Transactions of the Philological Society Volume 103:2 (2005) 147–170

SWADESH SUBLISTS AND THE BENEFITS OF BORROWING: AN ANDEAN CASE STUDY[1]

By April McMahon[a], Paul Heggarty[a], Robert McMahon[b] and Natalia Slaska[c]
[a]*English Language, PPLS, University of Edinburgh;* [b]*East of Scotland Molecular Genetics Service, Western General Hospital, Edinburgh;* [c]*Department of English Language and Linguistics, University of Sheffield*

Abstract

Although borrowing and contact are recognised as important factors in language histories, there is no clear and agreed way of dealing with their effects on methods like traditional lexicostatistics. We argue that one promising approach involves subdividing standard meaning-lists into more and less conservative sublists. Differences between trees or networks generated from these sublists may then indicate borrowing. This approach has been tested on Indo-European, but here we apply it to languages of the Andes in an attempt to answer the vexed Quechumara question: are Quechua and Aymara genetically related, or linked only by contact? Our evidence suggests that contact is the more likely explanation for parallels between Quechua and Aymara.

1. Borrowing: a curse or a blessing?

Language classification, particularly when it involves the lexicon, suffers from a difficult duality. On the one hand, we represent families and subgroups using trees, which by their nature are designed only to show descent with modification from a single historical source. In addition, methods like lexicostatistics are based

[1]The research reported here was funded by the AHRB (grant AN6720/APN12536), and we gratefully acknowledge their support. We also thank two reviewers for their helpful and constructive comments. Additional supplementary material can be found at http://www.philsoc.org.uk/transactions.asp

on judgements of whether items compared are plausibly cognate: non-cognates are excluded. On the other hand, however, it is well known that speakers, and therefore languages, borrow from one another; and undiagnosed or misdiagnosed loans can obscure the familial signal and lead to erroneous classifications. But borrowing is nonetheless part of the histories of many languages: some, like pidgins, creoles and mixed languages, would not exist at all without contact situations and their linguistic consequences.

Many leading scholars in historical and comparative linguistics have acknowledged borrowing as a real factor in language histories. Pulgram (1995: 233), for instance, argues that this makes the vocabulary less than ideal for classification, noting that '…words may wander easily, sometimes eagerly, from their original family into any other, they may become naturalized and thereby hide their origins.' Hall (1960: 152) stresses the ubiquity of borrowing, arguing that 'There is not and has not been for thousands of years a 'pure' language, in the sense of one without any borrowings from a foreign language.' Hoenigswald (1990: 12) is also at pains to point out that contact situations '…are in no way exceptional or morbid; on the contrary, they are all-pervasive'; and although he suggests that in principle borrowings must be identified and excluded before the comparative method can be applied, he is also well aware that this is sometimes easier said than done (1960, 1973).

How, then, are we to deal with the fact that borrowings are a real and important aspect of the histories of languages, but that they also present an obstacle to accurate genetic grouping, and are incompatible with the family tree mode of representation? Embleton (1986) has amply demonstrated that simply opting for Swadesh-lists of basic vocabulary (Swadesh 1950, 1952, 1955) is not a solution, since borrowing still occurs in these basic meaning-lists: for the 200-item Swadesh list, she reports 12 borrowings from French into English, 16 from North Germanic into English, and 15 from Dutch to Frisian, for example. Likewise, attempts to exclude borrowings algorithmically (see again Embleton 1986) rely on prior diagnosis of loans; but we need a workable system particularly for cases where we are not sure whether we have loans or not.

In recent and ongoing work (see McMahon & McMahon 2003, 2004, forthcoming, Heggarty forthcoming, in preparation), we have taken a different approach to borrowing. We have used the extensive Dyen, Kruskal & Black (1992) database of 200-meaning Swadesh-lists for Indo-European, but crucially, following work by Lohr (1999), have subdivided this full list into two shorter sublists. These sublists are labelled the hihi list (composed of the most retentive, most reconstructible meanings, which seem maximally resistant to change, including borrowing), and the contrasting lolo list (which contains the least stable meanings in the Swadesh list, including those most amenable to borrowing). When these contrasting sublists are run through computer programs which draw and select trees (Felsenstein 2000), we find identical outputs in terms of branching order for the hihi and lolo sublists in cases where there have been no borrowings (though there may be some differences in branch lengths). When contact has taken place, on the other hand, we observe different trees for the different sublists: for the less conservative meanings, the borrowing language tends to move towards the language or group which is the source of the loans. These effects are even more obvious using more recent programs which construct networks, such as Network (Bandelt *et al.* 1995, Bandelt, Forster and Röhl 1999), and NeighbourNet (Bryant & Moulton 2004): these programs draw trees to show common ancestry, but introduce links, or reticulations between languages connected by contact. They therefore represent an interesting alternative to the family tree in representing the real, multi-dimensional complexity of language histories, relationships and interinfluences.

Our work so far, however, has focused on Indo-European. This is an essential first step in demonstrating the utility of new approaches, since it is incumbent on us to show we can capture what is known, before moving into the unknown. In this paper, we shall apply meaning sublists and network programs to a far less clear linguistic situation. We hope to show that the existence of borrowings, and in particular their apparent concentration in certain sections of basic vocabulary lists, can be a positive benefit in disentangling debated language histories.

2. The Andean Languages as Conundrum and Test-Case

Two main indigenous language families survive in the central Andes (Figure 1, supplementary material accessible at http://www.philsoc.org.uk/transactions.asp). Quechua is spoken by around 8 million people, mostly in Ecuador, Peru and Bolivia. Aymara[2] has some 1½ million speakers in northern Bolivia and neighbouring regions in Chile and the far south of Peru; a tiny isolated group in the central highlands of Peru still speak the related Jaqaru/Kawki language. Together Aymara and Jaqaru/Kawki form a family variously known as Aru, Jaqi, Aymaran, or simply Aymara. The most fundamental question of comparative linguistics for the Andes is whether these two families are ultimately related to each other, a long-running debate for which a special term has been coined: the 'Quechumara' question.

The question has arisen because of a raft of uncanny similarities between the language families, not least in their morphosyntactic and phonological systems (for a review, see Cerrón-Palomino (1995)). A number of the most striking parallels are found between the southern varieties of the families known to have been in intense contact over centuries, though others appear to go back to the proto-languages.

There are also enticing form-to-meaning correspondences in lexis: to take just one well-known example, Cuzco Quechua *yaĉa-*[3] *know* stands alongside Aymara *yati-*. For over a century, traditional estimates (see Cerrón-Palomino (2000: 311)) have claimed that between 20% and 30% of Quechua and Aymara vocabulary is held in common, as supported by the calculations in Adelaar (1986).

Moreover, early investigators were struck by a small number of *repeated* apparent correspondences of the type so important for the

[2]Terminology for Andean language names is notoriously inconsistent. We follow the arguments and proposal in Cerrón-Palomino (1993) for terms in Spanish, and here use the corresponding English terms. So we speak of the language families as a whole as Quechua and Aymara, and for any particular dialects or languages within them we always specify which, thus: Cuzco Quechua, southern Quechua, southern Aymara, central Aymara (i.e. Jaqaru/Kawki), etc.

[3]We follow here the convention in Andean linguistics by which the symbols <ŝ> and <š> represent respectively plain and retroflex fricatives; <čˇ> and <ĉ> plain and retroflex affricates.

comparative method, such as Cuzco Quechua [č] to Aymara [t] in the *know* example. On closer inspection, however, many were found to go back to identical sounds in Proto-Quechua and Proto-Aymara, in this case an assumed *[č] in **yaĉa-* ∼ **yaĉi-* (according to the reconstructions by Cerrón-Palomino (2000: 311)). Indeed, as reconstructions of the two proto-languages progressed, most form-to-meaning correspondences actually appeared more consistent with borrowing than with a remote common origin. We are left with precious few truly *regular* correspondences in *different* sounds at the level of the proto-languages: certainly not enough to constitute compelling evidence of common origin. Clearly, however, the historical context in which both families developed has been characterised by indisputably intense, long-term contact; indeed, the most plausible homeland hypotheses have for both these families converged on regions of the central coast and highlands of Peru. Latest views (reviewed in Cerrón-Palomino (2000: ch.7)) see first Aymara, then Quechua, spreading southwards (and a separate branch of Quechua northwards at some point). This scenario broadly fits with the degrees of divergence between the most different members of each family. For Quechua this has been likened informally to the differences between Spanish and Portuguese, or Spanish and Italian; the difference within Aymara between the main southern variety of Aymara and Jaqaru is generally felt to be rather greater.

What is not in doubt is the quality or quantity of the correspondences. The **yaĉa-* ∼ **yaĉi-* example is just one of many hundreds of equally clear and often identical form-to-meaning correspondences that seem to go back all the way to Quechua and Aymara proto-forms. Matches so numerous and so close clearly exclude chance as an explanation: whether they reflect contact or common origin, there is unquestionably *some* very direct connection between the Quechua and Aymara language families.

Torero (2002: 154) has recently suggested that there is no *demonstrable* relatedness between the families, so that all such evident correspondences are to be ascribed to contact and convergence. Typically, however, the tone most specialists adopt on this issue remains non-committal and balanced. Cerrón-Palomino is pointedly careful not to claim that the parallels he identifies, however striking, convincingly favour either explanation over the

other: 'In sum, the question of the common origin of the two languages is not a closed case' (Cerrón-Palomino (2000: 337)); and more recently still, Cerrón-Palomino (2003: 22) 'The convergence hypothesis … should not and cannot rule out its counterpart of remote genetically common origin.' (Throughout this article, translations of all quotations are Heggarty's.) Despite a radical difference of emphasis between these two leading Peruvian specialists in the Andean languages, both nonetheless articulate the closest thing there is to a consensus on the Quechumara question among linguists. The key lies in Torero's word *demonstrable*.

On the one hand, the nature of many of the correspondences between Proto-Quechua and Proto-Aymara is beyond what – at least to most comparative and historical linguists used to Indo-European languages – appears consistent with contact alone. Many specialists sign up to almost identical phonological inventories for the two proto-languages, for instance. So if one wishes to invoke a convergence scenario, it needs to be more of the intensity suggested by Dixon (1997) for Australian languages, and much stronger than the more superficial contact effects found in Indo-European, even in the Balkans.

On the other hand, any signal in the Quechumara data may be beyond the limits inherent in the comparative method. The further one looks into the past, the weaker any surviving signal of clear correspondences consistent only with common origin, until eventually we reach a point where apparent correspondences become indistinguishable from chance similarities. The picture is all the more confusing when known intense population contacts in more recent times bring a host of new superficial interactions between the languages to overlay, swamp and confuse any old, weak signal of common origin that there *might* originally have been. Unfortunately for our purposes, this appears very much in line with the historical scenario we can imagine held sway in the Andes over the last millennium or more.

Our traditional techniques, then, can no longer establish that these languages *demonstrably* – beyond reasonable doubt – go back to a single common ancestor; though importantly, nor can they exclude that. So, four decades after the first serious linguistic approaches to the question, and following a great deal of research

and debate, the jury is still out. Novel methodological approaches have been tried, notably by Adelaar (1986) and Campbell (1995), but Campbell (1995: 195) himself recognises that they still leave us working with 'evidence [that] is suggestive, but also falls short of confirming the proposed relationship'. As Cerrón-Palomino (2000: 337) sums up: 'Despite [their] attempts… we are left with the impression of having made little progress in elucidating the problem'. In particular, Campbell (1995: 181) notes that 'As Parker (1973: 109) indicated, unless a satisfying way of distinguishing loanwords from plausible cognates is determined, there will never be a consensus in opinions concerning the genetic relationship or lack thereof between these two families.'

It is in this context that our own research should be seen. For to make progress, we shall need some new techniques that offer novel and/or more refined approaches to the linguistic data. While all this makes the Quechumara issue a particularly hard nut to crack, it also makes it an ideal test-case for any proposed method that seeks to help elucidate the question of common origin *vs.* convergence.

3. A New Approach to Measuring Similarity in Lexical Semantics

3.1. Cognates or Loanwords?

Given the current state of our knowledge, all we have is identifiable form-to-meaning parallels like **yaĉa-* ∼ **yaĉi-* ('know') or **warmi* ∼ **marmi* ('woman') which may have materialised by common origin or contact. It therefore remains quite unclear *a priori* whether Quechua and Aymara are related at all. But this would appear to rule out any form of lexicostatistics, since that method still rests on the notion of cognacy: in principle, it can only be meaningfully applied to languages already known to be genealogically related.

Not that this has always deterred attempts to apply just that method here. Apart from Torero's (1972) glottochronology for Quechua varieties only (and work by Hardman on Aymara, which is cited by Torero (1972: 54) but which we have been unable to

access), the only major lexicostatistical study of the Andean languages is Büttner (1983), who did indeed seek to use his lexicostatistical data as 'evidence' that the Quechua and Aymara families are related (and, he specifies, at the stock level). The circularity of Büttner's conclusions has been pointed out by many critics, including Torero (2002: 149) and Cerrón-Palomino (2003: 372–373), and can be traced back to the fundamentally flawed assumption that the meanings in the Swadesh lists are necessarily immune to borrowing (see Embleton 1986).

Not only do we need a novel methodological approach to apparent form-to-meaning correspondences; we also need a term that does not imply either explanation, unlike *cognate*, *borrowing*, or even *correspondence*. Kessler (2001) talks simply of *historically connected* words: we propose the term *correlate*, which we prefer due to its formal similarity with *cognate*.

Our *correlate* is a cover term for any striking form-to-meaning correspondence more convincingly attributable to some (unspecified) historical connection than to chance. Rather than attempt to decide in advance between common origin or borrowing, we prefer to use a different tool with 'diagnostic' power for telling apart signals of common origin and contact (see section 4).

We are well aware that most specialists in the Andean languages consider a large number of the known correlates to be evident loanwords. The problem is that there are also plenty of words on which analyses disagree, which may be prejudged depending on one's position on whether Quechua and Aymara are related. The only solution is the methodological one Adelaar (1986: 380) points out, seconded by Torero (2002: 155) who cites him as follows: 'Any progress in the question of Quechua-Aru [*i.e.* Aymaran] relationships "presupposes methodologically that all preconceived ideas about genetic relationship be entirely abandoned"'. For the purposes of our method, all Quechua $\sim$ Aymara correspondences are correlates, nothing more.

We should also add that Spanish loanwords – which *are* mostly very easily identifiable – are scored in this study as 'no data': they are determined mostly by social factors, especially endangerment and remoteness, and can only serve to mask and confuse the origins of and relationships between the Andean languages.

3.2. Sensitive Measures of Similarity in Lexical Semantics

The second key methodological flaw that condemned Büttner's (1983) study was his approach to the data. Firstly, he used secondary data from the relatively poor documentation then available on the Andean languages. In our study we use primary data from Heggarty's own fieldwork in the Andes between 2001 and 2004, which can be consulted on the internet at http://www.quechua.org.uk.

Moreover, Büttner's approach to the data was dangerously haphazard: witness comments like Cerrón-Palomino's (2003: 372): 'astonishingly naïve choices … arbitrary in his selection'. However, Büttner was not helped by the inflexibility of the traditional lexicostatistical method, nor by the Swadesh 200 meaning-list, since many meanings prove to be unrealistic and unsuitable for the Andean languages.

Criticisms of lexicostatistics are hardly news to comparative and historical linguists. Overcoming them requires us to confront the most basic and widely criticised flaw in the method: its simplistic insistence on 'one meaning one word'. From the beginning lexicostatistics was led by the *data format*: it represents the linguistic data using only binary all-or-nothing values (needed in order to use the lexicostatistical results for glottochronology, which we steer well clear of).

If we have the benefit of prior application of the comparative method, along with a fairly secure and well documented history of the languages concerned, we can still use traditional lexicostatistics, since we will be able to arrive at reasoned cognacy judgements, and use our computational methods to test for any exceptions: this is the approach we have used for Indo-European (McMahon & McMahon 2003, 2004, forthcoming). However, if the comparative method is not applicable, as in the Andean situation, we need a different approach; and here we use a method developed in Heggarty (in preparation)[4]. In one sense, this is a 'refined lexicostatistics'; but it starts anew from first

[4]Individual scholars may apply to us as of now for an advance copy of the relevant chapters.

principles for comparing and quantifying languages. The goal is to represent as meaningfully as possible in numbers the wider concept of the *degree of similarity between languages in their lexical semantics*.

Granted, the method is based ultimately on how far different languages use lexemes that are or are not correlate, but there the similarity with lexicostatistics ends. Crucially, this method departs from the earlier assumption of 'one meaning, one lexeme'. Instead, Heggarty's database format represents, in as detailed, sensitive and balanced a way as possible, the exact nature of the overlaps and differences between languages in their lexical semantics. More refined quantification results simply follow from this, as we shall see. As before, however, the real utility and new insights that such quantifications provide lie in what we can do with our understanding of linguistic data and relationships once they have been encoded into numerical representations. In particular, the raw results can be synthesised and converted into a range of representations which help us interpret what the patterns of similarity between the Andean languages really mean for their origins and historical development.

3.3. An illustration of the method

Since the method and calculations we use are much more complex than for traditional lexicostatistics, there is not space here to cover them in any real detail, so we refer interested readers to the full exposition in Heggarty (in preparation). We are limited here to an overview of the specific mechanisms built into our database and processing program to make them sensitive to the three main levels or 'scopes' for partial overlap between languages in their lexical semantics that the binary approach of lexicostatistics cannot accommodate.

The most important scope for partial overlap involves correlate lexemes patterning differently across the various SUB-SENSES of a list-meaning, and indeed more indirectly RELATED SENSES. The list-meaning *sun* in (1) provides a useful illustration of some of the many different types of overlap possible, for just four of our Andean varieties.

(1) 'Sun' in four Andean varieties.

- Atalla Quechua: has *inti* in the *celestial object* sense, but *rupa-y* in the sense of *sunlight/heat of the sun*
- Chetilla Quechua: has *rupa-y* only; *inti* is unknown (except indirectly through Spanish)
- Laraos Quechua: has *inti*; *rupa-y* only in the related verb root *be hot* (*sunny*), *burn*
- Puki Aymara: has *inti* only; *rupa-y* is entirely unknown, not even as *be hot*, *burn*

From (1) a whole range of complex patterns and degrees of overlap emerge between the various pairs of languages, which can usefully be expressed in terms of INTELLIGIBILITY. Indeed we can arrange the pairs in a descending scale of mutual intelligibility, as in (2).

(2) Descending intelligibility scale for Andean varieties.

- Laraos ~ Puki Full correlates *inti* in all senses of *sun*.
- Laraos ~ Atalla Full correlates in the *celestial object* sense; in the *sunlight* sense the Atalla speaker would understand a slightly different sub-sense, and the Laraos speaker only a rather more different related meaning (*burn*).
- Puki ~ Atalla Full correlates in the *celestial object* sense; in the *sunlight* sense the Atalla speaker would understand a slightly different sub-sense, and the Puki speaker would not recognise the Atalla root at all.
- Chetilla ~ Atalla In the main *celestial object* sense the Atalla speaker would understand a slightly different sub-sense (*sunlight*), and the Chetilla speaker would not recognise the Atalla root at all; in the secondary *sunlight* sense they share full correlates.

- Chetilla ~ Laraos — In both the main *celestial object* sense and the secondary *sunlight* sense the Laraos speaker would understand only a quite different though still related meaning (*burn* or *be hot/sunny*), while the Chetilla speaker would not recognise the Laraos root at all.
- Chetilla ~ Puki — No correlates in either sense: each variety has only one lexeme, not correlate to the other's, nor even recognised as related meanings.

For these six cases, our representation and quantification system produces a descending scale of intelligibility ratings in line with the unquantified linguistic scale above, as follows: 1, 0.83, 0.78, 0.56, 0.17 and 0. Note that traditional lexicostatistics, as in Dyen *et al.* (1992), would only give figures of 0 or 1 for all of these relationships. An overview of quite how our model came to these figures is set out in detail in Figure 2. Simplifying the procedure somewhat, the method recognises what is effectively a scale of different degrees of overlap in meaning between correlates:

- *full*, precise intelligibility;
- *close* overlap in close sub-senses of the same list-meaning that gives still high intelligibility;
- *indirect* overlap in a rather different meaning but still one related closely enough within the same general semantic field to contribute to at least a certain degree of intelligibility;
- *no* intelligibility at all.

In the absence of a clear linguistic case for more specific figures, these categories are by default represented in the simplest numerically equal steps, namely as overlap ratings of 1, 2/3, 1/3 and 0 respectively.

Figure 2 also illustrates how our method incorporates the fact that one of the two sub-senses of *sun* is rather more basic than the other: namely the *celestial object* sense, rather than the narrower focus on the *light*/*heat* that comes from it. Respecting our default weighting principles, this relationship is represented by the simplest possible unequal ratio, 2:1.

Our model also accommodates MULTIPLE TRUE SYNONYMS for any single list-meaning, as for example for *head*, for which our Chacpar

LANGUAGES	SUB-SENSE A: *celestial object*						SUB-SENSE B: *sunlight/heat*						WEIGHTING COMPONENTS		OVERALL OVERLAP WEIGHTING
IN BOTH INTELLIGIBILITY DIRECTIONS: LANGUAGE[1] → LANGUAGE[2]	ENTRY IN LANGUAGE[1]	WHICH IN LANGUAGE[2] IS RECOGNISED AS: SAME SUB-SENSE[A]	DIFFERENT SUB-SENSE[B] ONLY	RELATED MEANING	UNRECOGNISED	AVERAGE FOR BOTH INTELLIGIBILITY DIRECTIONS	ENTRY IN LANGUAGE[2]	WHICH IN LANGUAGE[1] IS RECOGNISED AS: SAME SUB-SENSE[B]	DIFFERENT SUB-SENSE[A] ONLY	RELATED MEANING	UNRECOGNISED	AVERAGE FOR BOTH INTELLIGIBILITY DIRECTIONS	SUB-SENSE A WEIGHTED × 2/3	SUB-SENSE B WEIGHTED × 1/3	= SUM OF TWO WEIGHTED SUB-SENSE COMPONENTS
Laraos → Puki	*inti*	1				1	(*inti*)	(1)				(1)	2/3	(1/3)	1
Puki → Laraos	*inti*	1					(*inti*)	(1)							
Laraos → Atalla	*inti*	1				1	(*inti*)		2/3			1/2	4/6	1/6	5/6 = 0.83
Atalla → Laraos	*inti*	1					*rupa-y*			1/3					
Puki → Atalla	*inti*	1				1	(*inti*)		2/3			1/3	6/9	1/9	7/9 = 0.78
Atalla → Puki	*inti*	1					*rupa-y*				0				
Chetilla → Atalla	*rupa-y*		2/3			1/3	(*rupa-y*)	1				1	2/9	3/9	5/9 = 0.56
Atalla → Chetilla	*inti*				0		*rupa-y*	1							
Chetilla → Laraos	*rupa-y*			1/3		1/6	(*rupa-y*)			1/3		1/6	2/18	1/18	3/18 = 0.17
Laraos → Chetilla	*inti*				0		(*inti*)				0				
Chetilla → Puki	*rupa-y*				0	0	(*rupa-y*)				0	0	0	0	0
Puki → Chetilla	*inti*				0		(*inti*)				0				

Figure 2. Calculations of degrees of similarity ('overlap') for four Andean language varieties in their lexical semantics for the list-meaning *sun*. 1 = identity; 0 = total difference.

Quechua informant uses both *uma* and *piqa*, while Cuzco Quechua speakers have only *uma*, a pattern of correlates one can represent symbolically as AB *vs.* A. Multiple synonyms can also be weighted against each other for their relative frequency and importance.

Although the concept of intelligibility is built into our model, we use a 'linguistically informed' version. In order to measure similarity only in lexical semantics, we abstract away from known sound changes that may have left different phonetic reflexes of correlates across different languages. (We measure similarity in the quite distinct field of phonetics using a completely different method, also set out in Heggarty (in preparation); and see Kessler, this volume). Furthermore, intelligibility measures just the *passive* overlap between the systems: this fails to reflect other differences, for example in the frequency of shared synonyms in active use. To represent differences in *active* overlap (e.g. the same synonyms but used with different frequencies in different languages), comparisons require corresponding adjustments to the overlap figures, normally small reductions. The result is a scale of different degrees of synonym overlap, as shown in Figure 3. Again, even those approaches to lexicostatistics that allow multiple synonyms, such as Dyen *et al.*'s, would represent the first three all as 100% similarity.

Lexicostatistics is also blind to partial overlap between lexeme pairs composed of MULTIPLE MORPHEMES, some but not all of which are correlate, the standard example being the Spanish and French lexemes for 'heart', namely *corazón* and *coeur*. Here the roots are correlate, but the Spanish lexeme includes a suffix-*azón* (< Latin -*atĭo*, -*ōnis*) that has no corresponding morpheme in the French lexeme. Cognacy (and our *correlation*) is an attribute of *morphemes*, not necessarily of whole lexemes. We therefore must segment lexemes into their component morphemes, and ask whether those individually, not the whole lexemes, are correlate. In such cases too the overall overlap rating emerges at less than 1 but greater than 0. The exact value depends again on weightings assigned to the morphemes where necessary to reflect their relative contribution to overall meaning and intelligibility. In *corazón* ~ *coeur*, for example, the suffix -*azón* would be weighted less than the root, and overall overlap rated at 2/3.

LANGUAGE	CORRELATES PRESENT (AS SYNONYMS)			SIMILARITY RATING BETWEEN LANGUAGE[1] AND LANGUAGE[2]
language[1]	A	B	-	
language[2]	A	B	-	1
language[2]	A	-	-	$^2/_3$
language[2]	A	-	C	$^1/_2$
language[2]	-	-	C	0

Figure 3. Measurements of synonym overlap: various scenarios.

In the Andean languages, multiple morpheme forms are often compound nouns, e.g. Jaqaru *šim(ɨ) ts'aaka* for *tooth*, literally *mouth bone*. Complex partial overlaps of this type are surprisingly frequent across many Andean varieties, not least Ecuadoran Quechua, quite plausibly as a result of a hypothesised period of *koine*development (Muysken 1981).

3.4. Correlate Plausibility Levels

Finally, the Quechumara question asks whether certain apparent form-to-meaning correspondences really constitute 'correlates' sufficiently striking to call for an explanation other than chance; and/or whether proposed reconstructions are convincing, or highly speculative. Our data-set specifies all debatable cases along a 0–7

scale expressing levels of 'plausibility': how far does the degree of phonetic similarity between correlate sets appear to constitute a correlation significantly greater than chance? Assessments are based on a number of principles, and derived from known sound changes in the Andean languages. They also draw on Cerrón-Palomino's (2000: 311) categories of obvious loanwords, very probable cognates, probable cognates, and obviously unrelated forms.

To give some outline examples, the cases above where two varieties both have *inti*, and therefore identical forms, would score 7; *inti* ~ *rupa.y* would score 0, at the other extreme, since there is really no basis for assuming correlateness in this case; *p'iqi* ~ *piqa* is rated at 5, but **quλu* ~ **urqu* (=3) and **huma* ~ **qam* (=2) are less convincing and more speculative. We are well aware that these characterisations are in part subjective and impressionistic; but scalar measures minimise the impact of such subjectivity on the figures, causing generally at worst a shift of the order of 0.1 to 0.2, rather than the wild swing of 0 to 1 which would arise from any misidentification in traditional lexicostatistics. These plausibility ratings allow different sets of results to be produced for each level, *i.e.* for a range of more speculative and more conservative approaches to possible correlations in the data: here, we adopt the fairly conservative position that any relationship rated at 5 or above counts as correlate.

3.5. Adjusting the Meaning-List for the Andean Languages

Finally, our study needs to be sensitive to the nature of the languages it is applied to. As many linguists who have sought to apply lexicostatistics outside Europe have pointed out, Swadesh-lists need considerable adaptation for use with structurally very different languages originating in very different cultures. Perhaps the best-known example is Matisoff's (1978, 2000) CALMSEA list, a well-chosen acronym: **C**ulturally **a**nd **L**inguistically **M**eaningful for **S**outh-**E**ast **A**sia. For this study Heggarty devised our own 'CALMA' meaning-list incorporating items **C**ulturally **a**nd **L**inguistically **M**eaningful for the **A**ndes.

The result has been a quite radical reworking and pruning of Swadesh's lists, as set out in detail in Heggarty (in preparation). Of

our total 150 list-meanings (see Figure 4 of our supplementary material available at http://www.philsoc.org.uk/transactions.asp), 85 are in Swadesh's 100 meaning-list, and another 30 in his 200 meaning-list. The remaining 35 are meanings often used in other lexicostatistical studies, which we brought in deliberately in order to ensure a balance between subsets of meanings known to be particularly stable over time, and others known to be much less stable, allowing analogues of our hihi and lolo lists for Indo-European to be developed. Embleton (1986) has pointed out the loss of resolution entailed where lists include fewer than 200 meanings; but since our method differs from traditional lexicostatistics in measuring *degrees* of similarity between 0 and 1 for any one list-meaning, we obtain much greater resolution from each meaning, which arguably compensates for covering rather fewer of them.

4. Andean applications

We turn now to an application of our more refined method to the Quechumara question. Our database here involves the CALMA 150-meaning list, for 14 varieties of Quechua; 3 varieties of Aymara; and Kawki and Jaqaru, which are typically classified as independent Aymara languages. This combination of lists and varieties with our more sophisticated techniques for scoring correlateness will provide a graded rating of similarity between varieties and languages; but that is all. Any network plotted from these results will be purely a phenogram, giving information on distance, and not a phylogram, which tells us about the history of the different groups. And yet if we are to make any progress towards answering the Quechumara question, is it not precisely insight into the more likely history that we need?

The answer, again, lies in our use of sublists, which allow us to place a historical interpretation on our phenetic results. We have excerpted from Heggarty's Andean database two groups of 30 items corresponding to our Indo-European hihi and lolo sublists, the former being most retentive, and the latter most prone to change and borrowing. These sublists are shown in (3), and though membership is not identical with the hihi and lolo lists for

Indo-European, the overlap has been maximised as much as possible given the different compositions of the Swadesh and CALMA lists (overlapping items are shown in bold).

(3)a. Andean hihi list, 30 items

one	**two**	**three**	**four**	**five**
I	**thou** (you sg.)	**not**	**ear**	**tongue**
tooth	**foot**	**fingernail (claw)**	heart	**name**
day	**night**	**sun**	**star**	shadow
wind	**salt**	green	**new**	**come**
eat	**sleep**	live (be alive)	**give**	sew

b. Andean lolo list, 30 items

year	**left** (hand side)	face	**mouth**	lip
neck	(upper) **back**, shoulder	skin (human)	breast	**bird**
tail	**wing**	**man (male adult)**	**river**	**stone**
bread	branch	**grass**	**rope**	red
straight	sick (be ill)	empty	**heavy**	far (away)
hot	**walk**	swim	**think**	**push**

These Andean sublists can be shown to be differentially affected by borrowing in the same way as our parallel sublists for Indo-European. Spanish borrowings can be identified relatively readily in all the Andean languages and varieties, and we find an average of 2.7% Spanish loans in the hihi sublist, and 6.7% in the lolo sublist, nearly three times as high. This difference is significant at the $p < 0.001$ level (paired t-test; $t = -4.1$, $df = 18$).

We can now use these sublists to generate networks showing degrees of similarity among the Andean languages and varieties. Since we are using graded, scalar data to obtain overall similarity scores, we require a network program which handles distance-based rather than character-based data (see McMahon & McMahon (forthcoming, Chapter 6) for further discussion). The graphs in Figure 5 were therefore generated using NeighbourNet (Bryant and Moulton 2004).

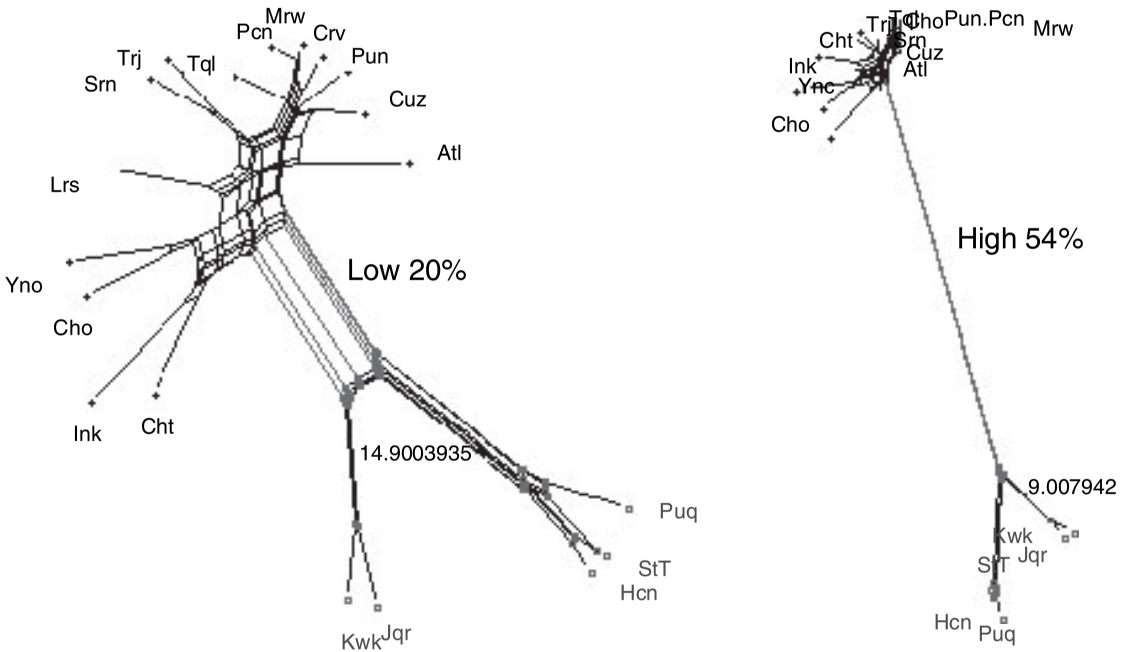

Figure 5. Comparison of lexical distances between the Quechua and Aymara groups of Andean languages, lolo sublist on the left, hihi on the right, drawn using NeighbourNet. (Key to varieties in Figure 1, supplementary material at http://www.philsoc.org.uk).

It is very clear in both these graphs that the 14 Quechua dialects cluster together, at the top of the networks; so do the three Aymara varieties at the bottom, plus Jaqaru and Kawki, which however constitute a separate branch within Aymara. The most interesting aspect of these graphs, however, is the calculation of distance between the root nodes for those two Quechua and Aymara groups, which in the lolo network is just over 20%, while for the hihi one it is 54.4%, nearly three times as high. Lexical distance here is simply another way of expressing percentage non-correlateness, with 10% approximating to a distance of 3 lexical items. What this means, then, is that the distance between Quechua and Aymara is considerably less for the lolo sublist, which is typically more changeable, than for the hihi items, which are generally more resistant to change.

This is not the pattern we find in cases where we know, or can at least hypothesise with reasonable confidence, that languages are related. In the same graphs in Figure 5 above, we see that within the Aymara cluster, the lexical distance from Kawki and Jaqaru to the Aymara root is 14.9% for the lolo graph on the left, but 9%, rather

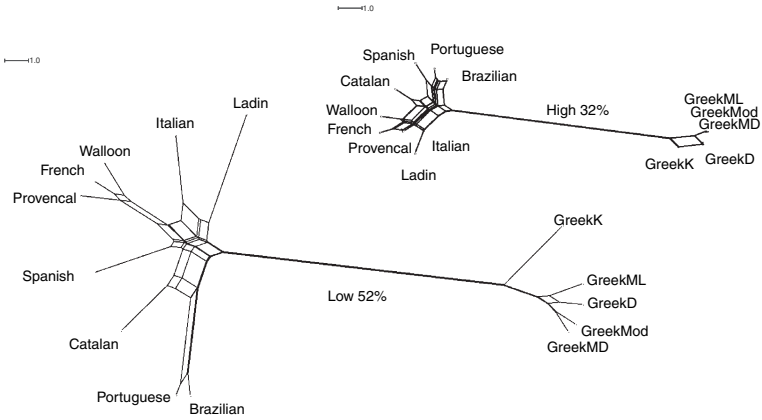

Figure 6. Comparison of lexical distances between Greek and Romance, lolo sublist on the left, hihi on the right, drawn using NeighbourNet (not to scale).

lower, for the hihi graph on the right. As shown in Figure 6, we find similar results in the case of Greek and a sample of Romance languages (from the Dyen, Kruskal & Black 1992 database), chosen simply because they represent a good comparison with Quechua and Aymara, as two Indo-European groups of comparable sizes and overall lexical distance from each other.

The comparable distances between Greek and Romance are 52% for the lolo sublist and 32% for the hihi group: these results were obtained using traditional, cognate-based, binary lexicostatistics, but the pattern is nonetheless the same as for Aymara compared with Kawki and Jaqaru. For the two Aymara groups, and the two Indo-European ones, we therefore find greater distance for the lolo sublist, and greater similarity for the hihi meanings. This is precisely the opposite of the pattern shown in Figure 5 for Quechua compared with Aymara. It would appear, then, that our figures argue against common ancestry, against the hypothetical Quechumara family, and for a relationship of contact alone. Common ancestry corresponds regularly to calculations of greater distance for the lolo subgroup than for the hihi one, simply because the lolo items, by definition, are more likely to change. Comparing Quechua with Aymara, we find instead three times as much distance is apparent for the hihi than for the lolo items. We conclude that if, as in this case, two groups show less affinity in the sublist which is more stable, then contact seems the most appropriate explanation, and common origin can at best be remote.

These results, of course, reflect work in progress, and must be taken as indicative and requiring further investigation, rather than absolutely conclusive. None of these calculations or networks proves that Quechua and Aymara never shared a common ancestor, but they do indicate that contact is the main determinant of the lexical similarities between the two groups. Campbell (1995: 195) objects that many criticisms of the Quechumara hypothesis are 'beside the point' in providing only evidence *for* contact (which everyone involved already accepts as a contributory factor), and not specifically *against* common origin. Here we can go further, however: common origin does appear to be a distinctly more strained explanation for our results than contact *alone*. Even if Quechua and Aymara did go back ultimately to a common origin,

our comparisons, at least on initial analysis, suggest that we would have to place any Quechumara language even further back in time than Proto-Indo-European. For more detailed discussion, see Heggarty (forthcoming).

In his methodological approach to the Quechumara question, Campbell (1995: 182) specifically sought to exclude possible loanwords arguing that 'Only quite basic vocabulary should be compared, lexical items most unlikely to be loans', and that 'forms which are quite similar phonetically should be discarded as possible loans'. However, if we attempted to remove loans from our data *a priori*, or took the view that Swadesh-type lists are so resistant to borrowing that the presence of loans should not be an issue, we could not use sublisting as a technique for discovering likely loans and would lose an opportunity to cast light on language histories which cannot be illuminated using more traditional methods of comparison. If, on the other hand, we recognise that borrowing is both inevitable and interesting, and develop methods for finding and representing its contribution, we may be able to answer the more recalcitrant questions of historical linguistics, like the status of Quechumara. It is in such cases, we would suggest, that we observe the benefits of borrowing.

English Language
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
14 Buccleuch Place
Edinburgh EH8 9LN
Email: April.McMahon@ed.ac.uk

References

Adelaar, Willem F. H., 1986. 'La Relación quechua-aru: Perspectivas para la separación del léxico', *Revista Andina* 4:2, 379–426.

Bandelt, H.-J., Forster, P., Sykes, B. C. & Richards, M. B., 1995. 'Mitochondrial portraits of human populations using median networks', *Genetics* 141, 743–53.

Bandelt, H.-J., Forster, P. & Röhl, A., 1999. 'Median-joining networks for inferring intraspecific phylogenies', *Molecular Biology and Evolution* 16, 37–48.

Bryant, David & Moulton, V., 2004. 'NeighbourNet: An agglomerative algorithm for the construction of planar phylogenetic networks'. *Molecular Biology and Evolution* 21, 255–65.

BÜTTNER, THOMAS TH., 1983. *Las Lenguas de los Andes centrales*, Madrid: Ediciones Cultura Hispánica.
CAMPBELL LYLE, 1995. 'The Quechumaran hypothesis and lessons for distant genetic comparison', *Diachronica* XII:2, 157–99.
CERRÓN-PALOMINO, RODOLFO, 1993. 'Quechuística y aimarística: una propuesta terminológical', *Alma Mater* 5, 41–55.
CERRÓN-PALOMINO, RODOLFO, 1995. *Quechumara: estructuras paralelas de las lenguas quechua y aimara*, La Paz, Bolivia: Centro de Investigación y Promoción del Campesinado.
CERRÓN-PALOMINO, RODOLFO, 2000. *Lingüística Aimara*, Cuzco, Peru: Centro Bartolomé de las Casas.
CERRÓN-PALOMINO, RODOLFO, 2003. *Lingüística Quechua*, Cuzco, Peru: Centro Bartolomé de las Casas.
DIXON, R. M. W., 1997. *The Rise and Fall of Languages*, Cambridge: Cambridge University Press.
DYEN, ISIDORE, KRUSKAL, JOSEPH B. & BLACK, PAUL, 1992. 'An IndoEuropean Classification: A lexicostatistical experiment', *Transactions of the American Philosophical Society* 82.
EMBLETON, SHEILA, 1986. *Statistics in Historical Linguistics*, Bochum: Brockmeyer.
HALL, ROBERT A., 1960. *Linguistics and Your Language*, New York: Anchor Books.
HEGGARTY, PAUL, forthcoming. 'Enigmas en los orígenes de los idiomas andinos: nuevos métodos aplicados a las preguntas sin contestar'. *Revista Andina* 40.
HEGGARTY, PAUL, in preparation. *Measured Language*, to be published by Blackwell.
HOENIGSWALD, HENRY M., 1960. *Language Change and Linguistic Reconstruction*, Chicago: University of Chicago Press.
HOENIGSWALD, HENRY M., 1973, *Studies in Formal Historical Linguistics*, Dordrecht: Reidel.
HOENIGSWALD, HENRY M., 1990. 'Does language grow on trees? Ancestry, descent, regularity', *Proceedings of the American Philosophical Society* 134, 10–18.
KESSLER, BRETT, 2001. *The Significance of Word Lists*, Stanford: CSLI Publications.
LOHR, MARISA, 1999. *Methods for the Genetic Classification of Languages*. PhD dissertation, University of Cambridge.
MATISOFF, JAMES A., 1978. *Variational Semantics in Tibeto-Burman*, Philadelphia: Institute for the Study of Human Issues.
MATISOFF, JAMES A., 2000. 'On the uselessness of glottochronology for the subgrouping of Tibeto-Burman', in Colin Renfrew, April McMahon and Larry Trask (eds.) *Time Depth in Historical Linguistics*, Volume 2. Cambridge: McDonald Institute for Archaeological Research, 333–371.
MCMAHON, APRIL & ROBERT MCMAHON, 2003. 'Finding families: quantitative methods in language classification', *Transactions of the Philological Society* 101.1, 7–55.
MCMAHON, APRIL & ROBERT MCMAHON, 2004. 'Family values', in Christian Kay, Simon Horobin and Jeremy Smith (eds.), *New Perspectives on English Historical Linguistics, Volume 1: Syntax and Morphology*, Amsterdam: Benjamins, 103–123.
MCMAHON, APRIL & ROBERT MCMAHON, forthcoming. *Language Classification by Numbers*, Oxford: Oxford University Press.
MUYSKEN, PIETER, 1981. 'El Quechua del Perú y Ecuador: una vision comparativa', paper presented at the Congreso Internacional en Homenaje a Andrés Bello, Panama.

PARKER, GARY, 1973. 'On the evidence for complex stops in Proto-Quechua', *International Journal of American Linguistics* 39, 106–110.

PULGRAM, ERNST, 1995. 'Proto-languages in prehistory: reality and reconstruction', *Language Sciences* 17, 223–39.

SWADESH, MORRIS, 1950. 'Salish internal relationships', *International Journal of American Linguistics* 16, 157–67.

SWADESH, MORRIS, 1952. 'Lexicostatistic dating of prehistoric ethnic contacts', *Proceedings of the American Philosophical Society* 96, 452–63.

SWADESH, MORRIS, 1955. 'Towards greater accuracy in lexicostatistic dating', *International Journal of American Linguistics* 21, 121–37.

TORERO, ALFREDO, 1972. 'Lingüística e historia de la sociedad andina', in A. Escobar (ed.), *El reto del multilingüismo en el Perú*, Lima, Peru: Instituto de Estudios Peruanos.

TORERO, ALFREDO, 2002. *Idiomas de los Andes – Lingüística e Historia*. Lima, Peru: Editorial Horizonte/IFEA.