Studying contact without detailed studies of the languages involved: A non-philological approach to language contact

Author(s): Mark Donohue

Proceedings of the 38th Annual Meeting of the Berkeley Linguistics Society (2014), pp. 92-120
General Session and Thematic Session on Language Contact
Editors: Kayla Carpenter, Oana David, Florian Lionnet, Christine Sheil, Tammy Stark, Vivian Wauters

Studying Contact without Detailed Studies of the Languages Involved: A Non-Philological Approach to Language Contact

MARK DONOHUE
The Australian National University

Studies of contact have revealed that all kinds of language material can, in the right circumstances, be borrowed from one language to another. Detecting, describing, and analyzing such situations typically involve the detailed study of at least two languages. An alternative involves detecting contact situations through database analysis. This cannot supplant the detailed work that requires detailed descriptive work in particular fields, but can allow us to examine large enough samples of languages that we can start to better understand, through calibration against known histories and other non-linguistic data types, likelihoods of different 'social contact' scenarios resulting in different kinds of linguistic traces, and also allow for the more targeted investigation of specific areas and language-to-language interactions. I shall describe the method, and illustrate its application in a number of case studies in regions for which we have good samples of language data.

1 Too Many Language (Contact Situation)s, Not Enough Time

In this paper I address the question of how we, as a discipline, might have a chance of identifying more of the language contact situations that exist around the world, and propose steps towards a solution. This will involve calibrating the results from the computational analysis of multivariate and multidimensional data.

I shall discuss the nature of contact and relatedness, and then propose some operational heuristics that, while they do not automate the detection of areality, certainly to make the objective detection of such patterns more objective.

I shall not try to list bibliographically the numerous studies of contact and contact languages; suffice to say that contact appears to be as universal an ingredient in the synchronic and diachronic make-up of languages as is descent, and descriptions of contact are as varied as are the contact situations themselves.

2 'Contact'

In the linguistic sense 'contact' studies have multiplied enormously in the last decade, with approaches ranging from the social to the individual being promoted. A number of consequences of 'contact' have been described and catalogued, and a number of definitions of what might be a contact situation have been put forward. For the purposes of this paper I shall define 'contact' as being:

- a circumstance in which two linguistically distinct societies influence each other;
- facilitated by some portions of at least one society having some competence in the language of the other society;
- detectable and describable on the basis of linguistic data.

A stricter definition of 'contact' will include the requirement that a likely contact source can be identified (eg., Thomason 2009); this is necessary to avoid reclassifying as 'contact' elements of change in a language that arose from language-internal processes of change. Since, for most parts of the world, we do not have written data on ancient languages relevant to those found in contemporary distributions nor do we have means of inferring that data other than through the process of historical reconstruction, which would not in many cases distinguish between contact with a now-vanished language and independent developments, I shall not impose this constraint, useful though it is.1 We have other means of treating changed elements of a language (from a diachronic perspective) as suspicious or not suspicious, vis-à-vis putative language contact scenarios.

Evidence for contact can be detected in many ways. At the outset, evidence for contact may be present in one or more subsystems of a language (with a non-random distribution). A language may display contact in only one sub-part of the lexicon (for instance, lexical items with the semantic category of 'tools'), without affecting the language structurally, either phonologically or morphosyntactically. Alternatively, an extreme example of contact would be thorough-going change throughout the lexicon, in the phonological system, in the forms and functions of bound morphemes, and in terms of the syntactic structures found. This might well be a good representation of what is detected when a community shifts language without strong first-language interference.

We also detect contact effects by the presence of features are are not expected to be the 'natural' result of internal language developments. To fully explore this possibility we must have an idea of what level of variation is 'normal' in a language family, and then explore the appearance of variation that lies beyond this normal range. This reflects the view, present in some work on diachronic linguistics, that language family membership is in part a function of whether the language has a close enough typological 'fit' (see, for instance, discussion in Noonan 2010). While this is not part of the methodology espoused in the classical comparative method, it is in a sense necessary in order to be able to discuss the problematic case of pidgin and creole languages. It is not hard to find examples of the application of this sort of principle within standard historical linguistics. In terms of sound change, we would not be surprised to find correspondences of the sort shown in (1), nor would it be unusual to identify a chain of sound changes of the form shown in (2). Other correspondences, such as are similarly generally judged to be 'natural', or at least plausible, are easy to find (such as (3)).

(1) b:p, p:f, f:h or h:Ø
(2) *b > p > f > h > Ø
(3) b:m:v:w

[1] Of course, when we do have attested records of an ancient language, or the witness of a donor language, we can identify contact by identifying the source(s) of the unexpected features in the borrowing language (e.g, Thomason 2009).

On the other hand the proposal that the history of a language can be better understood by positing a *b > t change would require convincing documentation and argumentation; a change such as *b > e would be even more exceptional, and require even more convincing argumentation, rather than simply noting the putative correspondences. This illustrates the existence of a 'range of variation' that linguists work with when evaluating possible language relationships. Typologically, the same principles can be applied to the kinds of changes found. Verb-initial languages are known to be susceptible to variation in their word order, sometimes leading to a change to subject-initial order; but a change of VOS > OVS is not expected, nor is an SVO > VOS change. Voiced stops in one language might correspond to voiceless, aspirated, prenasalized or imploded stops in a relative, but a correspondence of the form aspirated:imploded is not so expected. We do not expect phonological systems to show such a wide range of variation. Importantly, as noted above, different subsystems of the language can show variable levels of contact-affectedness. The history of the lexicon, that part of a language which the comparative methods pays attention to, with the careful methodology of that approach, need not match the history of the phonology, including accent, innovations, and local areal 'trends', and this too can logically (and attestedly) be independent of the history of the morphosyntax, including both inherited quirks and acquired patterns.

Any or many of these traits can be 'askew,' compared to the expected range of variation for a particular language family or subgroup. This tells us that something other than uninterrupted intergenerational transmission was going on (following Noonan 2010). Adding in the requirement that we find regular correspondences (in the lexicon, phonology and morphology) between languages, we can arrive at a simple factorial typology of language relations, shown in Table 1. While there are numerous 'exemplary' languages, showing the typological profile of their family as well as the regular (sound, morphological) correspondences that cement relationship in that genealogical unit, and while there are many pairs of languages which cannot be considered to be related at all, less attention has been paid to the 'plus-minus' languages, those that satisfy one of the criteria in Table 1, but not the other. Note that Table 1 provides a crude, but operational, means of discussing possible language contact: if a language is a 'plus-minus' language with respect to its relatives, contact should be suspected.

Table 1. Kinds of language relations, defined across two binary dimensions

	+ regular correspondences	–regular correspondences
+ 'typological fit'	genealogically related languages	(contact-affected languages?)
– 'typological fit'	(pidgin/creole languages?)	unrelated languages

The mention of 'typological fit' in Table 1 is already inadequate; having broken down our language data into the lexical, the phonological and the morphosyntactic, there is no reason not to break things down even more. The lexicon can be broken down by semantic categories such as Body parts, Kin terms, Pronouns, Animals, Plants, 'Human' plants, Natural world, Tools, Properties, Colors, Demonstratives, Locations, Numerals, Verbs, Interrogatives, etc., and these can then be independently investigated for contact effects. The phonology can be broken up into different natural classes, such as total number of consonants, total number of vowels, total number of tones; number of plosives, number of nasals, fricatives, liquids; number of high vowels, low vowels, front vowels, front rounded vowels; presence of level

tones, rising tones, falling tones; presence and productivity of contrastive phonation types. The morphosyntax should be broken up into different constructions and categories, which are more salient than 'whole-language typology.' Some categories would include head-marking, dependent-marking; presence of ergative case(s) or accusative case(s); presence of a passive voice, or applicatives; use of verbal agreement, inflectional tense, evidentiality, etc.; marking of gender or clusivity, etc.

In short, to study contact objectively, we simply need to examine exhaustive lexical, structural and typological data for each language in the comparison set. This is true; but it is not a methodological step forward, as it certainly does not propose any time-saving elements. It is tempting to just pre-select the features to examine, but then we run the risk of (consciously or unconsciously) 'cherry-picking' the data to reach a certain set of conclusions (see discussion in Donohue, Wichmann and Albu 2008). An objective attempt to detect contact must examine different sub-systems of a language, and for each sub-system examine as much data as possible; and this involves typologizing languages according to many dimensions of variation, and in a way that allows for rapid (computational) evaluation.

3 Speeding Up the Contact Discovery Process

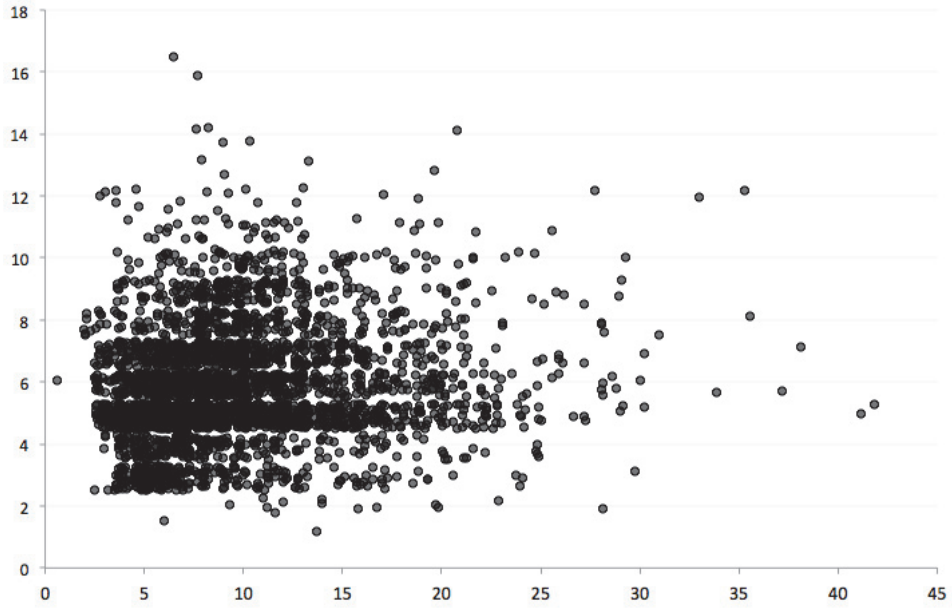
This data can be fed into a clustering algorithm (in this paper I use Splitstree – Huson and Bryant 2006). Such algorithms are designed to take large amounts of data and produce 'best fit' clusters for the input. Importantly, such algorithms cannot automatically detect relationship or non-relationship, but can simply detect a degree of relationship (along the dimensions examined) between pairs or clusters of languages. If the data evaluated is maximal, that is as inclusive as possible of the variation found in the language along the relevant dimension, then any relationship detected between two languages that cannot be attributed to a genealogical relationship must be assumed to represent one of (a) random chance, (b) the reflection of universal tendencies, or (c) the relics of contact.

3.1 Typology and Distance

A vast amount of work shows that linguistic traits are subject to distance decay effects; that is, the further it is between any two points, the less similar they will be, on average. This has been demonstrated repeatedly for lexical similarities (eg., the summary in Donohue et al. 2012). Holman et al. (in press) have shown similar effects for typological traits. Similarly, a vast number of publications shows the correlation of lexical similarity with distance, with complicating social factors (for an only partial list, see discussion in Nerbonne 2009 (and many other works), Donohue et al. 2012).

For instance, in Figure 1 we have a representation of languages (shown as individual dots) classified according to two dimensions. On the x-axis we have a measure of how many oral, egressive stops the languages contrast, and on the y-axis we see the number of contrastive vowel qualities are present; these two variables are approximately independent, as can be assessed by an examination of Figure 1, and so a typology based on these two variables is not vacuous. This is only one of many ways to typologize languages according to non-binary variables in two dimensions, and while crude it clearly represents an improvement over a typological classification measured along only one dimension.

Figure 1. Languages classified by size of plosive inventory and size of vowel quality inventory



It can and should be argued that a category such as 'stops' or 'vowel qualities' is too broad; languages do not, in contact situations, borrow (or lose) a number of vowels, or a number of stops. More appropriate would be to divide the stops into different variables for places and manners; to separate the 'place' variable into actual places (e.g., bilabials, linguo-labials, dentals, alveolars, alveolar affricates) and manners (e.g., voiced, voiceless, ejectives, implosives, prenasalized). When this is applied, the result yields approximately 40 dimensions of variation, with greater or lesser degrees of independence. (Similar decomposition of vowels will yield approximately 30 dimensions of variation.)[2]

3.2 Quantifying Multiple Dimensions of Typological Distance

What is the 'typological distance' between the different stop systems outlined in Table 2? That depends on our coding. If we code the oppositions present in the languages, then systems a. and c. are identical except on the dimensions [voiced] and [prenasalized]. If we code according to the phonemes present, then the two systems differ in six ways.

[2] Note that the approach I am advocating here is not based on looking at actual phonemes (or allophones). Especially given the variation found between individual linguists in coding data, or between separate linguistic communities for coding phonological contrasts, the existence of a contrast is more robust than coding and testing the nature of that contrast. For instance, whether different linguists have recorded a language as showing a /t̪/≠/t/, /t̪/≠/ʈ/ or /t/≠/ʈ/ contrast, all linguists would agree that there is a contrast in place for coronal stops. Similarly, the identity of a phoneme as /ts/, /tʃ/, /ʈʂ/ or /tɕ/ (to name just a few possibilities) is less important than the number of similar affrictes it contrasts with. The dimension of contrast is a more stable feature than the points of contrast in cross-linguistic comparison.

Table 2. Six small (oral) stop inventories

a.	p	t	k
	b	d	g

b.	p	t	k
	b	d	

c.	p	t	k
	mb	nd	ŋg

d.		t	k
	mb	nd	ŋg

e.	p	t	k
		d	

f.		t	k
	b	d	

Table 3. Values of *phonemes* from Table 2 quantified.

	p	t	k	b	d	g	mb	nd	ŋg
a.	1	1	1	1	1	1	0	0	0
b.	1	1	1	1	1	0	0	0	0
c.	1	1	1	0	0	0	1	1	1
d.	0	1	1	0	0	0	1	1	1
e.	1	1	1	0	1	0	0	0	0
f.	0	1	1	1	1	0	0	0	0

Table 4. Values of *oppositions* from Table 2 quantified.

	Bilabial	Alveolar	Velar	Voiceless	Voiced	Prenasalized
a.	2	2	2	3	3	0
b.	2	2	1	3	2	0
c.	2	2	2	3	0	3
d.	1	2	2	2	0	3
e.	1	2	1	3	1	0
f.	1	2	1	2	2	0

When these data are analyzed into networks, the different coding decisions are apparent in the different configurations they generate. While most of the differences between the plosive systems in Table 2 represent structural differences, systems a. and c. are different in ways that would be expected variants across different dialects, or different closely-related languages, varying only in the emicization of prenasalization. In the network based on segment identities in Figure 2 they are as far apart as it is possible to be; a. and c. form a clade only if b. and e. are also included. When coded for oppositions, in Figure 3, a. and c. appear as divergent sisters. Coding for oppositions, then, leads to analyses that reflect structural phonological, rather than surface phonetic, differences in languages. This has good and bad points, but certainly overcomes the between-linguist differences that plague studies based on secondary sources (inevitable in any large comparison). Can we make an informed decision about which of these two coding choices should be made? The answer is no; there is no single universally appropriate way to code, with the choice dependent on what is revealed by the different choices.

Figures 2 and 3. Network analysis of the two tables of data in Tables 3 and 4.

Figure 2. Network analysis based on segment identities

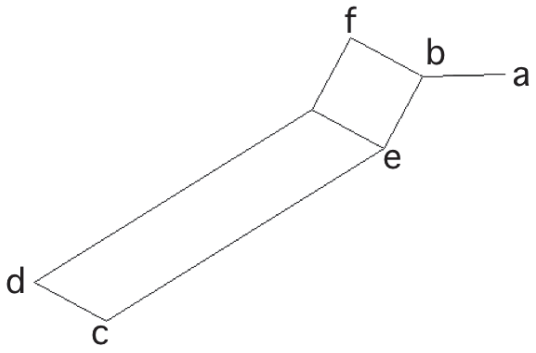
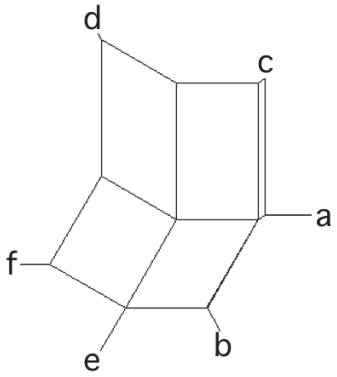


Figure 3. Network analysis based on segment oppositions



When this approach has been applied to the entire phonology, which is a finite system, we can avoid all claims of selectional bias. This is harder with morphosyntax, though the use of a feature set that was not created for a particular purpose also avoids the potential for this charge. One such database is the World Atlas of Language Structures feature set; the Syntactic Structures of the World's Languages project has a similar, overlapping set of features with different excursions. What is important is that the features selected form an objective set, with hopefully near-exhaustive coverage of at least some subsystems.

4 Illustrative case studies

When we apply these principles to whole language phonologies, or to large selections of morphosyntactic data (such as the feature set used in *WALS* – Haspelmath et al. 2005) we can find a series of useful heuristics for detecting contact.[3] We can examine this for a couple of case for which we have known histories, and then exemplify the method with more 'exotic' data.

Using the same simple method, we can arrive at hypotheses about possible contact events even in the absence of a hypothesis about where the contact may have come from (that is, examining data from within one genealogical (sub)group alone), or we can examine possible 'leanings' towards un-affiliated languages, to show where contact-induced change has applied.

4.1 Detecting Contact Without an Out-Group

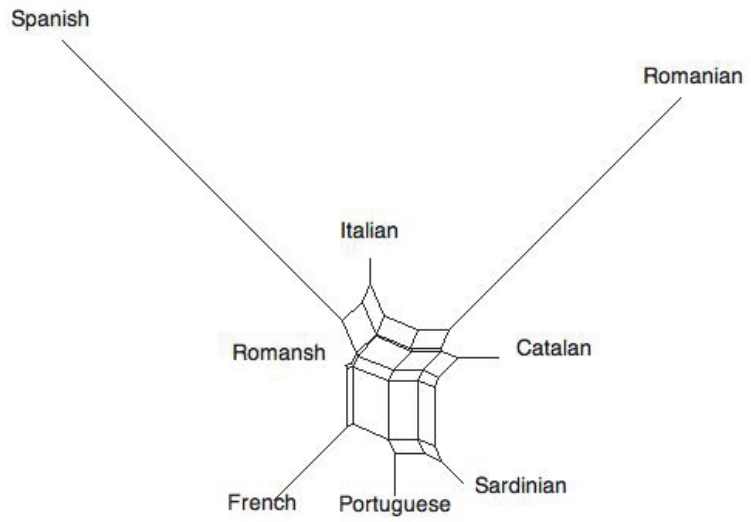
Examining languages that are known to subgroup together, we can easily find which of those languages are more or less 'typical'. While we do not yet have a strong theory about the level of change that is expected, and while such a theory would almost certainly have to be qualified endlessly to take account of local conditions, we can certainly examine relative

[3] I am not claiming that the features included in *WALS* are in some way 'optimised' for typological comparison, but we do note that they have been selected for typological breadth, and even more importantly they are a set of features that has been chosen independently of any particular study (and so cannot be accused of selectional bias – Donohue et al. 2010).

degrees of 'typicality': that is, the degree to which an 'essential nature,' to quote Noonan (2010), is preserved in common between the languages under examination.

In Figure 4 we see the network that results from clustering ~200 morphosyntactic traits (essentially the morphosyntactic features used in *WALS*) for eight Romance languages of Europe. Here we see a relatively tight cluster at the bottom of the figure, in an area containing six of the languages; above that, in the figure, are two outliers, Romanian and Spanish. The tight cluster tells us what the expected range of variation is; Spanish and Romanian show us that some languages of this (sub)family exceed these levels of variation. Based on this data alone, we would *not* expect to find a significantly contact-affected story in the histories of Romansh, Italian, Catalan, Sardinian, Portuguese or French; at least, we would not as strong a level of contact as we are led to suspect for Romanian and Spanish. When we examine non-linguistic historical records we find these hypotheses are confirmed: Spain has a history of long occupation by the Moors, and Romanian is known to have been influenced through contact with Dacian and later Slavic.

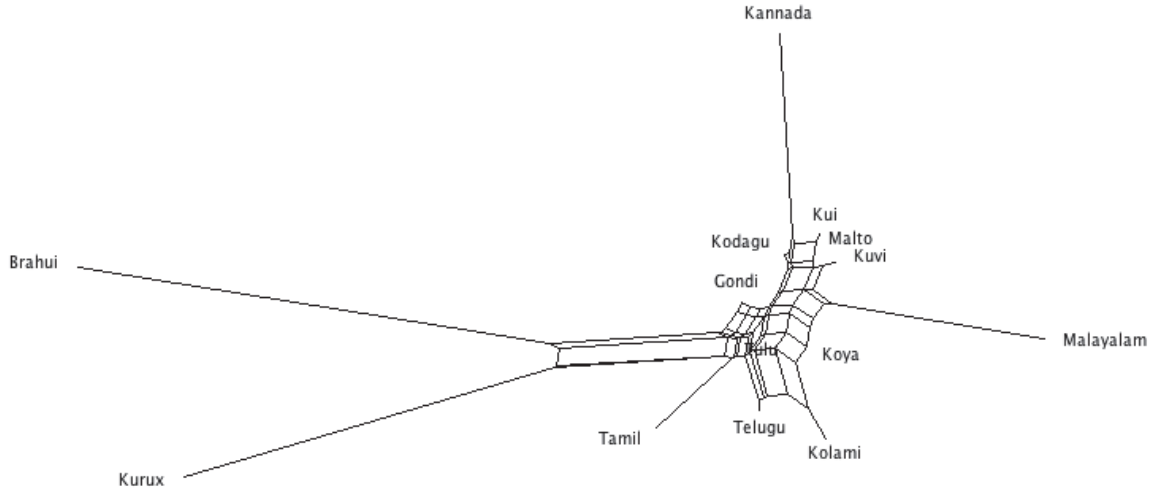
Figure 4. Network analysis of the *WALS* morphosyntactic traits of eight Romance languages.



In Figure 5 we see the results obtained when examining Dravidian by applying a similar methodology to that conducted for Romance. While most members of this family are found in Southern India, three languages, Brahui, Kurux (/Oraon) and Malto are in the north of the subcontinent. While most of the languages are spoken by small, marginalized populations, four languages (Tamil, Malayalam, Kannada and Telugu) are official state languages with long literary traditions. Examining the network it is clear that the languages with long literary traditions, which are also those most influenced by the arrival of Indo-European languages via the Sanskrit written traditions, are divergent from the Dravidian 'canon' (though Telugu stays closest). Similarly we see Kurux and Brahui, the two most northerly languages, as highly divergent outliers. This clearly reflects the social circumstances that have denied them regular inter-Dravidian contact, since the arrival and ascendancy of the Indo-Aryan linguistic ecology in the north of South Asia. Perhaps less expected are the divergent positions of Malayalam

and Kannada; both are major state languages, but have their territories on the west side of India, not the east. There is thus some level of relative isolation, compared to the strongly Dravidian linguistic ecology that prevails in the south and east of India. Again, in terms of contact-induced change, the network in Figure 5 leads us to two suspicions: firstly, that Brahui and Kurux have undergone more contact-induced change than Malto, the other northerly Dravidian language; secondly, that Malayalam and Kannada have been excluded from as extensive contact with the other southerly Dravidian languages; and thirdly, that there has been extensive inter-Dravidian linguistic contact between the other Dravidian languages (those with a focus to the east of India).

Figure 5. Network analysis of the *WALS* morphosyntactic traits of fourteen Dravidian languages.



Note that the hypotheses that we can draw about Romance and Dravidian, on the basis of the clustering analysis of morphosyntactic traits, are hypotheses that can be made without knowledge of local geography or relevant history. Given that for both Romance and Dravidian we have good records, we can confirm the structure-based hypotheses: Romanian *is* geographically isolated from all other Romance languages, and *does* have a strong history of contact (initially with Dacian, and later with Slavic). Spanish is known to have been catastrophically affected by the Moorish invasions, affecting the language both directly and via the Basque-related contact that was enforced by the location of the Spanish court in refuge in the north-east of the country. Amongst the Dravidians we know that Brahui and Kurux are large populations that have been strongly isolated from other Dravidian languages, while engaging in extensive interaction with their Indo-European neighbors. Malto is as isolated from other Dravidians, but is less affected by Indo-European contact, as a small isolated tribe practicing swidden agriculture.

4.2 Traces of Contact Between Language Groups

When Romance is placed in a wider family context, expanding the sample to include with other Indo-European languages of Europe, we see interesting patterns when the clusters are compared to clades in the genealogical tree. We can see clusters that match traditional

subgroups, showing the results of the inheritance of shared innovations (which is the criteria for the classification of the languages as being part of the same subgroup) and also reflecting subgroup-internal contact (reflecting the tendency for genealogical units to share geographic, and social, affinity – see Donohue et al. 2012). We also see evidence for areal convergence that does not match with traditional subgrouping classification. Figure 6, from Donohue (2012), shows the clustering network arising from a comparison of WALS features in 36 languages of Europe. A striking pattern of convergence between the languages of the Balkans region, at the bottom of the figure, where we can see Romanian, Albanian, Greek, Macedonian and Bulgarian forming a loose (that is, highly reticulated) grouping. Bulgarian and Macedonian are only loosely affiliated with this cluster (reflecting the later appearance of Slavic languages in the Balkans), but the affinity of the other three languages is clear.

Figure 6. Network analysis of the *WALS* morphosyntactic traits of 36 European languages.

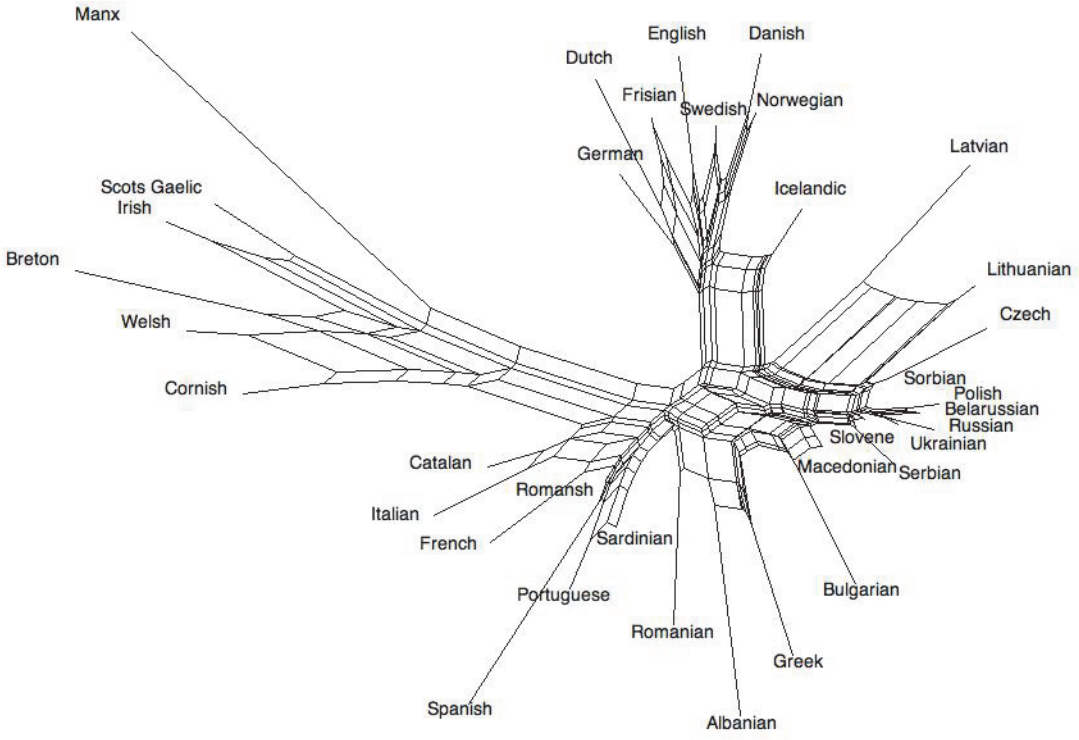


Table 5 shows the traits which are significantly different in the five identified 'Balkan' languages in the sample used here, the results of comparing both phonological and morphosyntactic databases and extracting those features that show a different distribution in the populations in, and out, of the Balkans (following the methodology in Bickel and Nichols 2012). Note that the presence, or number of central vowels, a trait frequently cited as being a feature of the Balkan linguistic area, does not test as being significantly higher in the Balkans than elsewhere in Europe.

In section 4.1 we saw that we are able to identify aberrant behavior within a group, and thus generate suspicions that external factors have played a role in shaping the structure of the

modern language. In this section we have seen that it is possible to identify convergence between different languages in the same region, where contact has taken place over time.

Calibrating against these known histories, we can see that examining the dendrograms allows us to make realistic predictions about broad aspects of social history. From this, we can be confident that the same techniques can be extended to families and areas for which we do not have written histories against which to calibrate.

Table 5. Balkans compared to the rest of Europe

Feature	Balkans compared to the rest of Europe
Plural pronominal suffix	higher
definite and/or demonstrative suffix	higher
prohibitive not normal imperative	higher
hortative morphology	higher
evidentiality, realised through tense paradigms	higher
objects indexed on the predicate	higher
word order use to form questions	lower

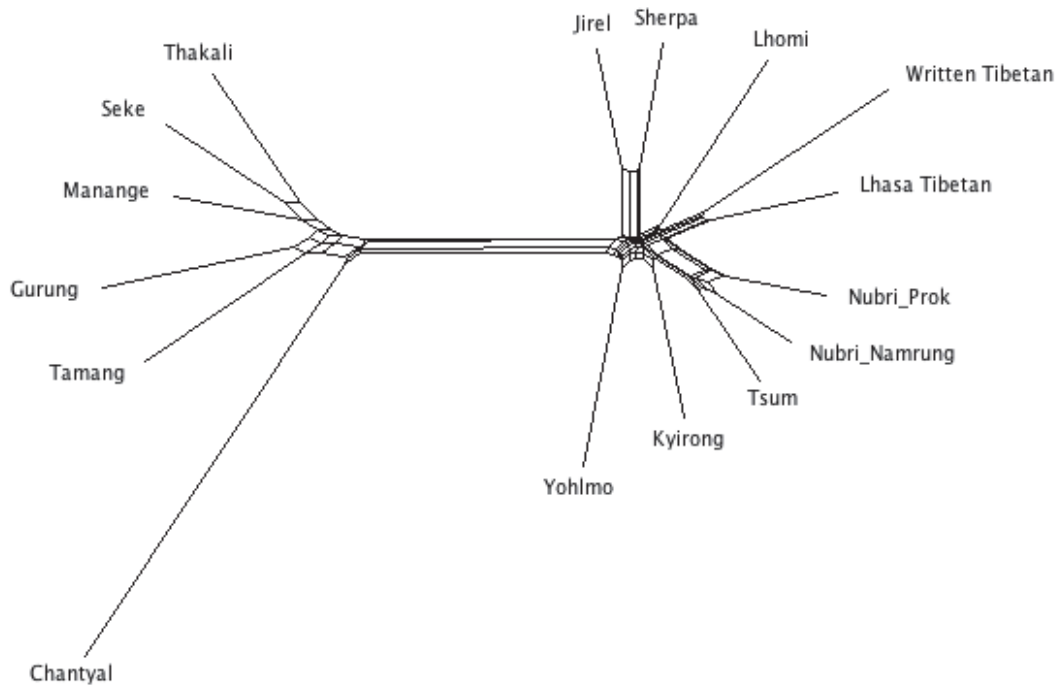
4.3 Contact in the Himalayas

Another, perhaps more traditional, example can be drawn from lexicostatistics. In Figure 7 we can see the clusters that emerge when we examine the lexicons of the geographically close Tamangic and Tibetan languages from north-central Nepal. Both are closely related subgroups within Tibeto-Burman, and both contain members that show evidence of contact. In the following figures we see the results of a comparison of cognacy across a 239-item wordlist.

In Figure 7 we see a network combining Tamangic and Tibetan, showing sixteen languages compared across all items in the wordlist. The lexical comparison clearly divides the languages into a Tamangic and a Tibetan group (Tamangic on the left). Within Tamangic there is a three-way division into Chantyal vs. the rest, strongly differentiated, and then a weaker split between Tamang and Gurung, the two large languages of the central hills, and Manange, Seke and Thakali, smaller languages of Himalayan valleys leading north towards Tibet (unfortunately insufficient lexical data is available for Nar-Phu, a Tamangic language of the Tibetan plateau which shows strong evidence of contact-induced change as a result of its existence on the edge of the Tibetan linguistic area).[4]

[4] The wordlist used contains 239 items: **Body parts**: body, head, hair, face, eye, ear, nose, mouth, tooth, tongue, breast, belly, arm, elbow, palm, finger, fingernail, leg, skin, bone, heart, blood, urine, faeces, knee, neck, liver. **Human relations**: name, man, woman, child, father, mother, older brother, younger brother, older sister, younger sister, son, daughter, husband, wife, boy, girl, person. **Pronouns**: I, you (informal), you (formal), he, she, we (incl.), we (excl.), you (pl.), they. **Animals**: fish, chicken, egg, cow, buffalo, milk, goat, horn, tail, dog, snake, monkey, mosquito, ant, spider, bird, louse, feather, yak (male), yak (female), fly (n.), horse. **Plants and food**: fruit, mango, banana, wheat, millet, rice, potato, eggplant, peanut, chilli, turmeric, garlic, onion, cauliflower, tomato, cabbage, oil, salt, meat, fat, seed, bark, barley flour, butter (yak). **Natural world**: sun, moon, sky, star, rain, water, river, cloud, lightning, rainbow, wind, stone, sand, mud, dust, tree, leaf, root, thorn, flower, earth, mountain, mountain pass, snow. **Tools and buildings**: village, house, roof, door, firewood, broom, mortar, pestle, hammer, knife, axe, rope, thread, needle, cloth, ring, path, fire, smoke, ash, gold, carpet. **Location**: above, below, this, that, these, those. **Numerals**: one, two, three, four, five, six, seven, eight, nine, ten, eleven, twelve, twenty, hundred, few, many, all. **Adjectives**: white, black, red, green, yellow, blue / turquoise, old, new, good, bad, wet, dry, long, short, hot,

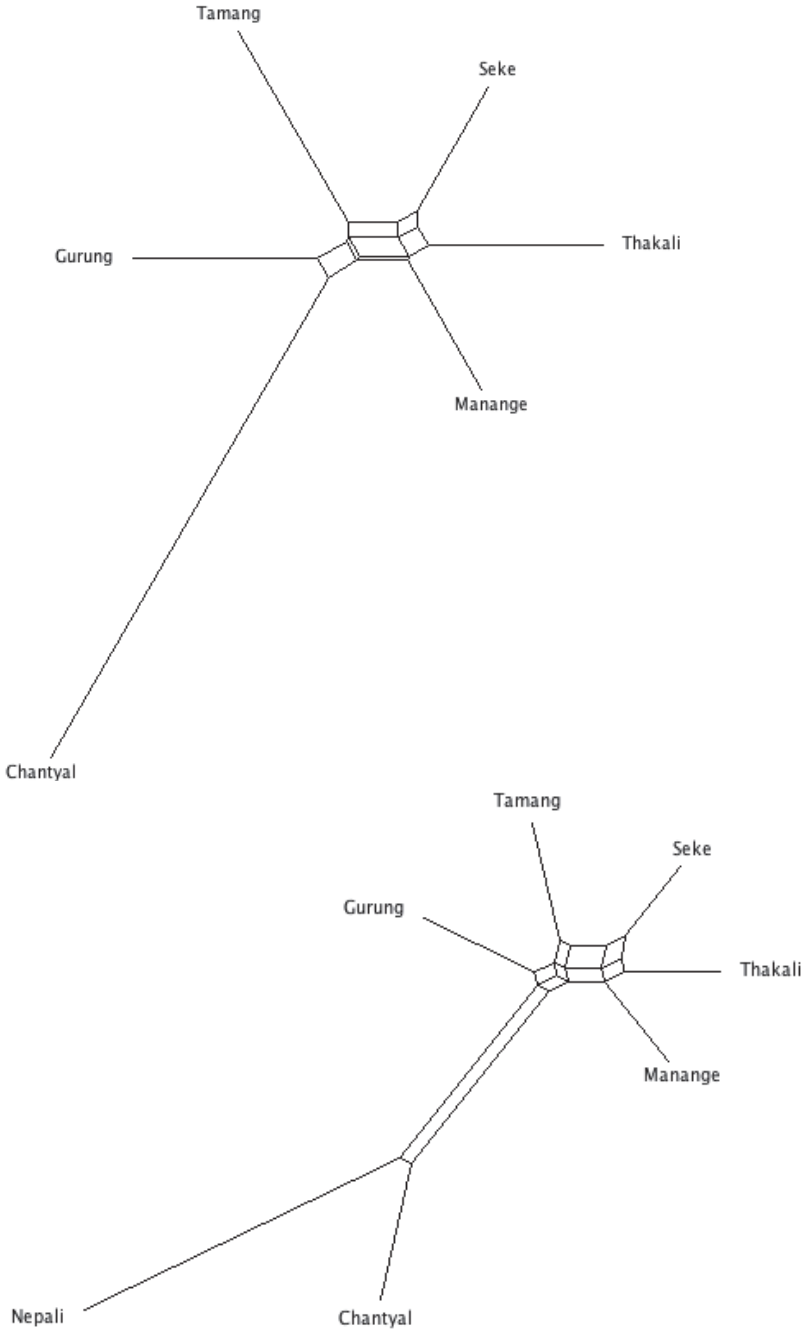
Figure 7. Network analysis of lexical similarities between Tamangic and Tibetan.



To better understand the divisions in the Tamangic group, we can examine the languages from Figure 7 in separate groups. Figure 8 shows the network for just the six Tamangic languages considered here. This representation largely confirms the comments made based on Figure 7, though there is not really any clear evidence that lexically Tamang and Gurung form a cluster. The position of Chantyal is made clear when we examine the second network in Figure 8, in which Nepali has been added. It is now clear that the lexical aberrancy of Chantyal with respect to the rest of its Tamangic relatives is due to extensive lexical contact with Nepali – exactly the position described in numerous papers by Noonan (e.g., 2003, 2008). The difference between the hill languages and the mountain valley languages might be attributed to contact with Nepali as well, though this is not as apparent as it is for Chantyal.

cold, right, left, near, far, big, small, heavy, light, same, different, whole, broken, full, round. **Verbs**: eat, bite, hungry, drink, thirsty, sleep, lie down, sit, give, burn (intr.), die, kill, fly (v.), walk, run, go, come, speak, hear, see, not, know, swim, stand. Interrogatives: who, what, where, when, how many, what kind. **Time**: day, night, morning, noon, evening, yesterday, today, tomorrow, week, month, year.

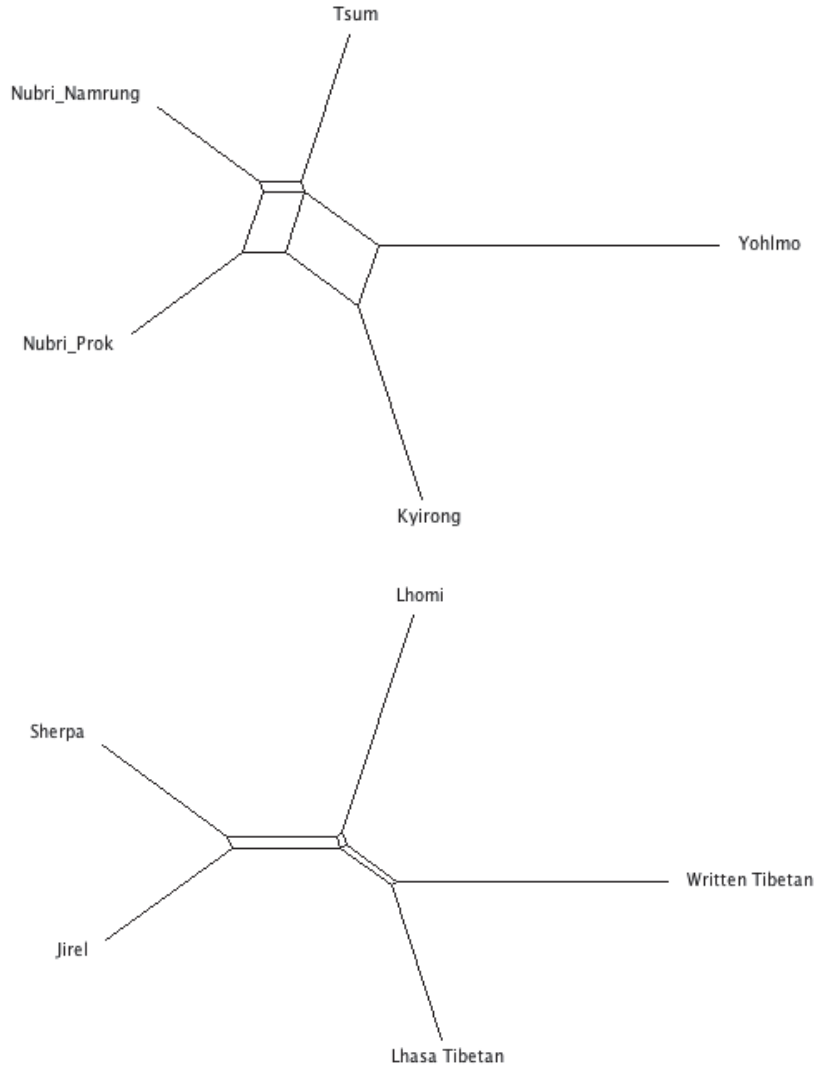
Figure 8. Network analysis of lexical similarities amongst the Tamangic languages (without and with Nepali).



In Figure 7 the Tibetan languages show a very approximate division into the eastern languages (top of the Tibetan cluster) and the western ones (bottom). These two geographically-based groups are shown separately in Figure 9. In the case of the western

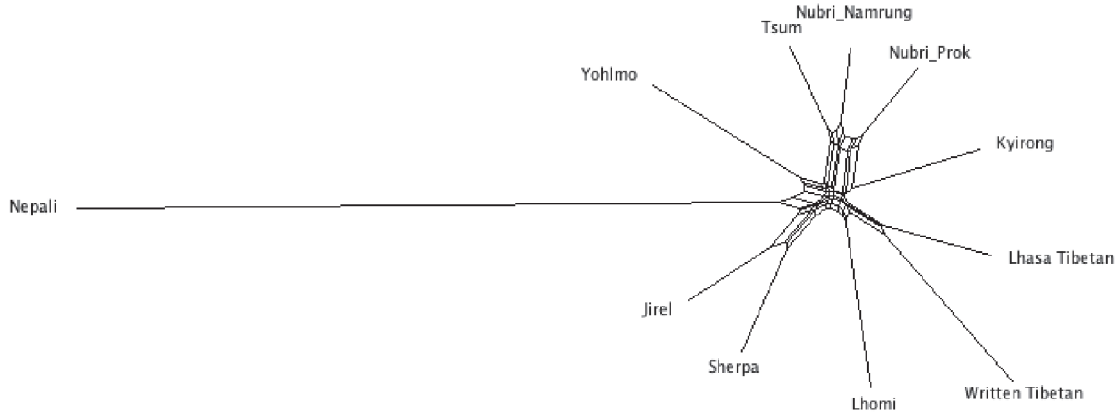
languages Yohlmo is the aberrant language; for the eastern languages Sherpa and Jirel emerge as the most divergent.

Figure 9. Network analysis of lexical similarities amongst the Tibetan languages (western above, eastern below).



As with the Tamangic languages, adding Nepali, the national language, to the sample is revealing, though less so thatn for the Tamangic languages. What is important is that the same languages that were identified on subgroup-internal grounds as being aberrant, Yohlmo, Sherpa and Jirel, are the languages that show the greatest convergence with Nepali, just as in Figure 10 they also showed the greatest convergence with Tamangic.

Figure 10. Network analysis of lexical similarities amongst the Tibetan languages.



These comparisons have all involved comparing whole wordlists. In Figures 11-12 we can compare the different clusters that arise from a comparison of different semantic domains. Compare Figure 11 with Figure 7; when we restrict the comparison to body parts there is no evidence of convergence between Yohlmo and the southern languages (Tamangic and Nepali); if any Tibetan language shows evidence of convergence in this semantic domain, it is Tsum, which in Figure 7 forms an overall cluster with the two (geographically close) Nubri varieties included in the sample. Among the Tamangic languages, all but Chantyal form a tight cluster, and Chantyal shows an unambiguous borrowing relationship with Nepali (indeed, as can be seen in all of the Figures here, from a lexical perspective there are few domains in which Chantyal appears as a Tibeto-Burman language at all).

Figure 11. Network analysis of lexical similarities in body parts.

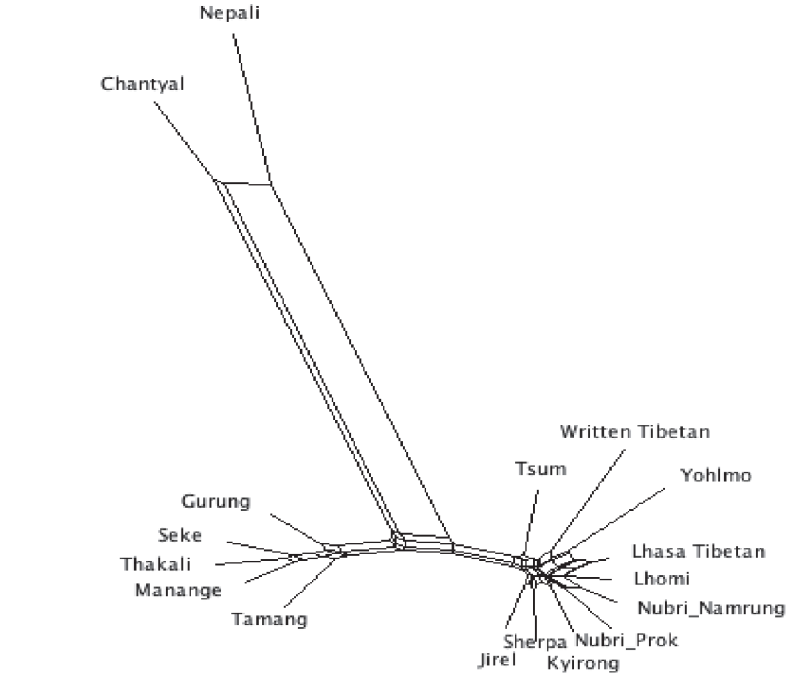
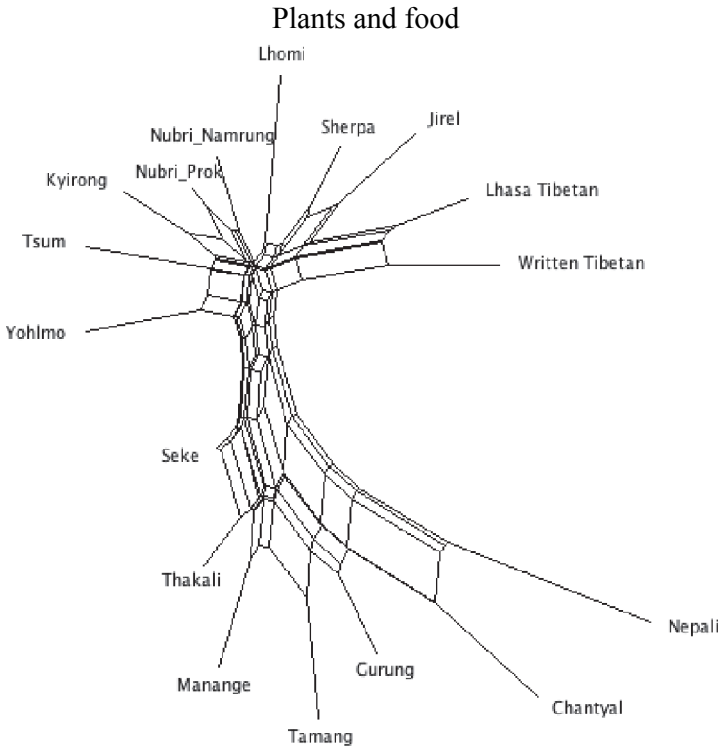


Figure 12 presents two more dendrograms, representing comparisons of two other semantic domains: plants and food (on the left) and tools and buildings (right). In these dendrograms Yohlmo is clearly strongly converging with the southern languages, especially for the 'tools' semantic domain. In the plants and food dendrogram we can see similarly see that the mountain valley Tamangic languages, Seke, Thakali and Manange, are much closer lexically to the Tibetan languages than their hill relatives. That these mountain valleys, above 3000m altitude, share a similar natural and plant ecology with the languages of the Tibetan plateau (~4000m altitude) makes it unsurprising that the lexicon for food that can be grown in their environments is more similar than it is with the southern languages (Tamang, Gurung, Chantyal and Nepali).

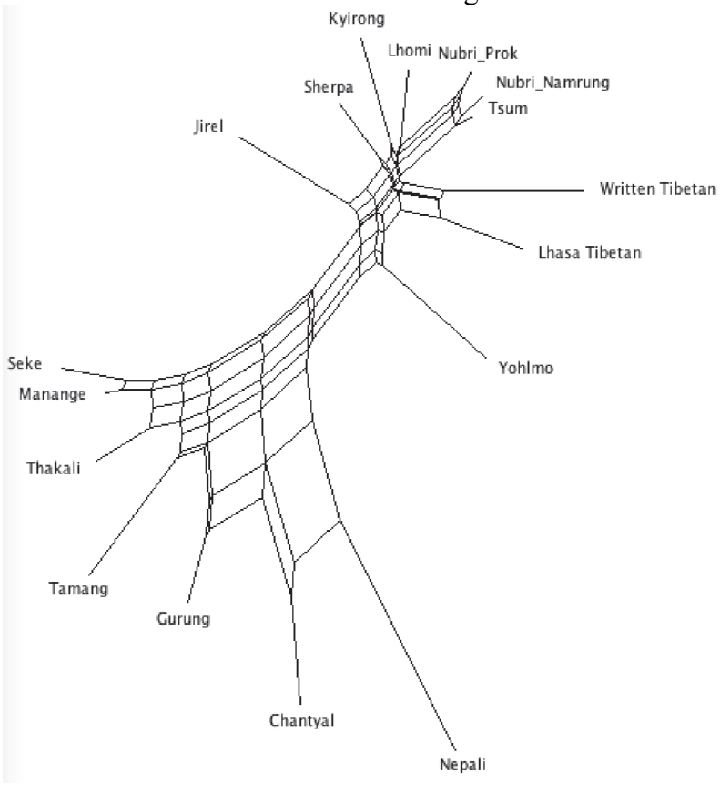
In terms of tools and buildings, none of the Tamangic languages show evidence of any convergence with the Tibetan languages, though the more southern and western languages to be more lexically similar with Nepali. Among the Tibetan languages Yohlmo is again closest to the Tamangic languages. Given the known history of Yohlmo in the Helambu valley (Clarke 1980a, 1980b), it is very likely that the patterns of convergence between Yohlmo, a close relative of (Lende) Kyirong, reflect contact with the pre-Tibetan language(s) of the Helambu area.

While the divergence of Yohlmo from a more general Tibetan profile for tools and building could be attributed either to convergence with Tamangic, or the presence of a common factor, Nepali influence, in both Tamangic and Yohlmo, the evidence from the plants and food dendrogram is less equivocal, with Yohlmo converging away from the other Kyirong Tibetan varieties, and towards Tamangic in preference to Nepali.

Figure 12. Network analysis of lexical similarities amongst the Tibetan languages: selected semantic domains



Tools and buildings



The examination of lexical data has again shown us that, even without an out-group to compare to, it is possible to develop realistic suspicious about contact histories in different languages. In Figure 7 it was clear even without the addition of the out-group Nepali that Chantyal had been heavily contact-affected in terms of its lexicon. In Figure 10 we can see that the effect of Nepali on the Tibetan languages has been much less than its effect on the Tamangic languages.

The discussion of subdomains of the lexicon is particularly interesting in light of the way the languages cluster when typological features, such as were examined in 4.1 – 4.2, are investigated. In Figures 13 and 14 we can see the clustering obtained in two subdomains of the phonology: all oppositions to do with consonants, and all oppositions not related to consonants (the full set of features examined is reported in Donohue et al. 2013). When the consonantal phonologies are examined we see three broad clusters: at the top of the diagram, the eastern Tibetan languages (plus Yohlmo) together with Thakali and Seke, the two Tamangic languages in this study that have been most influenced by Tibetan varieties (notably Mustang). The bottom right contains the three Tamangic languages spoken in the hills, and not in Himalayan valleys (plus, loosely, Nepali). The bottom left contains the western Tibetan varieties, plus Written Tibetan, plus Manange, the Tamangic language in most contact with these conservative varieties. Speaking of 'the phonology' is clearly not suitable, since the consonantal material tells a different story from that of the non-consonantal material. Examining the vowels, prosody and phonotactic conditions on the languages shows a very different picture, one in which there are two Tibetan poles (a conservative one on the

left, and an innovative one on the right), with Jirel, Sherpa and Yohlmo assimilating to the Tamangic core in the middle of the network.

We should note that, unlike the analysis of lexical items, the analysis of phonological oppositions offers nothing to suggest a close relationship between Chantyal and Nepali.

Figure 13. Network analysis of consonantal similarities.

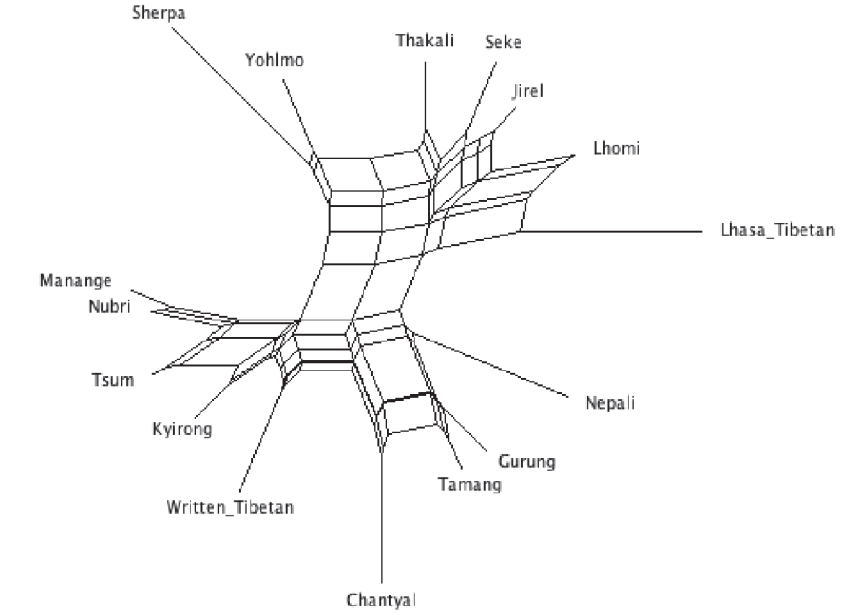
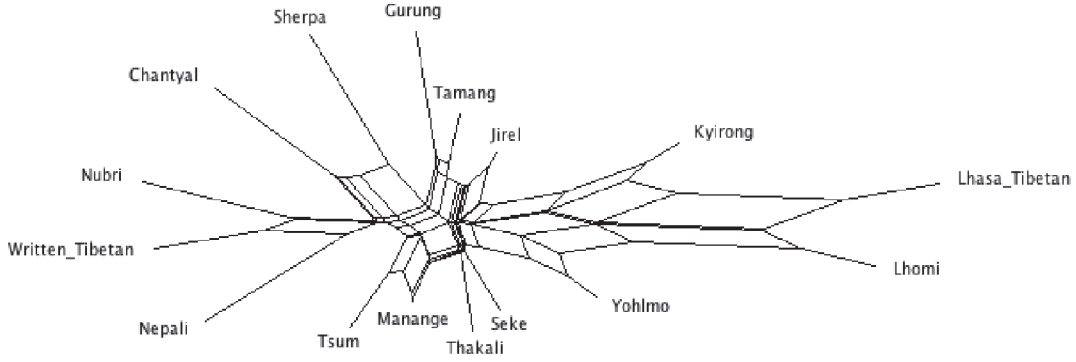
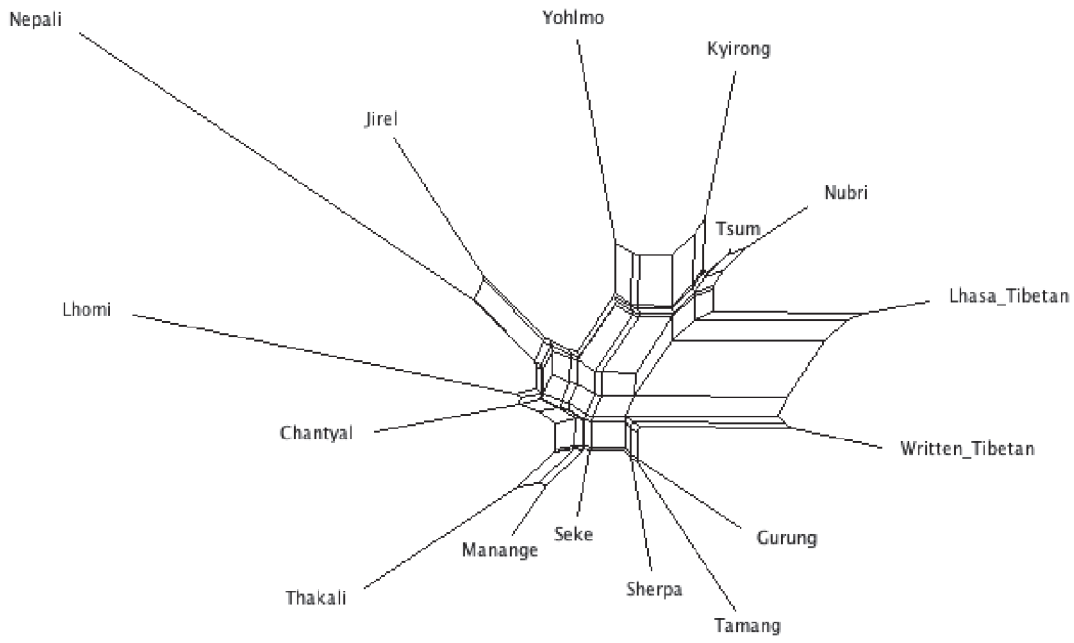


Figure 14. Network analysis of non-consonantal similarities.



When we examine broad morphosyntactic traits we can see that Sherpa is the Tibetan language most assimilated to the Tamangic profile, and Jirel is most affected by Nepali morphosyntactic patterns. Yohlmo is firmly embedded in a cluster with the other Kyirong-area Tibetan languages, with no evidence of assimilation to the southern languages.

Figure 15. Network analysis of morphosyntactic similarities.



Examining the Tibetan and Tamangic languages and their contact situation has shown that not only is the phonology different from the morphosyntax in terms of the kinds of clusters that emerge, but that sub-parts of the phonology have also been shown to reflect different histories. We have also seen that different kinds of linguistic data reveal different aspects of history: the consonantal system of Chantyal is firmly Tamangic, while the lexicon is clearly more strongly related to Nepali. For Yohlmo we can see a lexical pattern of assimilation to Tamangic, combined with a very conservative morphosyntax and consonant system, but a system of vowels and prosody that has been influenced by the Tamangic languages.

4.4 Exploring in Island Southeast Asia

The spread of Austronesian languages across a large portion of the world's surface has attracted much research in and across the disciplines of linguistics, archaeology, genetics and history. In this section I focus on the dispersal of Austronesian languages out of Taiwan and across Melanesia and Island Southeast Asia (ISEA).

Thanks to extensive records we have at least basic lexical data for a very large number of Austronesian languages, allowing great advances in our understanding of the linguistic history of these languages. While there is controversy about many of the proposed subgroups, two subgrouping facts are clear and uncontroversial: the Austronesian languages spoken outside Taiwan all form a single subgroup, Malayo-Polynesian, and the Austronesian languages of eastern Melanesia and Oceania form a single subgroup of Malayo-Polynesian, Oceanic.[5] Proto-Austronesian was spoken on the island of Taiwan, where all of the first-order subgroups of Austronesian are represented (e.g., Blust 2009). The Proto-Austronesian culture

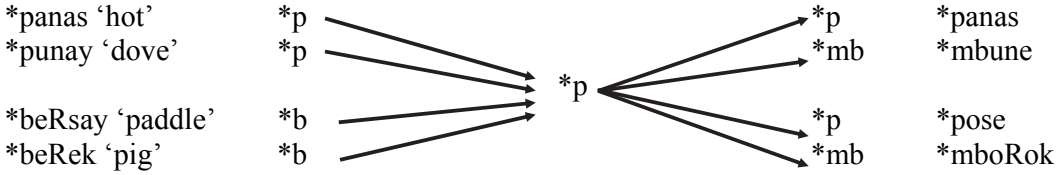
[5] Two languages of western Micronesia, Chamorro and Palauan, are exceptions to this, being Malayo-Polynesian but not Oceanic.

has been unambiguously linked to rice agriculture, other domesticates, a range of technological skills, and various cultural practices (Blust 2009, Pawley 2007). Proto-Oceanic reconstructions reveal some continuities compared to Proto-Austronesian and Proto-Malayo-Polynesian, but also many innovatory technologies (Pawley 2007). The location and cultural leanings that can be associated with Proto-Oceanic have been the subject of debate, and recent work on the heterogeneity of archaeological sites identified as being associated with Lapita culture, frequently claimed to have been tied to Proto-Oceanic, has emphasized the controversy (Donohue and Denham 2008, 2012, Specht et al. 2013). The questions we face, with respect to Proto-Oceanic, are summarized in (4).

(3) a. What is the relationship of Proto-Oceanic to the other Austronesian languages?
b. What is the relationship of Proto-Oceanic to the non-Austronesian languages of Melanesia?

Without disputing the Austronesian lexicon and sound correspondences with Proto-Malayo-Polynesian, we should point out that some of the defining criteria for Proto-Oceanic include the *irregular* correspondences found for the bilabial and velar stops. Figure 16 illustrates the correspondences that hold for four bilabial-initial nouns. In Proto-Malayo-Polynesian the words are distinguished by a voiced:voiceless opposition; in pre-Proto-Oceanic this opposition collapsed, and by the time of the Proto-Oceanic break-up the relevant parts of the lexicon had split into a voiceless vs. prenasalized opposition, without conditioning environment. Further, we should note that prenasalized stops are not an expected part of the phonology of the languages of Asia, including most of ISEA. They are, however, a feature of the languages of Melanesia, including north-east New Guinea (Donohue and Whiting 2011). (Another feature found in the reconstructed phonological inventory of Proto-Oceanic that is common in Melanesia, but not in ISEA, is the presence of rounded stops.)

Figure 16. Developments of labial stops between Proto-Malayo-Polynesian and Proto-Oceanic



These commonalities suggest that Proto-Oceanic includes substratal elements that have a provenance in the languages of eastern Melanesia.

Thanks to recent work describing the morphology and syntax of the languages we can meaningfully compare large amounts of data from a large sample of languages (e.g., Donohue 2007). Figure 17 shows a clustering analysis of morphosyntactic traits in Austronesian languages of ISEA, plus representative neighbors. Included are reconstructions of Proto-Austronesian and of Proto-Oceanic, as indicated. The regions marked on the dendrogram have been added based on the criteria in (4).

(4) a. Areas A and B are that part of the diagram that contains only Austronesian languages.
b. Area B is those languages most close, typologically, to Proto-Austronesian.

c. The two languages closest to Proto-Austronesian are Atayal and Paiwan, spoken on Taiwan.
b. Area D is that part of the diagram that contains only non-Austronesian languages.
e. Area E is that part of the diagram that contains the languages of mainland Southeast Asia.
f. Area C is the part of the diagram with a mixture of Austronesian and non-Austronesian languages.

There is certainly some continuity to the Austronesian languages, meaning that there is a block of languages that do not cluster with either the mainland Southeast Asian languages nor with the Melanesian languages. There is similarly a typological pole consisting of the (interior, highland) languages of New Guinea, with a few off-shore exceptions. Importantly there is a large typological region in which both Austronesian and non-Austronesian languages are found, intermingled, demonstrating that there is no strong genealogical reliability to the notion of typological traits. The languages in that part of the diagram labeled E could all be said to be leaning towards pidgin/creole-like structure.

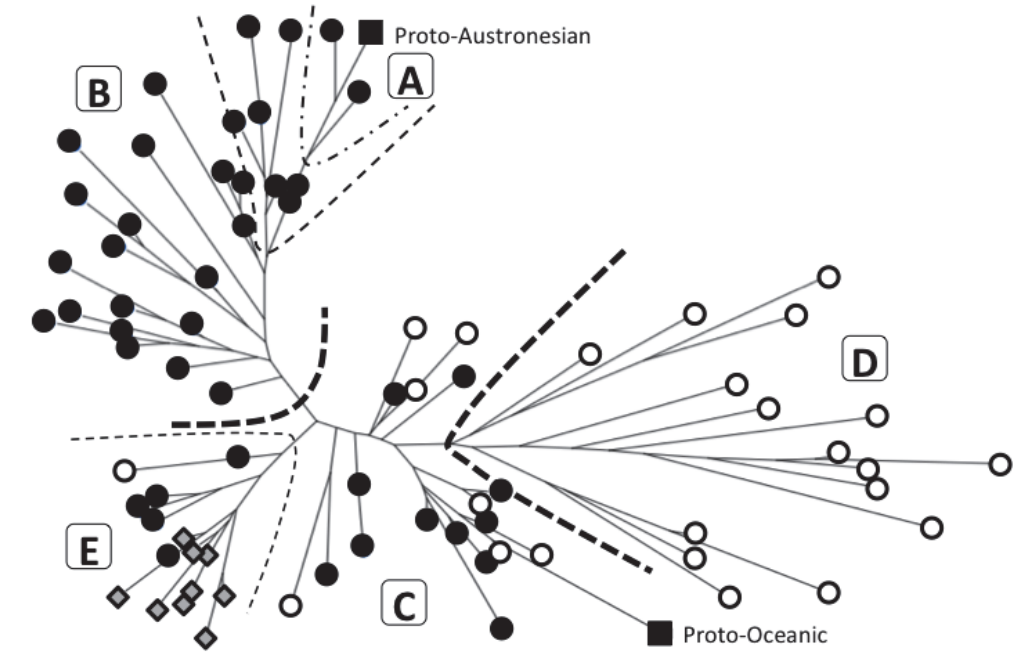
We can now examine the spatial distribution of the languages, with Map 1. There are two important point to note when examining the data in Figure 17 and Map 1.

- while cluster C and cluster E languages in Figure 17 are genealogically diverse, they can be modeled geographically without trouble:
    - cluster C is the buffer between B and D (remembering that Austronesian languages travelled along the north coast of New Guinea)
    - cluster E is found on mainland Southeast Asia, and as a buffer along the contact zone between clusters B, C and D
- while Proto-Austronesian *is* firmly embedded in a typological cluster of genealogically-related languages, Proto-Oceanic cannot be so defined.

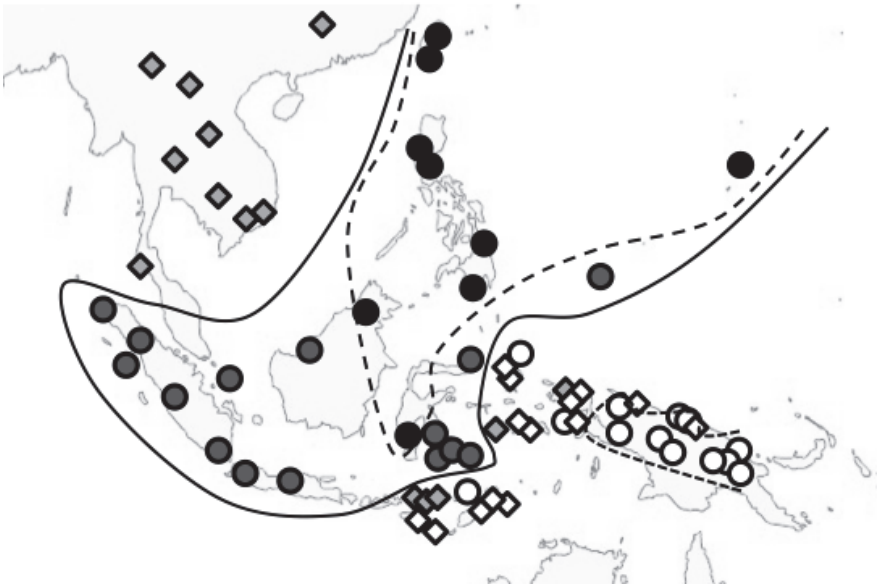
The significance of the position of Proto-Oceanic in the diagram is that it is clear that, from an Austronesian perspective, Proto-Oceanic appears to be drastically contact-affected. Given the evidence of a Melanesian 'accent' in Proto-Oceanic (witnessed by the development of prenasalized stop contrasts, discussed above), it seems (to preempt some of the discussion in section 5) that the demographic component of the arriving Austronesian culture must have been minimal. The morphosyntactic typology of Proto-Oceanic is not obviously recognizable as Austronesian, as defined in Figure 17, and the phonology fits better in north-east Melanesia than it does in the northern ISEA. (See the appendix for a guide to which language is where in the figure and map.)

This implies that even prior to the break-up of Proto-Oceanic, Proto-Oceanic was already strongly contact-affected. Elsewhere (Donohue and Denham in press) it has been suggested that we could more parsimoniously think of many of the Austronesian languages of eastern ISEA and Melanesia as being non-Austronesian languages of the region that have been partly relexified by contact with Austronesians. The level of morphological, phonological and lexical material shared between Proto-Malayo-Polynesian and Proto-Oceanic makes that an unlikely scenario for Proto-Oceanic, but a scenario in which Proto-Oceanic was the result of layers of language contact and language shift, conventionalized over a long time period, appears likely.

Figure 17. Clustering of 78 languages based on morphosyntactic features, focusing on Austronesian languages spoken west of New Guinea.



Map 1. The different typological language clusters from Figure 17 mapped out



Key: Black circle = Area A, Grey circle = Area B, White diamond = Area C, White circle = Area D, Grey diamond = Area E.

5 Conclusions

Contact can be detected in any area of a language, from simple lexical borrowings to more subtle patterns in the phonology and morphosyntax.

Even without detailed philological information we can detect contact scenarios, and even generate broad-outline social sketches that can direct the linguistic ecologies that will benefit from more detailed work. Table 6 (from Donohue 2013) presents possible demographic scenarios in broad outline, and the linguistic traces they might leave.

In this paper we have seen examples of many of these outcomes. The Romance data discussed in 4.1 shows Spanish acquiring elements of non-Indo-European morphosyntax as a result of the intense contact it has undergone with Arabic and Basque. With Romanian we see such heavy contact effects that it is hard to classify the language as Romance other than through the lexicon: the typology is convincingly Balkan and Slavic.

The Dravidian data similarly shows assimilation of the northern languages to non-Dravidian norms in their areas, again showing that we can detect a scenario in which a small but dominant outside group influences the language of the original inhabitants.

In 4.2 we managed to detect the Balkan *Sprachbund* without any philological data, comparing only general typological traits. In 4.3 we saw that using specialized sub-domains of linguistic data, rather than collapsing all the data from a particular domain, is revealing of kinds of contact scenarios that go beyond the broad typology shown in Table 6. And in 4.4 we examined the typology of the Austronesian languages of ISEA with the conclusion that contact, in the form of language shift, must have played a strong and early role in the formation of Austronesian.

Table 6. Different superimposition scenarios (Donohue 2013)

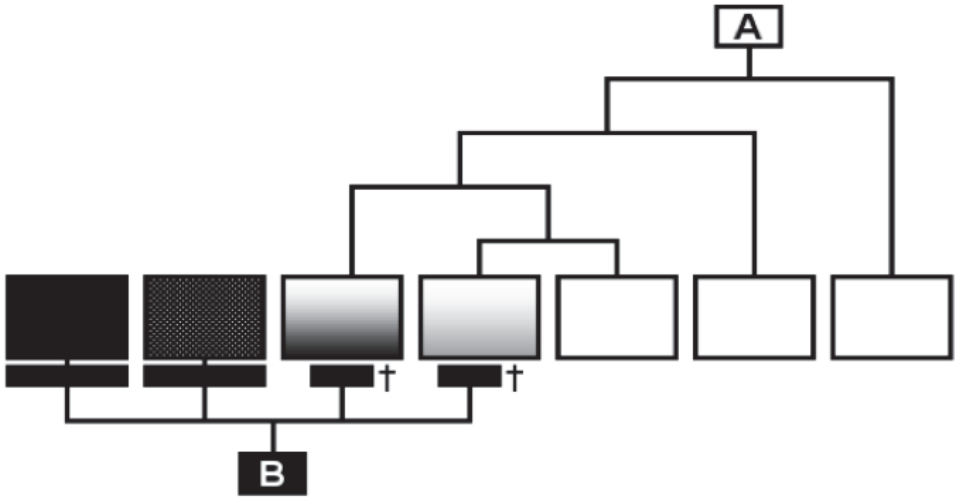
	population	
socio-politics	intruder populous	intruder minority
intruder dominant	1. **intruder L largely unchanged**	3. intruder L acquires **phonology substrate** from local L
	2. *local languages lost*	4a. *local languages lost* *or* 4b. local L acquires **morphosyntax overlay** from intruder L
intruder subordinate	5. intruder L acquires **morphosyntax overlay** from local L	7a. *intruder languages lost* *or* 7b. intruder L acquires **morphosyntax overlay** from local L
	6. local L acquires **phonology substrate** from intruder L	8. **local L largely unchanged**

Importantly, the methodology for detecting contact has been calibrated against scenarios that are well-understood. With a well-annotated database (eg., Haspelmath et al. 2005, Donohue et

al. 2013), and a simple clustering algorithm, we can detect contact even in areas where it was not previously suspected. Rather, we can detect suspicions of contact; any such suspicions would require a detailed examination of the languages concerned, such as was carried out for the Nepalese situation in 4.3.

When we examine data on a large scale it is hard to escape the conclusion that language shift is more common in the spread of language families than is generally discussed in the historical linguistics literature (see, e.g., Donohue and Denham 2011). This can be detected, as shown schematically in Figure 18, by detecting unexpected divergence in the typological profile of members of a language family.

Figure 18. Traces of earlier language ecologies survive in typological profiles



We might wonder what exactly we are detecting. Do the different subsets of features reflect different histories (inheritances, contact events)? Do the different clusterings tell us something about the stability of the different sets of features?

As a cautionary conclusion, we can finish with the observations that investigating typological features can be interpreted in ways consistent with known history. Equally, the method offers different results for different features, and is not a proxy for researching family relations (*à la* the comparative method), but it gives interesting insights into the language speakers' linguistic history. Wichmann and Saunders (2007) discuss the kinds of historical information that can be gleaned from typological analysis, and what information cannot be inferred; this paper extends points made by Wichmann and Saunders, both in detail and in application to different scenarios.

Importantly, in order to refine our heuristics that guide us to social scenarios we require more case studies that examine not only broad typological relationships, as in 4.1 and 4.2, but which go into details in multiple areas of analysis. As with most research projects, the interesting results emerge only when we are able to examine linguistic data not as monolithic 'black boxes' that can only yield a single 'sound bite' outcome, but rather treat each logically separate module of data separately, extending Oppenheimer's (2004) caution on the perils of too quickly combining interdisciplinary data to analysis that uses data from only one discipline.

References

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

Clarke, Graham. 1980a. Helambu History. *Journal of the Nepal Research Centre* 4: 1-38.

Clarke, Graham. 1980b. Lama and Tamang in Yolmo. In Michael Aris and Aung San Suu Kyi, eds., *Tibetan Studies in Honour of Hugh Richardson*: 79-88. Proceedings of the International Seminar on Tibetan Studies, Oxford, 1979. Warminster: Aris & Phillips Ltd.

Donohue, Mark. 2007. Word order in Austronesian: from north to south and west to east. *Linguistic Typology* 11 (2): 351-393.

Donohue, Mark. 2012. Typology and Areality. *Language Dynamics and Change* 2: 98-116.

Donohue, Mark. 2013. Who inherits what, when? contact, substrates and superimposition zones. In Balthasar Bickel, Lenore A. Grenoble David A. Peterson, and Alan Timberlake, eds., *Language Typology and Historical Contingency*: 219-240. Typological Studies in Language 104. Amsterdam: John Benjamins.

Donohue, Mark. 2013. Towards a Papuan history of languages. *Language and Linguistics in Melanesia* 31 (1): 24-41.
Available online at http://www.langlxmelanesia.com/issues.htm.

Donohue, Mark, and Tim Denham. 2008. The Language of Lapita: Vanuatu and an early Papuan Presence in the Pacific. *Oceanic Linguistics* 47 (2): 365-376.

Donohue, Mark, and Tim Denham. 2011. Language and genes attest different histories in Island Southeast Asia. *Oceanic Linguistics* 50 (2): 536-542.

Donohue, Mark, and Tim Denham. 2012. Lapita and Proto-Oceanic: thinking outside the pot. *Journal of Pacific History* 47 (4): 443-457.

Donohue, Mark, and Tim Denham. In press. Becoming Austronesian: mechanisms of language dispersal across Indo-Malaysia. In David Gil and John McWhorter, eds., *Austronesian undressed*. Canberra: Pacific Linguistics.

Donohue, Mark, Tim Denham and Stephen Oppenheimer. 2012. Uncoupling inheritance and diffusion: a lexical-based methodology detects social distance. *Diachronica* 29 (4): 502-522.

Donohue, Mark, Rebecca Hetherington, James McElvenny and Virginia Dawson. 2013. World phonotactics database. Department of Linguistics, The Australian National University. http://phonotactics.anu.edu.au. Accessed (Accessed 01 December 2013).

Donohue, Mark, Simon Musgrave, Bronwen Whiting and Søren Wichmann. 2011. Typological feature analysis models linguistic geography. *Language* 87 (2): 369-383.

Donohue, Mark, and Bronwen Whiting. 2011. Quantifying areality: a study of prenasalisation in Southeast Asia and New Guinea. *Linguistic Typology* 15: 101-121.

Donohue, Mark, Søren Wichmann and Mihai Albu. 2008. Typology, areality and diffusion. *Oceanic Linguistics* 47 (1): 223-232.

Driem, George van. 2001. *Languages of the Himalayas: An Ethnolinguistic Handbook of the Greater Himalayan Region*. Vols 1 and 2. Leiden: Brill.

Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie. 2005. *World Atlas of Language Structures.* Oxford: Oxford University Press.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, and E. Anthon Eff. 2014. Inheritance and diffusion of language and culture: A comparative perspective**.** *Social Evolution & History* 14.2

Huson, Daniel H. and David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23: 254–267. Software available from http://www.splitstree.org.

Nerbonne, John. 2009. Data-Driven Dialectology. *Language and Linguistics Compass* 3 (1): 175-198. DOI: 10.1111/j.1749-818x2008.00114.x

Nichols, Johanna and Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2: 760–820.

Noonan, Michael. 2003. Recent language contact in the Nepal Himalaya. In David Bradley, Randy LaPolla, Boyd Michailovsky and Graham Thurgood, eds., *Langauge variation: papers on variation and change in the Sinosphere and in the Indosphere in honour of James A. Matisoff*: 65-88. Canberra: Pacific Linguistics.

Noonan, Michael. 2008. Contact-induced change: the case of the Tamangic languages. In Peter Siemund and Noemi Kintana, eds., *Language contact and Contact languages*: 81-106. Amsterdam: John Benjamins.

Noonan. Michael. 2010. Genetic Classification and Language Contact. In Raymond Hickey, ed., *The Handbook of Language contact*, 48-67. Wiley Blackwell.

Oppenheimer, Stephen. 2004. The 'Express Train from Taiwan to Polynesia': on the congruence of proxy lines of evidence. *World Archaeology* 36 (4): 591-600

Pawley, Andrew K. 2007. The Origins of Early Lapita Culture: The testimony of historical linguistics. *Oceanic Explorations: Lapita and Western Pacific settlement* ed. by Stuart Bedford, Christophe Sand & Sean P. Connaughton, 17–49. Canberra: Australian National University.

Ross, Malcolm. 2003. Diagnosing prehistoric language contact. In R. Hickey, ed., *Motives for language change*: 174-198. Cambridge: Cambridge University Press.

Specht, Jim, Tim Denham, James Goff and John Edward Terrell. 2013. Deconstructing the Lapita Cultural Complex in the Bismarck Archipelago. *Journal of Archaeological Research*.

Thomason, Sally. 2001. *Language contact. An introduction*. Edinburgh: Edinburgh University Press.

Thomason, Sarah Grey. 2009. How to establish substratum interference. In Yasuhiko Nagano, ed., *Issues in Tibeto-Burman historical linguistics*: 319-328. Osaka: Senri Ethnological Studies 75.

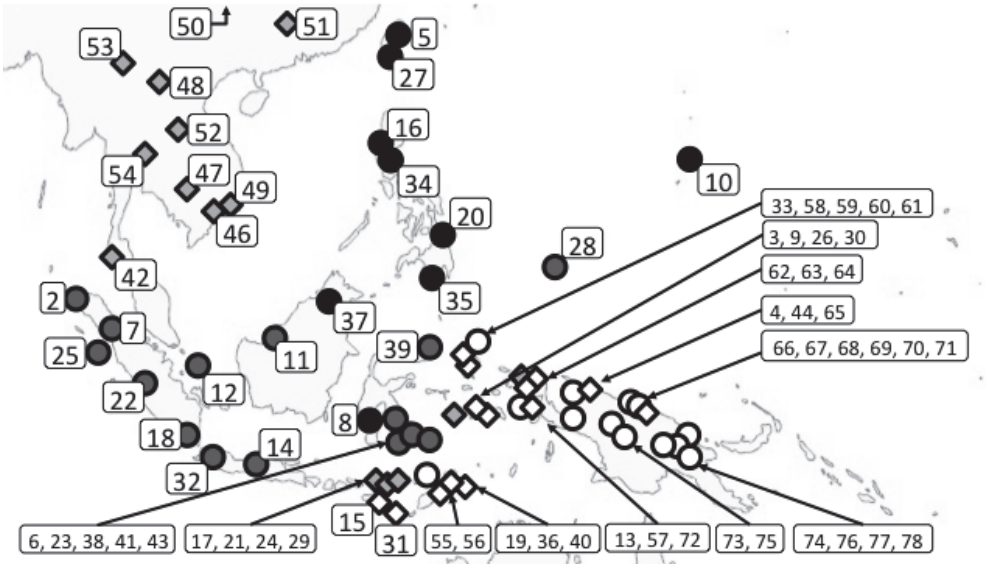
Thomason, Sarah Grey, and Terrence Kaufman. 1988. *Language contact, Creolization, and Genetic Linguistics*. Los Angeles: University of California Press.

Weinreich, Uriel. 1963. *Languages in Contact: findings and problems*. The Hague: Mouton.

Wichmann, Søren and Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. *Diachronica* 24: 373–404.

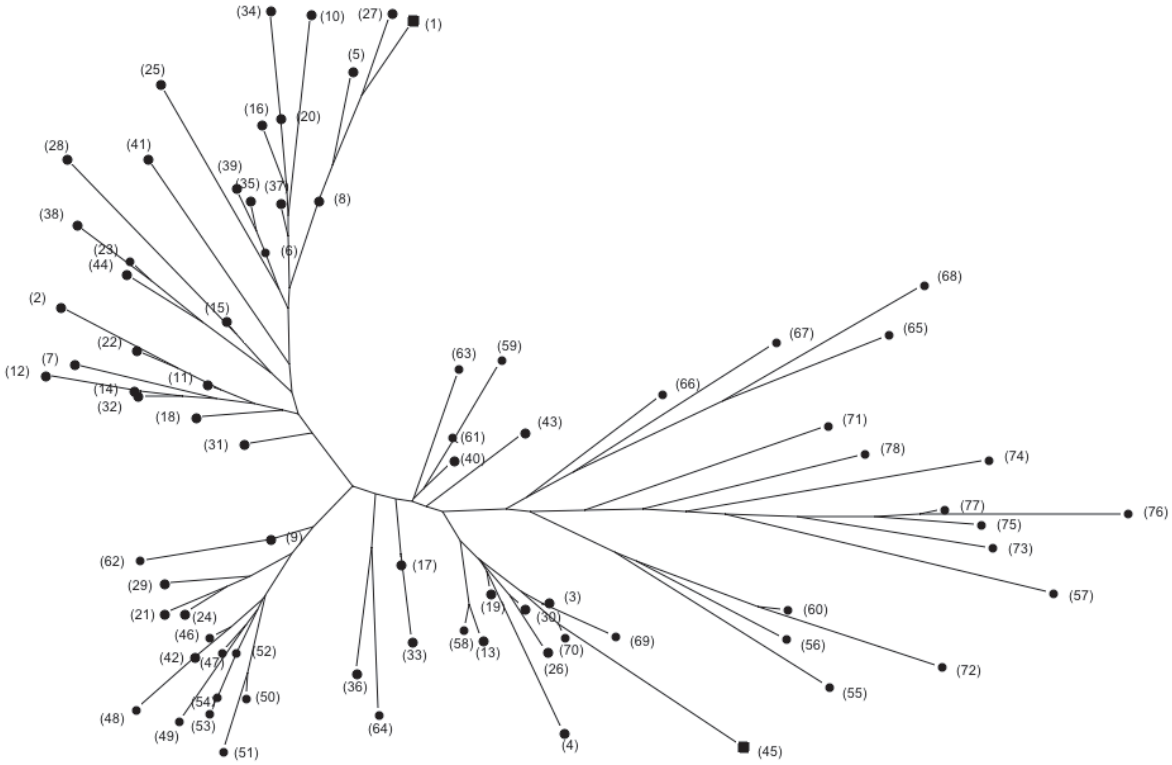
Appendix: Identities and Locations of Languages Discussed in 4.4.

Map 2. The different (modern) languages coded by region in Figure 17 and Map 1.



(Black circle = 'pure' Austronesian region, Grey circle = 'pure' Austronesian, but more distant from Taiwan. White circle = 'New Guinea' region; Grey diamond = mainland Southeast Asia region. White diamond = 'middle ground'.

Figure 19. Key to languages represented in Figure 17.



- Austronesian: (1) Proto-Austronesian; (2) Acehnese; (3) Alune; (4) Ambai; (5) Atayal; (6) Bajau; (7) Batak; (8) Bugis; (9) Buru; (10) Chamorro; (11) Iban; (12) Indonesian; (13) Irarutu; (14) Javanese; (15) Kambera; (16) Kapampangan; (17) Lamaholot; (18) Lampung; (19) Leti; (20) Mamanwa; (21) Manggarai; (22) Minangkabau; (23) Muna; (24) Ngada; (25) Nias; (26) Nuaulu; (27) Paiwan; (28) Palauan; (29) Palue; (30) Paulohi; (31) Sawu; (32) Sundanese; (33) Taba; (34) Tagalog; (35) Tboli; (36) Tetun; (37) Timugon; (38) Tolaki; (39) Tondano; (40) Tugun; (41) Tukang Besi; (42) Urak Lawoi; (43) Warembori; (44) Wolio; (45) Proto-Oceanic;
- Austroasiatic: (46) Chrau; (47) Khmer; (48) Khmu; (49) Vietnamese;
- Hmong-Mien: (50) Hmong Njua; (51) Mien;
- Kradai: (52) Lao; (53) Shan; (54) Thai;
- 'Papuan': (55) Kolana; (56) Tanglapui; (57) Iha; (58) Sahu; (59) Tidore; (60) Tobelo; (61) West Makian; (62) Abun; (63) Hatam; (64) Maybrat; (65) Yawa; (66) Dumo; (67) Isaka; (68) Sko; (69) Au; (70) Olo; (71) One; (72) Ekari; (73) Dani; (74) Kewa; (75) Una; (76) Fore; (77) Tauya; (78) Amele.

Mark Donohue
The Australian National University

mark.donohue@anu.edu.au