity until European contact. Comparable environment and settlement time to Vanuatu allows us to hold environmental and time variables constant while varying political scale
2 large-scale		
English	Australian and New Zealand	English will be taken as a representative large language for which we have good variationist data. Material will be gathered at two levels: a small rural community and at sample points across Australasia
Spanish	Latin America (Colombia, Peru or Chile)	A second pluricentric world-language with well-studied variations at national and local levels as well as impacts from indigenous American languages

To achieve the needed synthesis between documentary, variationist and modelling approaches, I will bring together researchers from quite different traditions. My own strengths are in recording undocumented languages, writing grammars and other analyses, and fitting what we discover into an overall comparative framework of how languages work. Two part-time postdocs (Dr Murray Garde and Dr Ruth Singer) will bring their deep knowledge of Bininj Gun-wok and Mawng respectively. Sociolinguistic expertise will be ensured through the participation of Professor Miriam Meyerhoff (Auckland; Vanuatu expertise) as an annual visitor plus input from Professor Catherine Travis (ANU; Spanish expertise). Modelling expertise will be brought in by the postdoc, with substantial extra input from DECRA Dr Simon Greenhill and regular ANU visitor Professor Russell Gray (Auckland). The annual forum debates (see Dissemination) will bring in an even broader set of contributors to this new synthesis.

Approach

Since our hypothesis is that the seeds of macrodiversity can be found in microdiversity, and that specific social conditions (including small scale of speech community, and particular types of egalitarian multilingualism) promote microdiversity, we need methods that can

- *detect and quantify differential levels of diversity*,
- *are as applicable to small as to large speech communities*,
- *gather the requisite social background data*,
- *measure social evaluation* (what people consider 'good speech') as well as *production* (what they actually say)
- *sum the findings from a matched number of variables*: since the total semiotic dispersal across individuals (i.e. the sum of all the ways they vary their speech) is too large to measure, we use a 'variation core' instead to see how far individuals vary from each other (in both production and evaluation) across the multidimensional space of their speech community
- *model these situations dynamically through computational simulations*, in terms of both linguistic variation and social factors, as well as testing the situations we sample against longer-term trajectories of change and diversification

The most difficult part, from an empirical point of view, is getting the detailed data from small-scale speech communities, which is one reason why a matched study of this type has never been done. For it to succeed, detailed expertise in the relevant languages is vital, since this will enable data-gathering to focus rapidly on loci of linguistic variation as well as greatly amplifying the efficiency of transcription (fluency can speed up transcription by a factor of 10!) and smoothing the logistics of field placement. ***Each of the five doctoral students will work in a small-scale speech community, on a different language, and will be guided by experienced linguists with knowledge of the languages and communities***: Evans, Garde and Singer in Arnhem Land, Evans in Southern New Guinea, Garde in Vanuatu, and Prof Catherine Travis (ANU) for Spanish. For Samoan we will draw on the expertise of Emeritus Prof. Andrew Pawley at the ANU plus collaboration with the National University of Samoa. *The roles of the various salaried researchers on the project are given in more detail in the section on Budget Justification.*

In more detail, our approach will proceed as follows:

(a) for each language we will gather ***quantifiable data on variation*** across a range of levels (sound, grammar, vocabulary, meaning). Some of this will focus on known shibboleths in the community, to make sure we get variables with maximum symbolic value. At the same time, for the sake of comparability, we sample a basket of typological variables likely to vary in comparable ways in most languages (e.g. 'she gives him the book' vs 'she gives the book to him'). This data will give us a representative sample of key patterns of language use for around 20 individuals per linguistic variety, stratified as far as possible for clan, age and gender. For a subset of individuals we will also get longitudinal data across a 5-year period, since recent work (59, 60) has shown adults to be much more dynamic in their language use than had previously been believed.

We will employ a number of methods to record and analyse individual speech in ways that yield comparable data sets. These include parallel naturalistic descriptions (family and village descriptions), word lists, elicitation sets of standardised grammatical features, a broad-spectrum story task designed to induce a range of speech styles (picture descriptions, dialogue, monologue), excerption of features from more free-form recordings (narrative on a freely chosen topic), and participant observation through general daily interaction. These different tasks bring out different degrees of self-monitoring vs relaxed speech.

The above speech-probes will enable us to develop a 'speech probe' for each individual that encapsulates their personal version of the language, including their own envelope of variation, based on around 20 minutes of transcribed material per person. Comparable materials will be gathered in each language in which speakers are fluent.

(b) ***integration of linguistic data-gathering with detailed information on the social background of all speakers sampled in (a)***: genealogies, clan affiliations, life histories, plus social network data will be recorded. This will allow us to identify individuals whose life-histories make them particularly interesting for the study of language change through the lifetime (e.g. women who have married into another speech community). Material will be transcribed in the individuals' language(s), thus yielding linguistic data at the same time.

(c) ***getting information on social evaluation as well as on production***, for all variables in (a) and across all categories in (b). In other words, for each variable, we don't just find out who uses it, but how it is evaluated (e.g. as innovative, conservative, characteristic of another group) across a sample of other members of the speech community. Part of this involves eliciting commentary on variation we have detected in (a): if we notice two variant pronunciations, we ask for judgments on these, typically eliciting opinions like 'both are OK', 'one is the shortcut way', 'that's how people from X village pronounce that word'. But, to make sure we detect as much socially-significant variation as possible, we will also play back controlled samples of stories and other material (including older recordings where available), and ask listeners to comment on any unusual features they hear.

(d) ***developing computational models.*** These are an integral part of making sense of the empirical data in (a)–(c). For example, they enable us to quantify structural dispersal across individuals in speech communities and measure the total distance between them (how many steps it takes to change one system into another). We can thus measure the relative degrees of variability across communities at different scales. We can also segregate out the contributions of production from evaluation, model the effects of different social (and acquisition) scenarios, and so forth.

An important part of the modelling will concern the emergence of 'rara' – unusual linguistic phenomena like the dual-non-dual organisation of number in Nen (9), widespread deponent verb agreement in Iwaidja and Mawng (61, 62, 63, 64) – addressing the question of whether they are equally likely to emerge in all types of speech community, or more likely to evolve in multilingual situations where the cost of learning and maintaining 'marked' structures is warranted by their value in signalling membership of one speech community rather than another.

For each of the four elements above, there will be close interaction between the design of fieldwork protocols and the development of the computational modelling, as set out in Table 3 below.

Table 3: Project Timetable and components.

Year	Canberra-based project work (numbered by month)	Fieldwork	Forum dialogue; themes
2014	1-3. Recruit and appoint most personnel (Project Officer, 2 PDRAs, 2 PGRs). 4-5. Develop first-pass field production protocol. 6-9. Field season I 10-12. Recruit first summer scholars. Archiving and transcription from first field season. Recruit PGRs3,4,5.	3. Garde and Evans make initial setup trip to Vanuatu. 6-6-9. PNG and Arnhem Land teams scope out variation, set up project sites, get genealogical data, record and transcribe initial material based on first production protocol. 11. Evans makes short visit to Samoa to negotiate field setup for PhD 4 with NUS.	4. Methods for studying variation in small-scale speech communities.
2015	1-5. Transcription and analysis of data from first field season. Develop definitive production protocol and identify variables to be tested in trial evaluation protocol. Input of initial field-derived variables into computer modelling. 6-9. Field season II, 10-12. Archiving and transcription from second field season.	1-2. Summer scholars (Eng) 2. Garde & Evans get trial Vanuatu production data 6-9. Field season (PNG, Arnhem Land, Vanuatu, Samoa, Colombia). Final production protocol run for c. 50% of individuals, plus initial data for evaluation protocol. In-field transcription.	4. Modelling variation, diversification, and evolution: scale, semiotic and process

Year	Canberra-based project work (numbered by month)	Fieldwork	Forum dialogue; themes
2016	1-5. Transcription and analysis of data from second field season. Synthesis of production protocol data for Arnhem Land and PNG. Develop definitive evaluation protocol. 6-9. Third field season, 10-12. Archiving and transcription from third field season.	1-2. Summer scholars (Eng) 6-9. Field season (PNG, Arnhem Land, Vanuatu, Samoa, Colombia). Final evaluation protocol run for 50%, gaps filled for production data plus further in-field transcription	4. Multilingualism and variation
2017	1-5. Transcription and analysis of data from third field season. Synthesis of evaluation protocol data for Arnhem Land and PNG. Preliminary testing of first-pass data against computational models 6-9 Field season, 10-12. Archiving and transcription from fourth field season.	1-2. Summer scholars (Eng) 6-9. Field season (PNG, Arnhem Land, Vanuatu, Samoa, Colombia). Final evaluation protocol recordings completed, in-field transcription.	4. Variation and integration at system level integration at system level
2018	1-4. Transcription and analysis of data from fourth field season. Synthesis of evaluation protocol data for Vanuatu, Samoa, English and Spanish. Full testing against models. 5-6. Field season. 6-7. Archiving and transcription from final field season. 8-12. Final analysis, model testing and write up	1-2. Summer scholars (Eng) 5. Field season (PNG, Arnhem Land, Vanuatu, Samoa, Colombia). Checking final data and gaps.	11. Project synthesis and dissemination

Notes:
1 the modeller will be Canberra-based throughout and will continue developing and testing the modelling during the field seasons each year.
2 because of the 1-year lag in the 2nd PGR cohort starting (i.e. 3 in 2014, 2015) the fieldwork schedule for the second cohort will run later –this is not shown here.
3 the exact time in the field will not be the same for all investigators or in all years: PhDs will stay the longest, in the second and third years of their projects, since they need more time to learn the language and also will be gathering the largest amount of data
4 transcription of field data is time consuming and requires several passes – typically transcription of the hardest parts in the field, where speakers can assist, followed by transcription of easier parts back in Canberra. This then often reveals further problems needing follow up the next field season.

CHOICE OF CASE STUDIES. Each case-study has been chosen so as to permute key cultural and linguistic settings. The major variables are given below.

- ***traditional economy*** (hunter-gatherers in Australia, small-scale agriculturalists in Vanuatu and Southern New Guinea),
- ***rules of marriage***, ***residence and social identity*** which impact upon the languages children are exposed to and their identification with multiple linguistic codes: clan and language exogamy with bilateral descent in Western Arnhem Land, sister-exchange and predominant village/language exogamy with patrilineal descent in Southern New Guinea, linguistic endogamy with bilateral residence in Vanuatu.
- ***degree of relatedness*** of the languages, which influences the degree of structural difference between the linguistic variants being thrown together, and hence the range of variant structures likely to be on the table for incorporation into local grammars. For *Southern New Guinea*, unrelated families are involved, bringing very different types of language into longstanding, intimate contact. In *Arnhem Land* we see contact between two ultimately-related but very different groups of languages: the Gunwinyguan and Iwaidjan languages diverge in many ways and the Iwaidjan languages are 'un-Australian' in many ways (65): there is thus an interesting pattern of deep divergence reconnected by recent contact. The *Vanuatu* case is different again – the languages of Vanuatu are relatively closely related, but frenetic diversification in the last three millennia has created the

world's most linguistically diverse zone in terms of number of languages per capita, but without producing wide typological divergence between languages.

CONTROL STUDIES. As stated above, a central hypothesis of this project is that there is proportionally more intra-community variation in small-scale multilingual speech communities than in larger-scale communities.

To evaluate this we need comparable data sets from other speech communities, matched as closely as possible with the four communities above, but varying in terms of scale, contact and social organisation. To this end, we will obtain data from another, much larger Oceanic language – Samoan (c. 250,000 speakers), as well as from English and Spanish, two pluricentric world languages. By comparing Samoan and Sa in Vanuatu, we juxtapose two relatively closely-related languages that have been in the Pacific for comparable periods (around three millennia) but in societies which have seen radically different trajectories of political scale (small-scale polities in Vanuatu vs large centralised chiefdoms in Samoa; cf 66, 67, 68) and attendant linguistic diversity.

For the control studies we will gather data at two levels: (a) within a small-community of village size, so as to measure dispersal within the day-to-day community, (b) across widely-dispersed sample points at national levels (all three) and international levels (English and Spanish) pegging out a fuller range of variation across the larger speech community.

DATA-GATHERING FRAMEWORK. Data-gathering for the project will involve intensive and prolonged fieldwork in the relevant communities by the various members of the investigating team (Evans, postdocs, research associates, PhDs) all of whom will already have, or will need to develop, fluency and analytic sophistication in one or more of the languages of their setting. Typically each field linguist will spend 18 months at their field site, split over 4 field trips, complemented by intensive analysis back in Canberra.

The planning and execution of linguistic fieldwork is a well-understood process in which Evans has vast experience, and various technical innovations make recording and transcription easier. What is original to the current proposal is the close attention to variation. This means that data-gathering will be structured by language background (bilinguals and multilinguals, second-language speakers) and life history. This will give varying alignments of speakers with clan identity, local residence, and 'referee design' (69), i.e. the resonance of forms they use with norms from elsewhere. We hypothesise that the most influential innovators will be those with a cosmopolitan refereeing status (generating maximal variation through cross-pollination of their multilingual repertoire) but full clan identity and residence status (hence a social platform that gives status to their forms).

MODELLING METHODS. After transcription and analysis, data for each individual will be coded as values in multifactorial space, where each dimension is a given variable. This can be done whether we are dealing with grammatical constructions (*I like her* vs *she pleases me*), meaning ranges (does 'father' include 'father's brother', as in Dalabon), phonemes (is there a phoneme ŋ (=Eng. *ng*)?), word-structure (can ŋ occur word-initially, as in Idi *ŋən* 'I'), lexical choices (is the word for 'no' *kayakki* or *burrkyak*?). Variation may be intra-individual (I say both *coming* and *comin'*, variably), cross-individual (some say one, some the other, or say both but with different frequencies). Sets of codings like this allow us to position 'language slices' in multifactorial space and measure how close or distant they are – whether we are talking about 'languages' or individual varieties. We can do the same for each language someone speaks: someone speaking Nen, Nambu and Idi will generate three 'language slices', one per language. We can do the same for 'norm slices' – sets of judgments made by an individual.

We can create a separate multidimensional space representing people's social positioning (e.g. in Arnhem Land, X clan, Y matriphratry, Z community of upbringing) and cross-measure this against their linguistic characteristics. And we can build mathematical models of social networks in which individuals are connected in various ways – by kinship, coresidence and so on – and again examine the consequences for dispersal and convergence of all linguistic features of interest.

Modelling will not just be done after the fieldwork, however. Because of the vast amount of data we will gather, it is important to have a lean and focussed approach to what will be collected. The modeller (PDRA2) will work closely with the team in refining the fieldwork protocols at each stage, so that they are maximally well-matched to the models which need testing.

Significance and National Benefit

BREAKTHROUGH SCIENCE. The project will move the field in a bold new direction likely to mark a turning point in the way we study language, variation, diversity and change. Australia has had a world-leading reputation as the 'dawn-land of today's linguistics' (70) for its work on little-known languages, but by now the approaches it grew famous for in the 1980s and 1990s have been widely adopted worldwide and it is time to innovate in new ways. This project will renew Australia's leading reputation in linguistics by asking questions which are at the same time central and neglected, about the causes of linguistic diversity and disparity and why they vary so radically in different parts of the world. To answer them, we will expand the methods and foci of language documentation to look at variation, and combine them with powerful computational modelling to see how actually attested variation and change across individuals scales up to the diversification of whole languages under different patterns of intermarriage and multilingual engagement. For the first time, we will be able to examine the emergence of linguistic diversity at the micro-level in the sorts of small-scale societies that gave rise to the multitude of languages spoken around the world. Though there is increasing work on modelling the effects of change rates at the macro level over hundreds or thousands of years (71), noone has yet systematically examined the micro-processes that give rise to the macro-patterns we see at higher levels.

Labov (47) concludes his magnum opus *Principles of Language Change* by invoking the 'Historical Paradox' that '[t]he task of historical linguistics is to explain the differences between the past and the present; but to the extent that the past was different from the present, there is no way of knowing how different it was'. We cannot travel back into the past. But by studying what is going on in small-scale multilingual speech communities that can provide us with plausible models for how our distant ancestors lived and spoke, we have our best chance to get some glimpse of the factors that gave us the thousands of languages spoken across the world today.

UNDERSTANDING OUR REGION AND OUR WORLD. Australia is in a paradoxical situation: our traditionally monolingual, anglophone heritage is at odds with our location in the most linguistically diverse part of the world, and with growing bi- and multilingualism in our national population. Whether one looks inside to our own indigenous populations, or nearby to Pacific neighbours like PNG and Vanuatu, the message is the same: multilingualism and linguistic difference lie at the hearts of people's lives.

For such communities, the sort of variation focussed on in this project plays a vital part – whether seeking to maintain their language, develop school curricula or build resources like dictionaries or phone apps. Dictionaries or language teaching resources get stuck on debates about what is the 'correct' spelling or meaning. Speech technology needs to grapple with interpersonal variation for reasons as diverse as obeying voice commands or doing google searches. Smartphone apps or predictive spelling for people messaging in local languages need to be customised to local variation.

Language and languages occupy an ever more central role in the information age. Approaches that can handle variability lie at the heart of the next technological moves (e.g. variable-spelling searchers). Material of great community benefit, such as local-adapted writing systems, dictionaries, archived and commented recordings of the languages, will all be spinoffs of this project:. An ethos of sustainable linguistic research and community benefit has always been part of my research philosophy and each researcher on the project will be encouraged to find their own way of contributing in this regard.

Communication of Results

To maximise the impact of the project on linguistics and surrounding fields, we will:

(a) present in-progress findings at a range of national and international conferences in linguistics and neighbouring fields; all project members will make at least one annual presentation

(b) publish specific findings in appropriate journals in linguistics (*Language*, *Language Variation and Change,* etc.) and beyond (Science, *PLOS One* etc.)

(c) host an annual 'forum dialogue', which will involve two distinguished researchers each giving a thematically-organised 5-lecture series – one scholar in the morning, the other in the late afternoon, with detailed discussion and commentary. Projected session themes are listed in Table 3. Two student scholarships per year will be made available per year for interstate attendance, and video-broadcasts of the lectures will be streamed through the publicity arm of the ANU College of Asia and the Pacific

(d) at the end of the project, produce a series of specific monographs (one per case study) with a leading academic publisher, linked to a central browsable data-base, plus an overall monograph-length synthesis with a leading academic publisher

(e) all primary data will be archived in Paradisec, a digital archive co-hosted by ANU

(f) presentations of interim findings, research techniques and project directions will be made to relevant local institutions and communities in Arnhem Land, PNG, Vanuatu and Samoa

(g) findings will also be communicated to the general public through media interviews, popular articles, podcasts and a regular electronic newsletter all hosted on the project website.

References

Aikhenvald, Alexandra. 2010. *Language contact in Amazonia.* Oxford: Oxford University Press.

Baronchelli, Andrea, Nick Chater, R. Pastor Satorras & Morten H. Christiansen. 2012. The biological origin of linguistic diversity. *PLOS One* 7.10:e48029.

Bell, Allan. 2001. Back in style: reworking audience design. In Penelope Eckert and John R. Rickford (eds.) *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press. Pp. 139-169.

Blust, Robert. 2012. Hawu vowel metathesis. *Oceanic Linguistics* 51, 207-233.

Bybee, Joan and Rena Torres Cacoullos. 2009. The role of prefabs in grammaticization: How the particular and the general interact in language change. In Roberta L. Corrigan, Edith Moravcsik, Hamid Ouali and Kathleen Wheatley (eds), *Formulaic language*, vol. 1: Distribution and historical change, 187-217. Amsterdam: John Benjamins.

Collard, I.F. & Foley, R.A. 2002. Latitudinal patterns and environmental determinants of recent human cultural diversity: do humans follow biogeographical rules? *Evolutionary Ecology Research* 4:371-383.

Croft, William. 2010. Evolutionary Linguistics. *Annual Review of Anthropology.* 37:219–34

Croft, William. 2011. The continuity between crosslinguistic variation and language-internal variation. Presentation at workshop on 'Ecology, Population Movements, and Language Diversity', AFLICO Conference, Lyon, May 2011.

Currie, T. E. & Mace, R. 2009. Political complexity predicts the spread of ethnolinguistic groups. *Proceedings of the National Academy of Science* USA 106: 7339–7344.

Dahl, Östen. 2004. *The growth and maintenance of linguistic complexity.* Amsterdam: Benjamins.

Dixon, R.M.W. 2010. *Basic Linguistic Theory. Volume 1: Methodology.* Oxford: Oxford University Press.

Dorian, Nancy. 1994. Varieties of variation in a very small place: social homogeneity, prestige norms, and linguistic variation. *Language* 70:631-96.

Evans, Nicholas. 1998. Iwaidja mutation and its origins. In A. Siewierska and J. J. Song, eds., *Case, Typology and Grammar: In honour of Barry J. Blake.* Amsterdam: John Benjamins. Pp. 115-149.

Evans, Nicholas. 2000. Iwaidjan, a very un-Australian language family. *Linguistic Typology* 4.2:91-142.

Evans, Nicholas. 2003a. Context, culture and structuration in the languages of Australia. *Annual Review of Anthropology* 32:13-40.

Evans, Nicholas. 2003b. *Bininj Gun-wok: a pan-dialectal grammar of Mayali, Kunwinjku and Kune*. (2 volumes). Canberra: Pacific Linguistics.

Evans, Nicholas. 2007. Pseudo-argument affixes in Iwaidja and Ilgar: a case of deponent subject and object agreement. In Matthew Baerman, Greville G. Corbett, Dunstand Brown & Andrew Hippisley (eds.), *Deponency and morphological mismatches. Proceedings of the British Academy* 145:271-296.

Evans, Nicholas. 2010. *Dying Words: Endangered languages and what they have to tell us.* Maldon & Oxford: Wiley-Blackwell. The Language Library.

Evans, Nicholas. 2012. Even more diverse than we thought: the multiplicity of Trans-Fly languages. In Nicholas Evans & Marian Klamer (eds.) *Melanesian Languages on the Edge of Asia: Challenges for the 21st Century. Language Documentation and Conservation Special Publication No. 5*: 109-149.

Evans, Nicholas & Steven Levinson. 2009. The Myth of Language Universals. *Behavioral & Brain Sciences* 32: 429-448.

Fitch, Tecumseh. 2011. Unity and diversity in human language. *Philosophical Transactions of the Royal Society B: Biological Sciences* 366(1563):376-388.

Garde, Murray. 2002. Social deixis in Bininj Gun-wok conversation. Unpublished PhD Dissertation, University of Queensland.

Garde, Murray. 2008. Kun-dangwok: "clan lects" and Ausbau in western Arnhem Land. *International Journal of the Sociology of Language* 191:141–69.
Gorenflo, L.J., Suzanne Romaine, Russell Mittermeier and Kristen Walker-Painemilla. 2012. Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *Proceedings of the National Academy of Sciences of the United States of America* 109.21:8032-7
Greenhill, Simon. Demographic correlates of language diversity. To appear in Claire Bowern & Bethwyn Evans (eds.), *The Routledge Handbook of Historical Linguistics*.
Greenhill, S.J., Atkinson, Q.D., Meade, A., and Gray, R.D. 2010. The shape and tempo of language evolution. *Proceedings of the Royal Society London B*, 277, 2443-2450.
Hurford, J. 2003. The Language Mosaic and its Evolution. In M.H. Christiansen and S. Kirby, editors, *Language Evolution: The States of the Art*. Oxford University Press. Pp. 38-57.
Keller R. 1994. *On language change: the invisible hand in language.* London: Routledge
Keller R. 1998. *A theory of linguistic signs.* Oxford University Press.
Kerswill, Paul. 1996. Children, adolescents, and language change. Language Variation and Change 8.2:177-202.
Kulick, Don. 1992 . *Language Shift and Cultural Reproduction: Socialization, Self and Syncretism in a Papua New Guinean Village*. Cambridge: Cambridge University Press.
Labov, William. 1972. *Sociolinguistic Patterns.* Philadelphia: University of Pennsylvania Press.
Labov, William. 1994. *Principles of Linguistic Change, Volume 1: Internal Factors.* Wiley-Blackwell.
Labov, William. 2001. *Principles of Linguistic Change, Volume 2: Social Factors.* Wiley-Blackwell.
Labov, William. 2010. *Principles of Linguistic Change, Volume 3: Cognitive and Cultural Factors.* Wiley-Blackwell.
LaPolla, Randy. 2005. Typology and complexity. In J.W. Minett & W.S.-Y. Wang (eds.) *Language acquisition, change and emergence: essays in evolutionary linguistics*. Hong Kong: City University Press. Pp. 465-93.
Laycock, Don. 1982. Linguistic diversity in Melanesia: a tentative explanation. In R. Carle, M. Heinschke, P. Pink, et al. (eds.), *Gava': studies in Austronesian languages and cultures dedicated to Hans Kähler*. Berlin: Reimer. Pp. 31-37.
Laycock, Donald. 1982. Melanesian linguistic diversity: a Melanesian choice? In R. May and H. Nelson (eds.) *Melanesia: beyond diversity.* Canberra; RSPAS pp. 33-8.
Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PLoS One* 5.1: e8559
Mace, R. and Pagel, M. 1995. A latitudinal gradient in the density of human languages in North America. *Proc. Roy. Soc. Lond. (B)*, 261, 117-121.
Maclaurin, J. and K. Sterelny. 2008. What is Biodiversity? Chicago, University of Chicago Press.
Maffi, Luisa. 2005. Linguistic, Biological and Cultural Diversity. *Annual Review of Anthropology* 29:599-617.
McConvell, Patrick. 1985. The origin of subsections in Northern Australia. *Oceania* 56:1-33
Morphy Frances. 1977. Language and moiety: sociolectal variation in a Yuu:lngu language of North-East Arnhem Land. *Canberra Anthropology* 1.1:51-60.
Nettle, Daniel. 1999. *Linguistic Diversity.* Oxford: Oxford University Press.
Ostler, Nicholas. 2005. *Empires of the Word.* London: Harper Perennial.
Pagel, Mark. 2012. War of words: the language paradox explained. New Scientist, Dec. 11, 2012.
Pawley, Andrew. 1981. Melanesian diversity and Polynesian homogeneity: a unified explanation for language. In Jim Hollyman and Andrew Pawley (eds.), *Studies in Pacific Languages and Cultures in Honours of Bruce Biggs.* Auckland: Linguistic Society of New Zealand. Pp. 259-310.
Pawley, Andrew. 2007. Why do Polynesian island groups have one language and Melanesian island groups have many? Patterns of interaction and diversification in the Austronesian colonization of Remote Oceania.
Perkins, Revere. 1995. *Deixis, grammar and culture.* Amsterdam: Benjamins.
Pagel, Mark. 2012. War of words. *New Scientist*, 8 Dec 2012:39-42.
Poplack, Shana and Sali Tagliamonte. 2001. *African American English in the diaspora*. Massachusetts: Blackwell Publishers.
Raup, D.M., S.J. Gould, T.J.M. Schopf & D. S. Simberloff. 1973. Stochastic models of phylogeny and the evolution of diversity. *Journal of Geology* 81:525-542.
Sankoff, Gillian and Hélène Blondeau. 2007. Language change across the lifespan: /r/ in Montreal French. *Language* 83:560-588.

Shnukal, Anna. 1982. You're gettin' somethink for nothing. Two phonological variables in Australian English. *Australian Journal of Linguistics* 2.2:197-212.
Singer, Ruth. 2010. Mawng lexicalised agreement in typological perspective. I J. Wohlgemuth and M. Cysouw (eds.). *Rara and Rarissima: documenting the fringes of linguistic diversity. Empirical Approaches to Linguistic Typology* 46. Berlin: Mouton de Gruyter. 2010, pp. 327-342.
Singer, Ruth. 2011. Typologising idiomaticity: Noun-verb idioms and their relations. *Linguistic Typology* 15, 2011, pp.625-659.
Smith, Ian & Johnson, Steve. 1986. Sociolinguistic patterns in an unstratified society: the patrilects of Nganhcara. *Journal of the Atlantic Provinces Linguistics Association* 8: 29-43.
Stanford, James N. 2008a. Child dialect acquisition: new perspectives on parent/peer influence. *Journal of Sociolinguistics* 12.5:567-596.
Stanford, James N. 2008b. A sociotonetic analysis of Sui dialect contact. *Language Variation and Change* 20:409-450.
Stanford, James N. 2009. Clan as a sociolinguistic variable: three approaches to Sui clans. In Stanford & Preston (eds.), pp. 463-484.
Stanford, James N. and Dennis R. Preston (eds.), *Variation in Indigenous Minority Languages.* Amsterdam/Philadelphia: John Benjamins.
Sutton, Peter. 1978. *Wik: Aboriginal society, territory and language at Cape Keerweer, Cape York Peninsula, Australia.* University of Queensland: Unpublished Ph.D. Thesis.
Tadmore, Uri. Forthcoming. The grammaticalisation of generational relations in Onya Darat. In Randy LaPolla (ed.), The shaping of language: relationships between the structures of languages and their social, cultural, historical and natural environments.
Tagliamonte, Sali A. 2006. *Analysing Sociolinguistic Variation.* Cambridge: Cambridge University Press.
Thurston, W.R. 1987. *Processes of change in the languages of north-western New Britain*. Canberra: PL. B-99.
Thurston, W.R. 1992. Sociolinguistic typology and other factors effecting change in north-western New Britain, Papua New Guinea. In T. Dutton (ed.), *Culture change, language change: case studies from Melanesia*. PL C-120.
Torres Cacoullos, Rena and Catherine E. Travis. 2011. Using structural variability to evaluate convergence via code-switching. *International Journal of Bilingualism* 15(3): 241-267.
Trudgill, Peter. 1972. Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society* 1.2: 179-195.
Trudgill, Peter. 2011. *Sociolinguistic typology. Social determinants of linguistic complexity.* Oxford: Oxford University Press.
Wohlgemuth, Jan. 2010. Language endangerment, community size and typological rarity. in J. Wohlgemuth and M. Cysouw (eds.), *Rethinking Universals: How rarities affect linguistic theory*. Berlin: De Gruyter. Pp. 255-277.