On the polymorphemic genesis of some Proto-Quechua roots: establishing and interpreting non-random form/meaning correspondences on the basis of a cross-linguistic polysemy network

Emlen, N.Q.; Dellert, J.

On the polymorphemic genesis of some Proto-Quechuan roots

Establishing and interpreting non-random form/meaning correspondences on the basis of a cross-linguistic polysemy network

Nicholas Q. Emlen[1,2] and Johannes Dellert[2]
[1] Leiden University Centre for Linguistics | [2] University of Tübingen

In the Proto-Quechuan lexicon, many two-segment phonetic substrings recur in semantically related roots, even though they are not independent morphemes. Such elements may have been morphemes before the Proto-Quechuan stage (i.e., in Pre-Proto-Quechuan). On the other hand, this may simply be due to chance, or to phonesthesia. In this paper, we introduce the Crosslinguistic Colexification Network Clustering (CCNC) algorithm, as well as an accompanying test statistic, which allow us to evaluate our claims against a neutral standard of semantic relatedness (the CLICS[2] database; List et al. 2018). We obtain very strong statistical evidence that there are hitherto unexplained recurrent elements within Proto-Quechuan roots, but not within roots reconstructed for Proto-Aymaran, the proto-language of a neighboring language family whose members are otherwise structurally very similar to Proto-Quechuan, and which has therefore long been considered an obvious candidate for deep shared ancestry. Some of these elements are explainable as phonesthemes, but most appear to reflect archaic Quechuan morphology. These findings are consistent with an emerging picture of the early Quechuan-Aymaran contact relationship in which Quechuan structure was reformatted on the Aymaran template.

Keywords: Quechuan, Aymaran, reconstruction, historical linguistics, polysemy networks, colexification, CLICS[2], Crosslinguistic Colexification Network Clustering (CCNC), archaic morphology, phonesthemes

https://doi.org/10.1075/dia.16041.eml | Published online: 5 August 2020
Diachronica 37:3 (2020), pp. 318–367. issn 0176-4225 | e-issn 1569-9714

1. Introduction

The Proto-Quechuan (PQ) lexicon exhibits a curious pattern: many phonetic sequences are shared by semantically related roots, even though they are not analyzable as synchronically independent morphemes. Consider, for instance, the reconstructed Proto-Quechuan roots in (1), from Emlen (2017), and the roots from modern Quechuan languages, in (2), which are not distributed widely enough in the family to justify reconstruction in Proto-Quechuan. All of these roots have to do with hanging, tying, and cord, and they all begin with /wa/ (the sources of lexical data from these languages, as well as the language abbreviations, are described in §5).

(1) Proto-Quechuan

*wata-	'to tie, repair'
*wanku-	'to wrap, bundle'
*waya-	'loose, to loosen'
*wayu-	'hanging fruit, to hang, to mature (fruit)'
*warku-	'to hang up'
*watu	'strap, cord, belt'
*waʎqa	'pendant'
*waska	'rope'

(2) Modern Quechuan languages

walt^{h}a-	'to wrap up, swaddle baby' (CUS)
wayʎunk'u-	'to swing' (CUS)
waʎq^{h}i	'flaccid' (BOL)
warpu-	'to tie llamas together' (CUS)
waʎqi	'hanging bag' (ANC, JUN)
waqi-	'to hang' (ANC)

Similarly, the Proto-Quechuan roots in Table 1 end in phonetic strings that are identical to the four known Proto-Quechuan verbal directional suffixes: *-rpu 'downward motion', *-rku 'upward motion', *-rqu 'outward motion' and *-yku 'inward motion' (these suffixes are discussed by Parker 1973: 21–27; Adelaar & Muysken 2004: 231; Adelaar 2006; Adelaar 2013a: 58, among others; see discussion in §7.2). In all cases, the roots exhibit semantics consistent with the corresponding directional suffixes. Notably, these final strings often appear after recurrent initial strings. Other roots sharing these initial strings are also shown at the bottom of Table 1. Roots from particular Quechuan languages, which are not distributed widely enough to be reconstructed in Proto-Quechuan, are also shown in italics.

Table 1. Some recurrent initial and final phonetic substrings in Proto-Quechua

PQ suffix	/tʂu/: 'put'	/qa/: 'herd'	/wa/: 'hang'	/ya/: 'go'	/su/: 'take'
*-rpu 'down'	*tʂurpu- 'to take down object, take pot from fire, put pot on fire'	*qarpu-* 'to drive or herd downwards' (TAR, ANC)		*yarpu-* 'to descend' (TAR, ANC, PAC, JUN *yalpu-*)*	*sulpu-* 'to put down (e.g., a pot from a fire' (JUN)'
*-rku 'up'	*tʂurku- 'to put an object in a high place'	*qarku-* 'to drive or herd upwards' (TAR, PAC)	*warku- 'to hang up'	*yarku-* 'to climb' (TAR, ANC, PAC, JUN *yalku-*)	*sulku-* 'to draw water from a well' (JUN)
*-rqu 'out'		*qarqu- 'to expel, throw out, drive out of corral'		*yarqu-* 'to go out, leave' (TAR, ANC, PAC, JUN *yalqu-*)	*surqu- to remove, take out, extract
*-yku 'in'		*qayku- 'to lead indoors, drive into a corral'		*yayku- 'to enter'	
others	*tʂura- 'to put, place'	*qati- 'to herd animals, pursue'	[see (1)–(2)]	*yaʎi-* (CUS, ECU) *ʎaʎi- 'to be victorious, exceed, surpass'	

* Junín Quechua underwent a *r > l sound change, which explains the /l/ in some of these forms.

The presence of such roots in the Quechuan languages presents a puzzle. There are no independent morphemes with the forms /wa/, /tʂu/, /qa/, /ya/ or /su/ with these meanings in Proto-Quechuan nor in any modern Quechuan variety, and the leftover phonetic material found alongside such recurrent elements only occasionally resembles identifiable Quechuan morphology. In this paper, we explore the possibility that Quechuan roots like the ones in (1), (2) and Table 1 contain archaic Pre-Proto-Quechuan morphemes – such as *wa- 'cord; to hang, tie' – which have been lexicalized within mostly bisyllabic Quechuan roots.

Making sense of this pattern presents a number of methodological challenges. In particular, we must account for the possibility that the co-occurrence of phonetic substrings in semantically related roots is merely a coincidence. In any sufficiently large set of lexical data, we would expect to find some number of spurious form/meaning correspondences. For instance, if nothing were known about the history of English, we might observe that the semantically related English words *wicked, witch* and *wizard* share a two-segment substring /wi/, and on that basis

reconstruct an Early English morpheme *wi meaning something like 'evil, magic'. There are two methodological problems here. First, if we allow ourselves to be liberal in which concepts we consider to be semantically related, and base those semantic relationships on observations of the dataset itself (for instance, using an ad-hoc semantic category like 'evil, magic'), we are sure to come up with spurious clusters. For this reason, it is necessary to apply a neutral, external measure of semantic relatedness to the data. Second, as a matter of probability, short phonetic sequences are likely to appear in some number of semantically related roots by chance (a problem demonstrated by Ringe 1992). This is particularly true given the limited number of phonotactic possibilities presented by most languages, and given that the language-specific skews in phoneme distribution also show universal tendencies. Our solution to these methodological problems is to employ a neutral standard of semantic relatedness – that is to say, a standard that was developed independently of our research questions, and that is thus impartial to them. To this end, we use the CLICS² cross-linguistic co-lexicalization database (List et al. 2018). We then develop a statistical test for rejecting the null hypothesis that the apparent phonetic regularities in semantically related roots (like those in (1)–(2) and Table 1) are due to chance.

The test is based on what we will call the "Crosslinguistic Colexification Network Clustering" (CCNC) algorithm.[1] Put briefly, this method draws on a semantic similarity graph based on cross-linguistic polysemies as an external resource for identifying clusters of semantically related roots which share some relevant feature (in this case, two-segment phonetic sequences). We then use a simple statistic on the output of CCNC in order to measure whether lexemes within semantic clusters share that feature more frequently than we would expect by chance. The clustering method is described in §3, and our statistical tests for establishing the non-independence of shared features and semantic similarity are described in §4. Next, after describing our data in §5 – a list of 809 Proto-Quechuan reconstructed lexical roots and 259 Proto-Aymaran reconstructed lexical roots, which exclude Quechuan loans – we apply the CCNC method and statistical tests (§6).

Regarding the results of our experiments, we obtain very strong statistical evidence for hitherto unexplained recurrent elements within Proto-Quechuan roots, but no evidence for such elements within Proto-Aymaran roots or their English equivalents (§6.1). More precisely, we find that if the assignment of words to meanings were random, there would be a chance of less than 1 in 1,000 of observing a higher density of two-segment phonetic strings within semantically related Proto-Quechuan roots than we observe in the data ($p < 0.001$, a level normally considered highly statistically significant). Meanwhile, the odds of finding a more

1. The implementation is available at https://github.com/jdellert/ccnc.

extreme pattern in the Proto-Aymaran data would be 329 in 1,000 ($p = 0.329$), and for English, the odds would be 319 in 1,000 ($p = 0.319$). Both of these figures would commonly be interpreted as non-significant. In this way, we rule out chance as an explanation for the observed frequency of recurring two-segment phonetic substrings within semantically related Proto-Quechuan roots, while demonstrating that no similar effect is present in semantically related Proto-Aymaran roots or their English equivalents.

With chance eliminated as an explanation for the frequency of these sound-meaning correspondences in the Proto-Quechuan lexicon, it is next necessary to consider other explanations. One possibility is that the recurring phonetic substrings in semantically related roots do not, in fact, reflect archaic morphemes, but rather that they are phonesthemes. Phonesthemes are phonetic substrings that recur in semantically related roots, but which cannot be analyzed as separate morphemes (synchronically or historically); rather, the roots that contain them are formed through associative iconic influence. For example, John Firth observes semantic similarity among /tw/-initial English terms such as *twist, twirl, tweak, twill, tweed, tweezer, twiddle, twine* and *twinge* (1930: 186, cited in Kwon & Round 2015: 2); he claims that all these terms have to do with twisting. In these cases, the substring /tw/ is not a morpheme, because it cannot combine with other morphemes. The remnant phonetic sequences that follow that substring in each case are also not analyzable as morphemes.

It may well be the case that some of the terms we identify in Proto-Quechuan using the CCNC methodology contain phonesthemes. However, we believe that this explanation is inadequate to account for all of the of the sound-form correspondences that we observe in this paper, for three reasons. First, some of the subroot phonetic sequences do, in fact, co-occur with known Quechuan morphology, which makes their own status as morphemes more credible. For instance, the fact that the initial substrings in Table 1 (e.g., /tʂu/, /qa/, etc.) co-occur with phonetic material identical to the known Proto-Quechuan directional suffixes, and that they exhibit the corresponding directional semantics, suggests a morphologically productive process that cannot be simply dismissed as phonesthetic in the same manner as English terms like *twist, twirl*, etc. Second, a prominent characteristic of phonesthesia is that it often involves sound symbolism or at least sensory perception (like the /fl/ in *flash, flicker, flare*, etc., which refers to the visual experience of light (Kwon & Round 2015: 5)). The semantics of many roots identified by our methodology are not likely to be the basis of iconic associative influence (e.g., simple verbs of 'putting' or 'eating'). Third, CCNC identifies a strong pattern even on the small portion of the Proto-Quechuan lexicon for which we have reconstructible equivalents in Proto-Aymaran, whereas no effect at all is visible on an equivalent subset of English, despite the demonstrated prominence

of phonesthesia in that language. So, the pattern in Proto-Quechuan is demonstrably much stronger than we would expect from a relatively marginal process of root formation like phonesthesia. Also, as we discuss in §7.5, the total number of possibly historically polymorphemic roots in the PQ lexicon is likely even larger, since the CLICS² dataset is not an optimal framework for detecting the kind of semantic similarity we expect to find among Proto-Quechuan roots. For all of these reasons, we believe the roots we identify in this article are in fact just the tip of the iceberg, and that the presence of lexicalized archaic morphology is a widespread property of the Proto-Quechuan lexicon. At the same time, we acknowledge phonesthesia as a possible explanation for some forms, and we discuss relevant cases in §7.3.

While some of the adjoining phonetic material found alongside the recurrent two-segment phonetic strings is clearly identifiable (as in Table 1 above), one problem for our analysis is that we cannot explain all of it. For instance, in the /wa/-initial terms *warku- 'to hang up', *wata- 'to tie, repair' and *wanku- 'to wrap, bundle' in (1), we can interpret the remaining /rku/, but not the remaining /ta/ and /nku/. Ultimately, a full accounting of this phenomenon must eventually explain these phonetic remnants; it is likely that such remnants are due to diverse historical processes. However, we believe that this does not invalidate the patterns we identify in this first, exploratory identification of the phenomenon, particularly given the very strong statistical support we offer, and given that the whole phenomenon cannot be convincingly reduced to phonesthesia.

In §7, we go on to establish a list of proposed Proto-Quechuan clusters, drawing on a statistical analysis of Proto-Quechuan phonotactic patterns. Accounting for these Proto-Quechuan phonotactic patterns is important because some phonological processes appear to have applied to historically polymorphemic constructions, obscuring their original forms. We then propose reconstructions of nine archaic Quechuan roots on the basis of these patterns. We conclude in §8 by discussing the implications of our findings for Andean linguistic prehistory, and by outlining some new ways forward.

2. Relevance to Andean linguistic prehistory

As the foregoing discussion makes clear, our findings are relevant for understanding the early formation of the Quechuan lexicon. However, they also have implications for a different topic: the dynamics of early Quechuan-Aymaran language contact in the Central Andes. Explaining the striking lexical and structural resemblances between the Quechuan and Aymaran languages, which are spread across a vast and overlapping expanse of the Andean region, is one of the most enduring

and complex issues in South American historical linguistics. Between a quarter and a third of those languages' lexical material is either identical or nearly identical, and the languages exhibit striking typological similarities (explored in detail by Cerrón-Palomino 1994; Adelaar 2012a, 2017). Most observers before the 1950s and 1960s – from the 17th century Jesuit historian Bernabé Cobo (1890 [1653]) to the scholar-explorers of the 19th and early 20th centuries and beyond – assumed that these similarities were due to inheritance from a common ancestor language. This came to be known as the "Quechumaran hypothesis" (Mason 1950: 196–200; Orr and Longacre 1968). However, since the emergence of a rigorous historical-comparative tradition in Andean linguistics in the 1960s, most specialists in the region came to believe instead that the most obvious similarities between the families were the result of language contact (see Cerrón-Palomino 1987: 351–75; 2000: 298–337 for overviews of this development). Whether there exists a deeper genetic connection between the families, which can only be discerned by identifying more distant cognates that lie beyond those superficial similarities, remains an open question (Campbell 1995; Cerrón-Palomino 2000: 311–312).

While our understanding of the enigmatic Quechuan-Aymaran relationship has progressed steadily over the last several decades, much remains unknown. What has become clear most recently is that the contact effects which have emerged among those families are historically multilayered and geographically diverse. On the one hand, their mutual influence is already evident at the Proto-Quechuan and Proto-Aymaran stages, which means that we must posit even earlier stages before that period of contact, called Pre-Proto-Quechuan and Pre-Proto-Aymaran (e.g., Weber 1987: 35–48; Cerrón-Palomino 2000: 337; Adelaar 2012a). These stages are penultimate to the diversification of the clades from which we have data, and represent a time before the contact that transformed both the Quechuan and Aymaran lineages (or, at least, the earliest contact that we can currently detect; it is possible that there also exist earlier strata of contact between the lineages). On the other hand, subsequent periods of convergence have also taken place among Quechuan and Aymaran languages in various places after the diversification of those families. This situation requires that we make a distinction between the first period of Quechuan-Aymaran contact, before the respective proto-language stages – what Adelaar (2012b: 424, and elsewhere) calls the "initial convergence" – and more recent "local convergences" (ibid.) among their various daughter languages. It is not clear when the initial convergence took place, but it likely wasn't long before the families each spread across the region. Impressionistically, the internal variation in each family suggests that this probably wasn't more than a millennium or two before present (Heggarty & Beresford-Jones 2010: 166). The initial convergence surely took place after the development of agropastoralism in the Andes (3,500–5,500 years before present), since each

linguistic lineage already had separate, well developed lexical inventories for the practices, animals, cultivars, tools, and built structures associated with agropastoralism (Emlen & Adelaar 2017).

Recent progress on the Quechuan-Aymaran problem has made it possible to address two questions that loom over Andean historical linguistics today. First, what happened during the initial convergence? Second, what might Pre-Proto-Quechuan and Pre-Proto-Aymaran have been like before that period of convergence? These questions are important for our understanding of whether the two pre-proto-languages are ultimately related to each other, or to other languages in the region (which might be a more promising path forward at this time depth; see e.g., Adelaar 1986: 380; 2013b). In the last decade, several advances have been made in this respect. To begin with, we now know that the contact effects between Pre-Proto-Aymaran and Pre-Proto-Quechuan were quite asymmetrical. On the one hand, Adelaar (1986) argued that the direction of lexical borrowing was largely from Quechuan to Aymaran. This was confirmed by Emlen (2017), who, following and expanding upon Adelaar's methodology, found that the lexical items shared by both proto-languages have Quechuan phonological and phonotactic characteristics. On the other hand, the grammatical structure and perhaps also the phonology of Pre-Proto-Quechuan seem to have been reformatted on the Aymaran template. The latter process apparently led to a high degree of isomorphism in the languages' grammatical systems (Cerrón-Palomino 1994; Adelaar & Muysken 2004: 36; Adelaar 2012b; Muysken 2012), while the forms of the grammatical morphemes themselves remained mostly distinct.

Evidence for this process comes from a number of observations. Muysken (2012) observes that Aymaran languages exhibit great complexity in their idiosyncratic morphophonemic vowel deletion rules, in their phonemic inventories, and in their derivational morphologies, as well as notable irregularity in their person and tense marking systems. The corresponding features in Proto-Quechuan are patterned similarly to those Proto-Aymaran ones, but they are simpler and more regular, and some Quechuan suffixes are transparently built up from component morphemes while the corresponding Aymaran ones are not. To take another example, Proto-Aymaran had historically unrelated paradigms for its verbal and nominal person systems; Proto-Quechuan also had separate verbal and nominal person systems, but these can apparently be reconstructed to a single system (Cerrón-Palomino 1987: 137–144; Muysken 2012). In other words, both lineages have distinct verbal and nominal person paradigms, but the Quechuan lineage innovated this distinction relatively recently, while in the Aymaran lineage these paradigms came from distinct sources. According to both Muysken (2012) and Adelaar (2010: 242; 2012b: 462), such patterns suggest Aymaran as the template

for the remodeling of Quechuan structure. It may be possible to make the reverse argument for some cases, but so far it has not been attempted in the literature.

Adelaar (2012b: 463) proposes a specific dynamic for this restructuring. He writes that "Quechuan may have adopted an Aymaran model by reassigning elements from its own original morphemic inventory to borrowed functions". Indeed, several aspects of Proto-Quechuan morphology show signs of having been built up from smaller bits in ways that made Quechuan structure more Aymaran-like (for the earliest known version of this observation, see Uhle [1910] 1967: 48–49). For instance, the introduction of inverse markers in Quechuan allowed subject morphemes to be recruited for object reference (Adelaar 2009), an innovation that made the person system more similar to the Aymaran one. Similarly, some of the Proto-Quechuan tense and person marking system was assembled from a smaller set of morphemes, which can also be reconstructed (Adelaar 2011). It is not possible to reconstruct such processes in Proto-Aymaran. Finally, most relevant to this discussion, Cerrón-Palomino (1987: 191) and Muysken (2012) cite the presence of fossilized, archaic monosyllabic roots within Proto-Quechuan lexemes as evidence of the innovative character of that language's typological profile; monosyllabic roots are nearly absent in Proto-Aymaran and Proto-Quechuan, but were apparently more common earlier in the Quechuan lineage. The strong preference for minimally bisyllabic roots in the Proto-Quechuan lexicon may have emerged as a result of Aymaran influence.

This is the current state of thinking about the initial convergence: the Quechuan and Aymaran lineages arrived at a striking degree of structural isomorphism through notably different historical trajectories. However, while this general pattern has come into clearer focus in the last decade, it has not yet been possible to say much more about the nature of the Quechuan lexicon before the initial convergence, nor about the specific dynamics of that convergence. The findings presented in this paper support the hypothesis that some Quechuan morphemes have a historically polymorphemic origin, which is consistent with the Aymaran convergence hypothesis described above. These patterns are thus a step forward in our understanding of both the Pre-Proto-Quechuan lexicon and the initial convergence. They are also helpful for considering the Quechumaran hypothesis, because if we are correct that the presence of lexicalized sub-root elements is in fact a widespread property of the Proto-Quechuan lexicon, then the Proto-Quechuan lexicon and Proto-Aymaran lexicon (each purged of loans) would be most productively compared to each other after a broader analysis of such sub-root elements. Indeed, if anything is clear from our analysis, it is that we cannot be sure just how much historical morphological complexity lies below the surface of the Proto-Quechuan lexicon. This should urge caution during a

search for Proto-Quechuan's external relatives, including a comparison with the Aymaran lineage.

3. Crosslinguistic colexification network clustering (CCNC)

Chance similarity is an enduring problem in historical linguistics. As innumerable attempts at establishing macro-language families have shown (e.g., Nostratic, Salmons & Joseph 1998), for any pair of unrelated languages it is often easy to find many similar-looking word pairs (in some cases even with apparently regular sound correspondences) if we are generous enough with the meanings we allow to enter into such comparisons. Campbell & Poser (2008) provide a comprehensive discussion of this problem, showing for a range of examples that the degree of semantic leeway employed by proponents of macro-families would likely suffice to 'prove' the deep relationship of any pair of languages.

Historical linguists who work on individual language families take such risks into consideration when positing etymologies involving semantic change. A common means of evaluating the semantic plausibility of such etymologies is to compare them with similar semantic developments in other languages. This practice limits the combinatorial possibilities and thereby reduces the risk of chance similarity. However, it does not necessarily lead to a neutral standard of semantic similarity that both proponents and opponents of a given theory can agree on. In particular, even if it were required to justify each posited semantic shift by citing a parallel development in another language, the unpredictable nature of semantic change might still leave too many comparanda to choose from. This can make it difficult to avoid bias if the evidence of semantic relatedness is compiled by the same person who then uses it in service of their own hypothesis.

Claims of non-random structural patterns, which risk being based on researcher bias due to the possibility of chance phonetic overlap, should be reinforced through objective statistical tests with transparent baselines. This is possible whenever we can summarize the strength of our evidence in a number (statistic) that can be computed from the data, and can estimate the distribution of the statistic under the null hypothesis, i.e., on data as it would have been produced if the generating process did not exhibit the property we are testing for. For the problem of determining deep common ancestry, the building blocks for such tests have long been developed. The idea of rigorously quantifying the risk of chance similarities in multi-way comparisons of phoneme sequences was pioneered by Ringe (1992). Estimating the distribution of similarities under the null hypothesis by re-sampling from existing language data goes back to Kessler (2001). More recently, similar approaches have been used to assess questions of possible deep

phylogenetic relationships e.g. in the Central Solomons (Dunn & Terrill 2012) and in California (Haynie 2014). In order to prove an above-chance correlation between shared phonetic substrings and semantic similarity, as we do in this paper, we need to apply the ideas behind these works on an explicitly modeled neutral standard of semantic similarity. Statistical tests based on such a standard will be widely applicable to any similar question about possible non-random correspondences between form and meaning, which arise frequently in historical linguistics. For instance, the existing statistical tests of deep language relationship could become more sensitive if we allow a neutral standard of semantic similarity to guide the possible choices of comparanda.

In order to make cross-linguistic evidence of semantic similarity useful for statistical arguments, we need a mathematical structure which summarizes at least a part of our knowledge about which meanings tend to be expressed in similar ways, and which ones do not. While not an ideal fit, a good candidate for such a structure that builds on cross-linguistic regularities is synchronic polysemy – that is, the use of one word for more than one concept. A classic example of such a polysemy is the concept pair HAND and ARM, which in many languages are referred to by the same term. This relation between the concepts has been called "colexification" in the literature (François 2008). On the basis of this cross-linguistic evidence from the languages of the world, it is possible to make an empirically founded assertion that HAND and ARM are semantically related. Recent research on the regularities of semantic change (Steiner et al. 2011; Zalizniak et al. 2012; List et al. 2013, 2018; Münch & Dellert 2015) has established the possibility of arriving at a workable neutral standard for semantic comparison from the systematic compilation of such cross-linguistic polysemies.

The CLICS² cross-linguistic co-lexicalization database (List et al. 2018) is by far the most comprehensive effort so far to develop such a polysemy network. It provides data on frequent colexifications from more than 1,200 languages, with data available for more than 2,000 basic concepts. For instance, the database includes the information that the concept HAND has well-attested connections not only to ARM, but also to BRANCH, WING and FIVE; we could use this cluster of attested colexifications as an external and independently motivated guide towards comparanda in the search for distant cognates of a word for HAND. The CLICS² database can be used for purposes like ours by mapping the glosses from our dataset to standardized concept IDs represented in the database. In our case, this meant going through the reconstructed Proto-Quechuan and Proto-Aymaran lexical material and assigning the lexical items (like Proto-Quechuan *wata- 'to tie, repair' in (1) above) with CLICS² concept terms (like TIE). This process is explained in greater detail in §5.

Building on this kind of independent standard of semantic similarity encoded as a network, we can develop a general and automatable procedure for finding patterns which deviate, in some relevant feature, from a completely random mapping between meaning and form. Depending on the situation, we might want to investigate vowel alternations, shared consonant patterns, similarities in tone or, as in our case, shared sequences of phonetic segments. In every case, the patterns can be detected by clustering together words with closely related meanings (neighboring concepts in the polysemy network) that share the feature of interest. The general idea of statistical testing for non-random form-meaning correspondences has previously been explored in a different context by Blasi et al. (2016). Whereas we exploit semantic similarity to cluster across concepts in a search for non-random patterns in a single (proto-)language, they repeatedly test for patterns in the presence or absence of specific sounds in the words for individual concepts across many languages.

We now describe the core building blocks of the CCNC method. The procedure takes as input: (1) a mapping from words to a curated list of cross-linguistic concepts (the lexical data); (2) a network connecting those concepts (in this case, CLICS^2 reduced to the concepts for which data are available); (3) an arbitrary phonetic feature extraction function. The procedure returns clusters of semantically related words which share one of the extracted features. Several design decisions have to be made when implementing such a procedure, but we will still adopt the rather general term Crosslinguistic Colexification Network Clustering (CCNC) for the variant described in this article. The first core design decision is to only extract clusters which are centered on a single concept (i.e., WARM ⇔ HOT ⇔ SUN would form a cluster centered around HOT, whereas the concepts WEATHER ⇔ DAY ⇔ SUN ⇔ BRIGHT do not have a center connecting all of the concepts). This implies a restriction to patterns which can be described as a set of simultaneous developments among concepts which are immediate neighbors in the colexification network, without having to assume multiple stages of semantic developments which would have to develop over longer periods of time. Additional decisions we made were to extract clusters by decreasing size (i.e., the largest remaining cluster first), and to let each word participate in only one cluster.

In our implementation, these properties are enforced in a straightforward, albeit not optimally efficient manner. The algorithm proceeds in rounds, during each of which it extracts a single cluster. In each round, the algorithm visits every concept which still features unassigned words. For each concept, it collects all available words for the concept itself as well as its immediate neighbors in the network, builds a map from the relevant features to these words, determines the feature shared by the maximum number of these neighboring words, and remembers the resulting cluster in case it is larger than all clusters previously found in the cur-

rent round. After visiting all the concepts in this way, the maximum cluster of the round is added to the output, and the words in it are marked as already assigned to a cluster. The algorithm then proceeds to the next round, and continues in this manner until no new clusters are found.

4. Statistical test for non-random form/meaning correspondence

As in any quantitative hypothesis test, we need a test statistic (a meaningful summary of the data as a single number) with a known derivation under the null hypothesis. This allows us to quantify as a p-value the probability that we would see a value more extreme than the value on the actual data if the null hypothesis were true. We can then reject the null hypothesis if the p-value is lower than a previously stipulated significance level (usually 0.05). In our experiments, the null hypothesis will invariably be that there is no connection between semantic similarity and phonetic substring overlap in our dataset.

The purpose of our test statistic is to quantify the degree of semantic clustering of words displaying an overlap in the relevant phonetic properties. We chose a rather straightforward option by simply counting the number of word pairs connected by such clusters in the CCNC output. For instance, if we count every possible pair of terms within a cluster of five terms, we end up with a total of $\binom{5}{2} = 10$ pairings. A cluster of size two only connects a single pair of words – that is, $\binom{2}{2} = 1$ – which means that a cluster of size five would count as much as ten clusters of size 2. This is a natural definition from a mathematical point of view, while at the same time it reflects our intuitive judgments about how conspicuous we would perceive certain clustering scenarios to be. For instance, assume that our procedure finds one cluster of size 4, three clusters of size 3 and five clusters of size 2 in the data. This would result in a value of $1 \times \binom{4}{2} + 3 \times \binom{3}{2} + 5 \times \binom{2}{2} = 1 \times 6 + 3 \times 3 + 5 \times 1 = 20$ for the clustering statistic. If the largest cluster were reduced by one word (becoming a fourth cluster of size 3), the value of the statistic would drop to $4 \times 3 + 5 \times 1 = 17$. That is, we would need two additional clusters of size 3 to compensate for the loss of a cluster of size 4.

To estimate the distribution of this test statistic under different hypotheses, we adopted the well-established technique of creating pseudo-datasets of the same shape as the original data by randomly recombining the existing data in such a way that the null hypothesis can be expected to hold. In our case, we generated 1,000 pseudo-datasets by randomly shuffling the mapping between words and CLICS² concepts, making sure that the number of words assigned to each concept remains equal to the same number in the original data. This amounts to a very conservative estimate of the possible variation, as it trivially keeps the phoneme

frequencies and phonotactic patterns identical to the original data, which would be more difficult to do in a model which attempted to create plausible pseudo-datasets by generating new word forms. Due to the nature of our data, our application does not suffer from one of the most common problems of shuffling-based approaches, namely an underestimation of the expected similarity if words from different classes are shuffled despite the presence of category-dependent morphological material such as nominative or infinitive endings, or differences in root structure due to e.g., stress patterns. This will not be an issue for our data, as all forms are uninflected roots, and as we will see, slight phonotactic differences between word classes due to the existence of root-final consonants in non-verbal roots are not strong enough to make the original data recognizably different from a randomized mapping.

5. Preparing the Proto-Quechuan and Proto-Aymaran data for CCNC

The empirical basis of this paper is a large Quechuan and Aymaran lexical database. The database was used in an earlier paper (Emlen 2017) to reconstruct Proto-Quechuan and Proto-Aymaran lexical material, to disentangle the multi-layered history of lexical borrowing between the Quechuan and Aymaran lineages, to assign provenances to several hundred lexical roots, and to draw conclusions about the phonology of Pre-Proto-Aymaran. That analysis focused on the early Aymaran lineage, and this paper turns to the early Quechuan lineage.

The Quechuan data comprise more than 11,000 roots collected from 16 dictionaries and wordlists from across the family, chosen to cover as wide a genealogical and geographical range as possible. The Aymaran data include more than 10,000 roots from 10 sources across that family. The sources of data for the Proto-Quechuan and Proto-Aymaran reconstructions are listed in Emlen (2017). In addition, one more Quechuan source has been added to the corpus since the 2017 article: the Bolivian Quechua dictionary by Laime Ajacopa et al. (2007). Adelaar (2006) was also consulted for the Tarma Quechua data presented this paper.

The data described in Emlen (2017) were used for this paper to reconstruct 809 Proto-Quechuan roots and 259 Proto-Aymaran roots.[2] Roots were reconstructed in Proto-Quechuan (1) if they exhibited the sound correspondences known to have developed during the evolution of that family (Cerrón-Palomino 1987); and (2) if they were attested in at least one Quechuan variety of Central/

2. All the data used in this paper are available digitally, formatted as TSV files for input to our CCNC implementation, at https://github.com/jdellert/ccnc/que-aym-data. The relevant code can be found in the same repository.

Northern Peru (either Quechua I or the conservative Pacaraos, Cajamarca or Yauyos varieties), and in a variety of either Quechua IIB or IIC. This is the sharpest genealogical distinction in the family (Parker 1963; Torero 1964; Adelaar 2013a), and roots that are attested on both sides, all things being equal, likely descend from Proto-Quechuan (though there may be some exceptions due to later borrowing).

On the Aymaran side, roots were reconstructed (1) if they exhibited the sound correspondences known to have developed during the evolution of that family (Cerrón-Palomino 2000); and (2) if they were attested in both of the two extant branches of the family: Central Aymaran and Southern Aymaran. The Proto-Aymaran roots that show evidence of having originated in the Quechuan lineage (Emlen 2017) were then eliminated, leaving 259 Aymaran roots that likely descend from Pre-Proto-Aymaran. Note that all forms used in this article come from the database, and will not be cited by source.

A few of the Proto-Quechuan and Proto-Aymaran roots presented in Emlen (2017) have been eliminated from this analysis. Three of these are in response to Cerrón-Palomino (Forthcoming: 115–116), to whom we are grateful for his care and attention. He suggests that *kurku 'hunchback, hunch' is likely a loan of Spanish *corcova* 'hump', which is indeed plausible, if not certain. The term *wiraquča 'Andean deity' has been eliminated from the Proto-Quechuan list, and *q^hapaqa 'powerful, rich' from the Proto-Aymaran list, because of their association with the Inka period. Regarding the terms for which Cerrón-Palomino asserts a Puquina origin, we are not yet convinced of this claim, and prefer to leave them in the lists until more is understood about the history of the Puquina lexicon.[3] As for reconstructed forms that may be morphologically complex, we only regard this to be a relevant consideration when all of the putative morphemes are identifiable; indeed, the argument made in this paper is that many reconstructed roots are, in fact, historically morphologically complex. Note that none of the roots cited in Cerrón-Palomino's response figure into the analysis offered in this paper.

These Proto-Quechuan and Proto-Aymaran reconstructions were then mapped to the list of CLICS² concepts. This was a relatively straightforward process, though in some cases the granularity of distinctions in the CLICS² concept list was insufficient to capture the semantic nuances of the Proto-Quechuan

3. Note, for instance, that several of the terms of supposed Puquina origin listed by Cerrón-Palomino are almost universally attested in the Quechuan and Aymaran languages, but do not appear at all in the one surviving Puquina source (Oré 1607). The justification for their origin in Puquina is not based on linguistic evidence, but rather is deduced from ethnohistorical premises that require further scrutiny. More work on the Puquina lexicon is forthcoming from the Leiden Puquina Working Group (e.g., Mossel et al. 2020).

lexicon, and the same CLICS² concept had to be used for more than one lexical item. For instance, *wikʎu 'twisted, deformed', *wiksu 'twisted, cross-eyed, bow-legged', *wiqru 'lame, having an injured foot, bow-legged, twisted' and *wištu- 'twisted, crippled, to hobble, limp' were all assigned the CLICS² term TWIST. At the same time, some terms were too specialized to be mappable to a CLICS² concept (for example, there were no obvious concepts for Proto-Quechuan *tʂaʎpu- 'to submerge, immerse' and *tampa 'tangled, disheveled, unkempt'). These were not mapped, and thus were excluded from the analysis. Since the inclusion of variants would inflate the clusters, these variants were eliminated (for instance, for Proto-Quechuan *atʂpi ~ ašpi ~ aspi 'to dig, scratch', only *atʂpi- was used). In such cases, we chose the variants with the broadest distribution among the families' varieties, and those that followed known regularities in phonological change (i.e., *atʂpi- in this example, since lenition is more common than fortition). Mapping the Proto-Quechuan and Proto-Aymaran lexicons to the CLICS² concept list did not present any major difficulties, though it is, of course, impossible to arrive a completely objective mapping. Note also that we omitted from our experiments Proto-Quechuan terms that Andeanists have already identified as historically polymorphemic in the literature, since our argument will be most convincing if it is shown to go beyond what has already been said on the matter.

6. Analyzing the Proto-Quechuan lexicon using clustering significance tests

The CCNC algorithm identified groups of Proto-Quechuan lexemes that share a two-segment phonetic substring, and that appear together in a CLICS² semantic cluster. For instance, the roots *wata- 'to tie, repair' (CLICS² concept TIE), *waska 'rope' (CLICS² concept ROPE) and *wanku- 'to wrap, bundle, bandage' (CLICS² concept BUNDLE) are identified by the method because they share the two-segment substring /wa/, and because those three concepts are each linked by a single degree in the CLICS² network (and are thus considered semantically related by this neutral standard).

In order to validate the method, we ran exactly the same algorithm on the reconstructed Proto-Aymaran lexicon, expecting to find that semantic clusters sharing two-segment substrings are not more common in that language than we would expect by chance. We derive this by estimating the distribution of the test statistic from 1,000 samples of pseudo-data (that is, 1,000 lists in which the Proto-Aymaran phonological forms are reshuffled and randomly reassigned to the same set of CLICS² concepts). In order to achieve exactly comparable results, we reduce

both the Proto-Quechuan and the Proto-Aymaran data to those CLICS² concepts for which we have reconstructed roots in both proto-languages. This serves to exclude any possible effect of, e.g., different network densities for different concept samples. For instance, if we used the full datasets for both languages, and one set included more verbs than the other, this could lead to a denser network and therefore more comparanda for the language in question, because verbs tend to be more polysemous than members of other word classes (Gentner 1981). This intersection contains 147 concepts connected by 88 links. The connections among the 81 of these concepts that have any connection to another concept within the set is visualized in Figure 1.

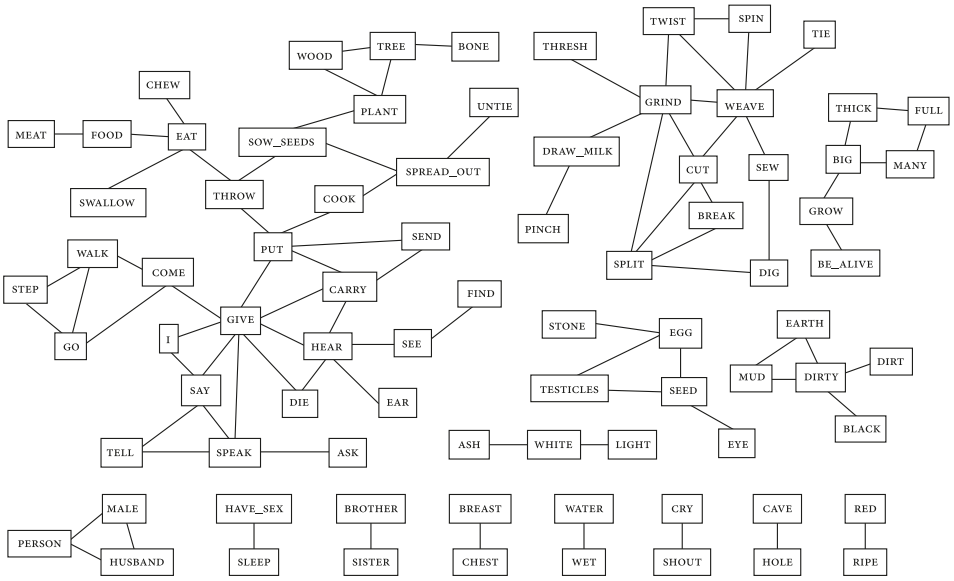

Figure 1. CLICS² subnetwork over the concepts with reconstructed roots in both Proto-Quechuan and Proto-Aymaran (66 additional concepts without any link not visible)

In order to statistically test our hypothesis, all we have to do is to apply CCNC to two different datasets. In the first experiment (§6.1), we perform tests on the Proto-Quechuan and on the Proto-Aymaran data in order to check whether, in either dataset, forms with shared two-segment substrings form semantic clusters more frequently than we would expect by chance. In the second experiment (§6.2), we then analyze the larger Proto-Quechuan dataset to test whether the shared segments occur word-initially.

6.1 Test 1: Frequency of recurrent two-segment substrings within semantic clusters

Running CCNC on the actual Proto-Quechuan data, with the criterion that clusters need to share a two-segment substring at any position, we arrive at the clusters visualized in orange in Figure 2. Note that some clusters only cover a single CLICS² concept – this means that there were multiple words mapped to the given concept, which share the phonetic substring in question. For this dataset, our test statistic reaches a value of 60. That is, there are 60 pairs of Proto-Quechuan words which share a two-segment substring and belong to the same semantic cluster.

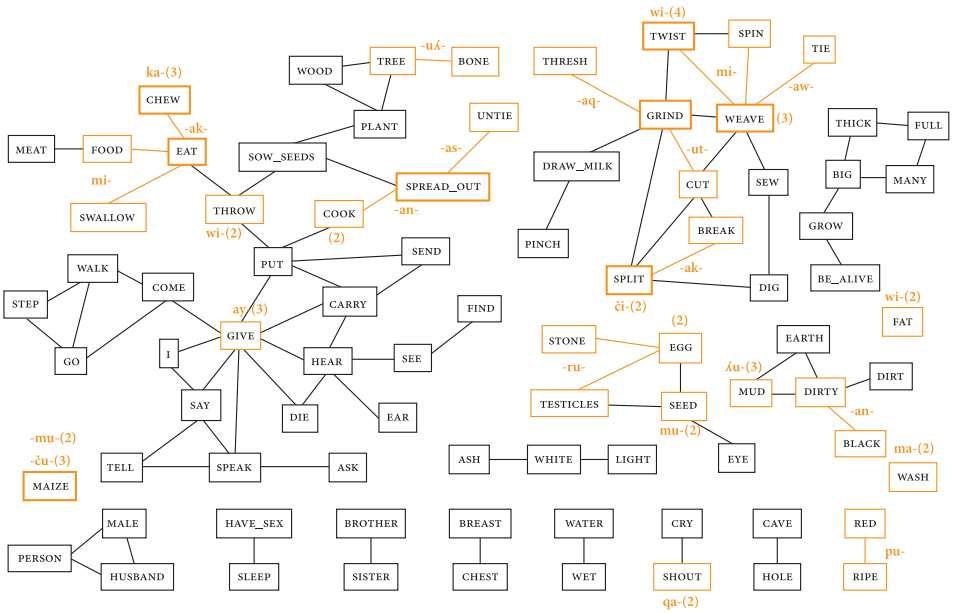

Figure 2. Clusters found by CCNC on the Proto-Quechuan data over the subnetwork, annotated with shared substrings and the number of roots mapped to the same concept (if any)

By contrast, the same procedure on the Proto-Aymaran data yields far fewer clusters, all of which are visualized in the same fashion in Figure 3. The value of the clustering statistic on the Proto-Aymaran dataset is 13.

The real question, of course, is whether either of those values significantly deviates from the values that we would expect if the relationship between forms and concepts were completely random. We cannot compare both values to a shared distribution: the different phonotactic systems of the two reconstructed languages can lead to spurious shared substrings being more or less likely. Furthermore, due to the larger dataset, more Quechuan roots tend to be available

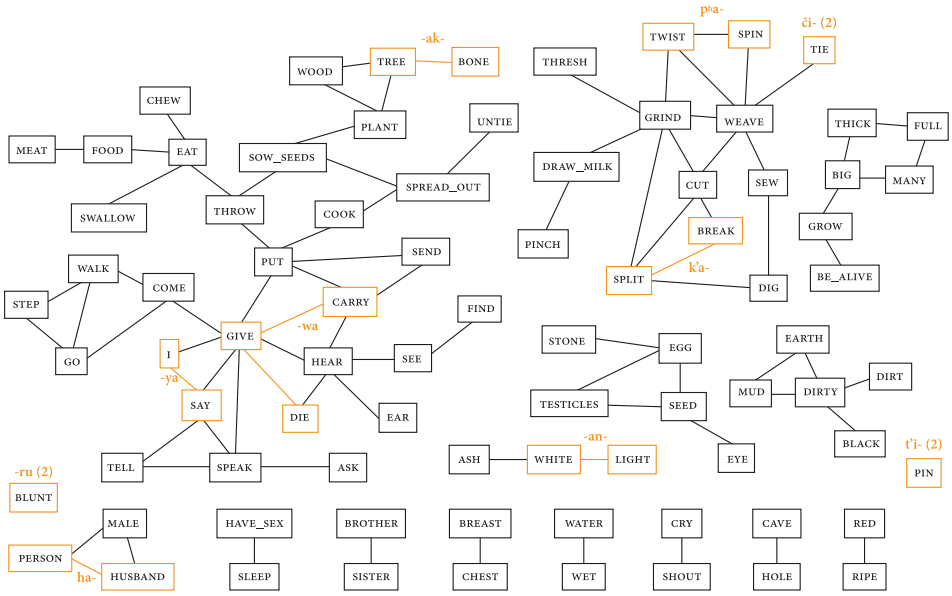

Figure 3. Clusters found by CCNC on the Proto-Aymaran data over the subnetwork, annotated with shared substrings and the number of roots mapped to the same concept (if any)

for a single concept than Aymaran equivalents. Instead, we have to shuffle both datasets separately in order find out which values of the clustering statistic we would expect in random assignments. The distributions of the clustering statistic across 1,000 pseudo-datasets are given in Figure 4 (for Proto-Quechuan) and Figure 5 (for Proto-Aymaran), with the observed values highlighted in yellow.

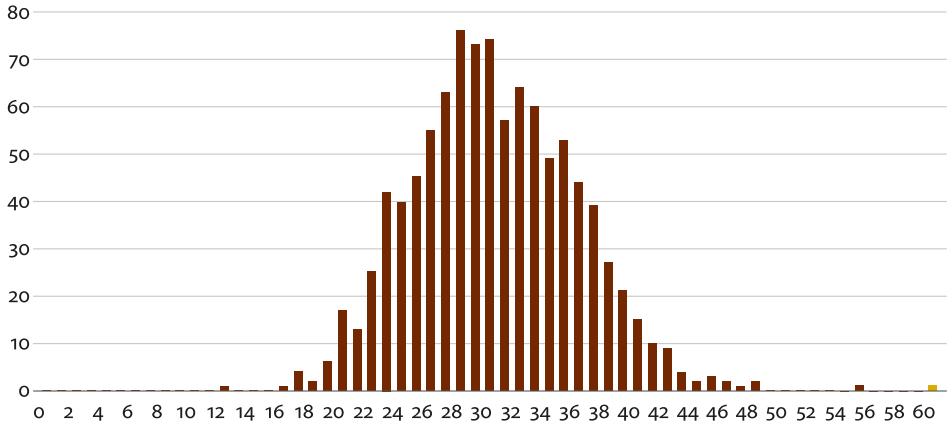

Figure 4. Distribution of the clustering statistic under the null hypothesis for Proto-Quechuan data (actual value marked in yellow at far right)

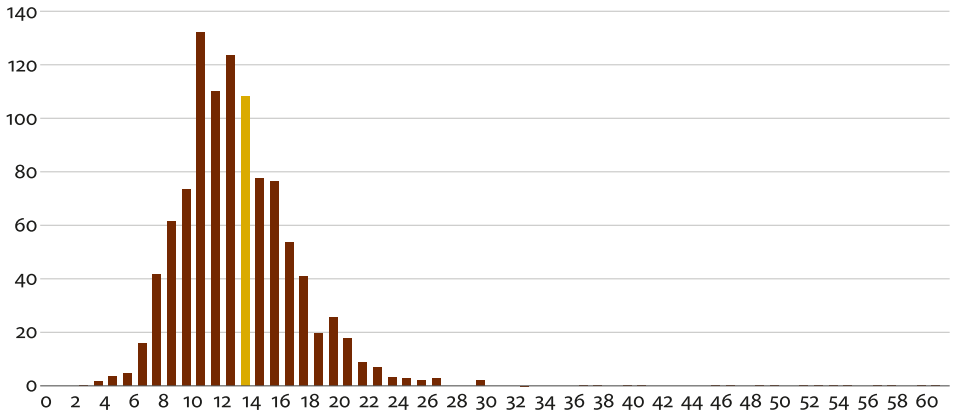

Figure 5. Distribution of the clustering statistic under the null hypothesis for Proto-Aymaran data (actual value marked in yellow)

It turns out that in 1,000 random resamples of the Proto-Quechuan data, we do not reach a single value that is greater than the value of 60 we received for the real data (Figure 4). This corresponds to $p < 0.001$ (at a z-score of 5.362), which is commonly read as extremely significant. By contrast, around a third of the Proto-Aymaran resamples led to higher values of the statistic than the observed 13, leading to a value of $p = 0.329$ (Figure 5). In other words, in a ranking of the values resulting from 1,000 pseudo-Proto-Aymaran reshufflings, the real Proto-Aymaran dataset would appear in position 329. This indicates that the observed value is well within the range expected under the null hypothesis – that is, it appears random. We have thus established that a clustering pattern as dense as we find in the Proto-Quechuan data would virtually never occur if there were no unexplained connections between form and meaning in the data. At the same time, the relationship between phonological form and concepts in the reconstructed Proto-Aymaran lexicon looks completely random, which means that the method does not suggest the presence of meaningful sub-root elements in that dataset.

As an additional validation, we executed the same procedure on English as well. English is known to exhibit phonesthesia, and its history is known well enough to rule out the presence of fossilized proto-morphemes. These characteristics make it a good additional point of comparison. On very similar data to the previously employed intersection of concepts reconstructible for Proto-Quechuan and Proto-Aymaran,[4] we arrive at a p-value of 0.319, which is very close

4. The only difference in the selection of concepts is that we had to exclude the concepts BROTHER IN LAW and GRINDSTONE due to the absence of monomorphemic equivalents in English. For each of the remaining 145 concepts, we picked a range of common near-synonyms to emulate the process of mapping the two proto-languages to CLICS^2 concepts. We lack the space

to what we observed for Proto-Aymaran. This reinforces our finding that Proto-Quechuan exhibits a highly divergent pattern which bears investigating.

6.2 Test 2: Position of meaningful substrings within Proto-Quechuan roots

Having established in our first test that this pattern in the Proto-Quechuan lexicon is not attributable to chance, we performed a further test on the larger dataset in order to learn where in the roots we expect to find the relevant recurrent elements.

Because it is not intersected with the Proto-Aymaran subnetwork, the CLICS^2 subnetwork for this Proto-Quechuan experiment is much larger than for the previous one, at 535 concepts, 369 of which are connected by 588 links (the rest being singletons). Again using 1,000 bootstrap samples generated by shuffling the assignment of words to CLICS^2 concepts, we tested two hypotheses by investigating distributions of the clustering measure on CCNC outputs. For the first hypothesis, we only clustered semantically related words sharing the same two-segment root-INITIAL substring, e.g., *waska 'rope' and *wata- 'to tie', and tested whether such root-initial overlaps occur significantly more often in semantically related roots than expected by chance. For the second hypothesis, we repeated the same test for two-segment root-FINAL clustering e.g., *čapra 'branch' and *waqra 'horn'. The clustering statistic for the root-initial substrings on the real data was 92, whereas for the root-final substrings it was 74. This might not seem like a large difference, but due to the different phonotactic possibilities in different positions in the word, there was again not a single value that was this high for the root-initial substrings on the bootstrap samples ($p < 0.001$, at a z-score of 5.752). On the other hand, 101 of 1,000 bootstrap samples led to higher values of the clustering statistic in root-final substrings than the real data ($p = 0.101$, which is non-significant). This means that the significant non-random signal in Proto-Quechuan reconstructions is concentrated in the initial segments of words, whereas the final elements of the stem appear random. The root-initial shared elements detected by CCNC provide quite a bit of material for further linguistic analysis.

7. Linguistic analysis of root-initial elements

Application of CCNC to the Proto-Quechuan data identified 130 roots that appear in the same CLICS^2 clusters of semantically-related roots and share initial two-

here to include our full English dataset, but we provide it together with the code at https://github.com/jdellert/ccnc.

segment substrings. (Note that this does not include the eight roots in the Proto-Quechuan list that were already identified as historically morphologically complex in the Quechuanist literature; those were omitted from the experiments reported here.) The algorithm found three clusters of four semantically related words; ten clusters of three semantically related words; and 44 clusters of two semantically related words. Together, the 130 lexical items in these clusters comprise 16.1% of the 809 total reconstructed Proto-Quechuan lexical items (which rises to 17.1% when the eight omitted roots are reintroduced) and 8.9% of the mapped CLICS^2 concepts. As we showed in our statistical analysis in §6.2, the probability of obtaining this density by chance is less than 1 in 1,000 ($p < 0.001$), which is commonly accepted grounds for discarding chance as an explanation. With chance ruled out, we now turn to other possible explanations for the patterns identified by the CCNC methodology. The two remaining possibilities are that these contain phonesthemes, and that they represent fossilized Pre-Proto-Quechuan morphology.

To review our discussion from the introduction, phonesthemes are phonetic substrings shared among semantically related lexical roots, such as /tw/-initial English terms like *twist, tweak*, etc., or /gl/-initial terms like *glimmer, glint, glisten*, etc. In these terms, /tw/ and /gl/ (as well as the remaining phonetic material that follows them) are not morphemes, because they do not combine with other morphemes. Instead, they were formed through associative iconic influence. Thus, reconstructing /tw/ and /gl/ as morphemes in an earlier period of English on this basis would be an error.

As we discussed in the introduction to this article, we believe it is possible that there are a few phonesthemes in Proto-Quechuan lexicon (see §7.3), but that this does not fully explain the phenomenon we have identified. First, several of the sub-root phonetic sequences identified by the CCNC methodology do, in fact, co-occur with known Quechuan suffixes (as in the case of *warku- 'to hang up'), which bolsters their case as genuine Pre-Proto-Quechuan morphemes. Second, many of the terms presented here simply do not share the kind of semantics that we would expect to be generated through sound symbolic or iconic associative processes. Phonesthemes often involve sound symbolism or relate more generally to sensory experience; it is hard to see why connections among simple terms for different kinds of putting, eating or spreading out fabric (for instance) would be best explained by this kind of iconic process. (Some of the roots in our analysis, on the other hand, are consistent with such semantics, as we discuss in §7.3.) Third, words formed through phonesthesis (e.g., *glimmer, glint, glisten*) tend to be relatively marginal components of a given language's lexicon, and would not likely create the magnitude of the effect that we observe in the Proto-Quechuan data. For instance, although English is known to exhibit widespread phonesthe-

sia, the degree of sound-meaning correspondence in that language is nowhere near pervasive enough to register a statistical effect like the one we observe in Proto-Quechuan (as we showed in §6.1). It is possible that phonesthesia is just vastly more prevalent in the Proto-Quechuan lexicon, but we do not see a reason to suspect that this is the case, given the well-established pervasiveness of sound symbolism in English (e.g., Blake 2017). Finally, as we discuss in §7.5, we suspect that the imperfect fit between the CLICS² dataset and this research question in fact leads to a substantial undercount of the number of roots partaking in this pattern. For these reasons, we believe that while phonesthesia may have been operative in some cases, it is not a convincing explanation for the pattern in general. This leaves us with fossilized morphology as the best explanation for most of these terms, and we indeed believe that this is a widespread property of the Proto-Quechuan lexicon.

In this section, we discuss the cases identified by CCNC one by one. However, identifying the specific Pre-Proto-Quechuan morphemes that might be lexicalized in these Proto-Quechuan roots is a complicated matter, for two reasons. First, we expect some number of spurious clusters even in the genuine Proto-Quechuan data, so these need to be evaluated case by case for plausibility. Second, there appear to have been phonological changes resulting from the phonotactic constraints of Proto-Quechuan, which need to be understood before we can reconstruct the exact form of the archaic Pre-Proto-Quechuan morphemes. We first turn to the issue of phonotactics.

7.1 Proto-Quechuan phonotactics

Some archaic Pre-Proto-Quechuan morphemes appear to have undergone phonological changes during their lexicalization within Proto-Quechuan roots, which has obscured their original forms. Two relevant phonotactic constraints, which may have caused these changes, bear mentioning. First, certain consonants rarely co-occur in the same Proto-Quechuan roots. For instance, *k and *q are two of the most common consonants in Proto-Quechuan, appearing in 225 (27.8%) and 186 (23.0%) of the 809 roots, respectively. If there were no restrictions on the co-occurrence of *k and *q, the chance that some roots would exhibit both consonants would be very high. However, there is only one Proto-Quechuan root that includes both consonants. Notably, this root is *qayku 'to lead indoors, drive into a corral', which appears to be formed from a Pre-Proto-Quechuan root *qa(ti)- 'to herd, move' and the Proto-Quechuan inward motion suffix *-yku (see below). A possible explanation for why we don't see more roots that include both *k and *q is suggested by Pacaraos Quechua, in which the reflex of this Proto-Quechuan root is *qayqu-*, indicating a process of consonant harmony (Adelaar 2006: 130).

Because *k and *q are two of the most common phonemes in the Proto-Quechuan lexicon, such consonant harmony may have affected a great many other Proto-Quechuan roots (if indeed many Proto-Quechuan roots are historically polymorphemic). Similar constraints appear to be at play in the co-occurrence of the affricates *č and *tṣ, and the sibilants *s and *š.

A second phonotactic constraint has to do with consonant adjacencies. Despite the fact that 46.6% of our reconstructed Proto-Quechuan roots feature adjacent consonants, many particular consonant combinations are entirely absent (e.g., sequences of the same consonant) or extremely rare (e.g., any combination of affricates and sibilants). This means that the forms of Pre-Proto-Quechuan morphemes may have undergone phonological changes at syllable boundaries (for instance, the deletion or modification of one of the two consonants) which obscures the historically polymorphemic Proto-Quechuan root's original form. Another possibility is that that the combination of Pre-Proto-Quechuan morphemes simply did not bring these consonants together in the first place, but this does not seem likely given the distribution of those consonants.

One phonological process that appears to have transformed the Proto-Quechuan lexicon is vowel deletion. To give just two examples, Adelaar (1987: 88) and Cerrón-Palomino (2000: 315) argue that the Proto-Quechuan verb *apta- 'to grasp, grab, carry in the hand, fist' comprises the Quechuan root *apa- 'to carry' and the Proto-Aymaran directional suffix *-pta 'upward motion', which became fossilized into a single Quechuan root. According to this analysis, the intervening vowel would have been deleted, and then the initial /p/ in the /ppt/ sequence would have also been deleted to reduce a sequence of three consonants: apa-pta- > appta- > *apta-. Another more straightforward example is Proto-Quechuan *wišʎa 'ladle', which appears to have been formed by adding /ʎa/ to *wiši- 'to pour, collect, transfer liquid or grains', and then by deleting the intervening vowel.

While isolated cases such as these are not probative, evidence of such vowel deletion across the Proto-Quechuan lexicon is ubiquitous. Consider the sets of roots in Table 1. Column 1 shows Quechuan (and Aymaran) (C)V.CV roots with affricate-initial second syllables. In the other four columns are (C)VC.CV Quechuan roots, in which the roots from column 1 were apparently appended with final CV sequences beginning with /k/, /t/, /p/, /q/ and /m/. If this scenario is correct, then the intervening vowel was deleted in all cases. Significantly, this process of vowel deletion is quite similar to what we find in all attested Aymaran languages, whereby certain suffixes predictably trigger deletion of the vowel that precedes them (Coler et al. 2020; see §8 below). Note that some of the apparent base forms in column 1 are in fact Aymaran rather than Quechuan (indicated in boldface type and with parentheses), while all of the terms in columns 2–5 are Quechuan. The implications of this point are discussed in §8.

Table 2. Some Proto-Quechuan (PQ) and contemporary Quechuan roots. Aymaran forms are bolded and indicated with (PA) for Proto-Aymaran, and (JAQ) for Jaqaru

1	2	3	4	5
	/kV/-final	**/tV/-final**	**/pV/-final**	**Others**
*sutṣu- 'to slide, slip'	*sutṣka- 'to slide, slip'			*hutṣqa-* 'to slip' (YAU)**
*katṣu- 'to bite, chew'	*katṣka- 'to gnaw, chew'	*kaštu- 'to chew'		
*ʎuču- 'to take off, strip, skin, slip off, remove'		*ʎušti- 'to peel, strip, denude' *ʎuštu-* 'to peel, strip' (CAJ, ANC, JUN)		
	*ʎučka- 'slippery, to slip, slide'	*ʎust'a-* 'to slip' (BOL)	*ʎušpi- 'to slip, to leak out, lick a plate or pot clean' *ʎusp'a* 'burnished' (BOL) *ʎusp'i-* 'to slip off' (BOL)	
***utṣu(ŋsa)* 'hole' (JAQ)**	*utṣku 'hole'	*ušti-* 'to dig, dig out' (ANC)		
***(h)ač'i- 'to carry (handful)' (PA)**	*ačku-* 'take with both hands' (ANC)			
***(h)itṣ'i- 'to scratch, rip, dig, scrape' (PA)**	*hitṣka-* 'to scrape, slice' (YAU, PAC, ANC, JUN)			
***atṣ'i- 'to dig, scratch' (PA)**			*atṣpi- ~ ašpi- ~ aspi- 'to dig, scratch'	*atṣmi-* 'to hoe furrows for the first time' (JUN) *atṣqa-* 'to hoe furrows for the first time' (JUN)

** In some Quechuan languages, /h/ is an inconsistent reflex of *s.

Of particular interest in Table 2 is the fate of the affricates after the vowel deletion took place. The affricates /tʂ/ and /č/ remained unchanged before /k/ in column 2 and before /q/ and /m/ in column 5, but both affricates underwent lenition before /t/ (column 3) and /p/ (column 4). In these two contexts, the affricate became /š/.

If this kind of lenition is indeed a genuine historical fact about Proto-Quechuan phonological change, we would expect to see the sequences /tʂt/, /čt/, /tʂp/ and /čp/ less often than predicted by those consonants' frequency in the lexicon. At the same time, we would expect /št/ and /šp/ to be correspondingly over-represented. Finally, we would expect the sequences /tʂk/ and /čk/ to occur about as frequently as predicted by those consonants' occurrence in the rest of the lexicon.

A statistical analysis of Proto-Quechuan consonant adjacencies is offered in Figure 6. It relies on computing the pointwise mutual information (PMI) score, a commonly used measure of the association between outcomes of variables, for each sequence of two consonants. The use of PMI scores for modeling phonotactics was previously explored e.g., by Szabó & Çöltekin (2013) for detecting vowel harmony patterns. The key idea is that in the absence of phonotactic restrictions, the distribution of consonant clusters would be predictable from the overall distribution of consonants, which would lead to all the PMI scores being zero. A negative PMI score for a consonant cluster means that it occurs less frequently than we would expect in the absence of phonotactic effects (marked in black), and a positive value means that it occurs more frequently (marked in white). Gray values fall within the expected range. In order to determine which of the PMI scores significantly differed from zero ($p < 0.05$), we employed the standard technique of bootstrapping in order to estimate distributions of the values on data beyond our particular sample of phonetic strings.

For the most part, the statistical findings in Figure 6 confirm what we predicted from the vowel deletion patterns and subsequent phonological changes posited above: /tʂt/, /čt/ and /čp/ are all significantly underrepresented based on what we would expect from their frequencies in the lexicon ($p < 0.001$ in each case). /tʂp/, however, is more common than we would expect ($p < 0.05$), which remains to be explained. Meanwhile, /št/ and /šp/ are indeed more abundant than predicted by those consonants' frequencies in the lexicon ($p < 0.05$ in both cases). Finally, /čk/ falls within the range of expected frequency, but /tʂk/ occurs more frequently than what is expected ($p < 0.05$), which also remains to be explained.

The phonotactics of Pre-Proto-Quechuan roots represent a promising path forward for our understanding of the Quechuan lineage's history (and the effects of Aymaran contact). We reserve a fuller exploration of that issue for a further paper.

	p	t	k	q	č	tṣ	s	š	ś	r	w	y	m	n	ñ
p	pp −1,631	tp −0,838	kp −1,821	qp −1,654	čp −1,112	tṣp 0,924	sp 0,597	šp 0,762	śp 1,035	rp 0,502	wp −0,373	yp −0,075	mp 1,235	np −1,305	ñp −0,089
t	pt −0,431	tt −1,217	kt −1,046	qt 0,590	čt −0,905	tṣt −0,396	st −0,733	št 1,118	śt −1,111	rt −1,160	wt −1,123	yt 0,865	mt −0,485	nt 1,707	ñt 0,118
k	pk −1,248	tk −0,086	kk −2,011	qk −1,844	čk 0,200	tṣk 1,079	sk 0,589	šk 0,360	śk 0,374	rk 0,467	wk −0,534	yk −0,259	mk −0,901	nk 1,772	ñk −0,279
q	pq −1,654	tq −0,472	kq −1,844	qq −1,677	čq −0,157	tṣq 0,358	sq 0,309	šq −0,432	śq 0,672	rq 0,304	wq 0,172	yq −0,508	mq −0,747	nq 1,193	ñq −0,112
č	pč −0,557	tč −0,905	kč 0,429	qč 0,750	čč −0,593	tṣč −0,084	sč −0,422	šč 0,128	śč −0,799	rč −0,848	wč 0,188	yč 0,064	mč 0,193	nč 1,108	ñč 0,430
tṣ	ptṣ −0,603	ttṣ −0,396	ktṣ −0,190	qtṣ −0,042	čtṣ −0,084	tṣtṣ 0,425	stṣ 0,087	štṣ 0,063	śtṣ −0,290	rtṣ −0,339	wtṣ −0,302	ytṣ −0,026	mtṣ −0,265	ntṣ 0,686	ñtṣ 0,939
s	ps −0,941	ts −0,733	ks 0,157	qs −0,378	čs −0,422	tṣs 0,087	ss −0,250	šs −0,274	śs −0,628	rs −0,677	ws 0,344	ys 0,209	ms 0,009	ns −0,615	ñs 0,601
š	pš 0,310	tš −0,757	kš −0,185	qš 0,268	čš −0,445	tṣš 0,063	sš −0,274	šš −0,298	śš −0,652	rš −0,701	wš −0,664	yš 0,173	mš −0,627	nš −0,062	ñš 0,577
ś	pś −0,721	tś −1,111	kś 0,608	qś 1,002	čś −0,799	tṣś −0,290	sś −0,628	šś 0,308	śś −1,006	rś −1,055	wś −0,445	yś 0,972	mś −0,980	nś −0,993	ñś 0,224
r	pr 0,336	tr −1,160	kr 0,158	qr −0,105	čr −0,848	tṣr −0,339	sr −0,677	šr −0,701	śr −1,055	rr −1,104	wr −0,092	yr 0,499	mr −0,460	nr 0,468	ñr 0,175
w	pw −1,330	tw −1,123	kw −0,541	qw −0,741	čw −0,245	tṣw 0,290	sw 0,659	šw −0,664	św 0,269	rw −0,087	ww −1,029	yw 0,755	mw −0,992	nw −1,005	ñw 0,212
y	py −0,088	ty −0,846	ky −0,223	qy −1,076	čy −0,534	tṣy −0,026	sy 0,189	šy −0,387	śy −0,741	ry −0,790	wy −0,753	yy −0,476	my −0,159	ny 0,534	ñy 0,488
m	pm −1,293	tm −1,086	km −0,498	qm −0,349	čm −0,774	tṣm −0,265	sm 0,707	šm −0,065	śm −0,399	rm −0,048	wm −0,992	ym −0,716	mm −0,955	nm −0,968	ñm 0,249
n	pn −1,305	tn −1,098	kn −1,496	qn −0,749	čn −0,786	tṣn −0,278	sn −0,615	šn 0,665	śn −0,993	rn −1,042	wn −0,030	yn 0,799	mn −0,968	nn −0,980	ñn 0,236
ñ	pñ −0,089	tñ 0,118	kñ −0,279	qñ −0,112	čñ 0,430	tṣñ 0,939	sñ 0,601	šñ 1,550	śñ 0,224	rñ 0,175	wñ 0,212	yñ 1,076	mñ 0,249	nñ 0,236	ññ 1,453

Figure 6. PMI scores for consonant clusters, with significant deviations from chance marked in black (less than expected) and white (more than expected); significance based on $p < 0.05$ over 1,000 bootstrap samples

7.2 Proto-Quechuan roots with identifiable fossilized Proto-Quechuan morphology

Before conducting the CCNC method described in §3, we first eliminated from consideration the Proto-Quechuan roots from the dataset that included candidates for fossilized Proto-Quechuan suffixes identified by other linguists (e.g., most of those listed in Table 1 at the beginning of this article). We did this to show that what has already been proposed in the literature does not explain all of the effect described in this paper. Rather, we showed with very strong statistical results in §6.1 that there are patterns to be explained beyond what was previously proposed. Now, having demonstrated the strength of the non-random signal even when roots with identifiable Proto-Quechuan morphology are removed from the sample, we can reintroduce those roots into our analysis and use them as additional evidence for specific candidate morphemes from Pre-Proto-Quechuan. If we can find a known Proto-Quechuan suffix like *-rku lexicalized alongside a likely Pre-Proto-Quechuan root like *wa- 'cord; to hang, tie', which was also identified by the CCNC method, this strengthens the case for the genuineness of Pre-Proto-Quechuan *wa-. It also weakens the possibility that these roots can be explained by phonesthesia.

A good place to start in identifying genuine Pre-Proto-Quechuan roots is with the four Proto-Quechuan verbal directional suffixes shown in Table 1 above. As pointed out by several Quechua specialists going back to the 17th century

(González Holguín 1607; Parker 1973: 22–23; Adelaar 1986; Cerrón-Palomino 1987: 191–192, 1989: 33–34; Weber 1996: 180; Adelaar & Muysken 2004: 190, 231; Adelaar 2006), these four directional suffixes appear to be lexicalized within some Quechuan roots. Among the clusters of Proto-Quechuan roots identified by the experiment in §6.2, there are two that each include two verbs with apparent lexicalized directional suffixes. The first, shown in Table 1 at the beginning of this article, is a set of verbs beginning in the phonological substring /qa/ that refer to herding animals and moving objects: *qarqu- 'to expel, throw out, drive out of corral' and *qayku- 'to lead indoors, drive into a corral' (note also the Proto-Quechuan verb *qati- 'to herd animals, pursue'). Also in Table 1 above, Proto-Quechuan directional suffixes *-rpu (downward motion) and *-rku (upward motion) can be found in a second set of verbs as well, following /tʂu/, which refer to putting and placing objects. These include *tʂurpu- 'to take down object, take pot from fire, put pot on fire' and *tʂurku- 'to put an object in a high place' (note also *tʂura- 'to put').

While these sets of /qa/- and /tʂu/-initial roots appear to share fossilized morphology, it is not immediately clear what should be reconstructed in Pre-Proto-Quechuan. One possibility is *qa- 'to herd, move' and *tʂu- 'to put'. However, evidence from Proto-Quechuan phonotactics discussed in §7.1 suggests that *qati- 'to herd animals, pursue' and *tʂura- 'to put, place' could also have been the base forms. In this scenario, the changes summarized in Table 3 would have obtained, consistent with the phonotactic constraints described in §7.1:

Table 3. Possible phonological changes in the formation of some Proto-Quechua verbal roots

	***qati- 'to herd'**	***tʂura- 'to put'**
1. suffixation	*qati-rqu-*	*tʂura-rpu-*
2. vowel deletion	*qatrqu-*	*tʂurrpu-*
3. reduction of CCC sequence	*qarqu-	*tʂurpu-

This explanation is more parsimonious than positing *qa- and *tʂu- for four reasons. First, these phonological changes are consistent with the observed phonotactic patterns of Proto-Quechuan. Second, the terms *qati- and *tʂura- are already reconstructed in Proto-Quechuan, so we wouldn't have to posit new Pre-Proto-Quechuan forms *qa- and *tʂu-. Third, *qati- 'to herd' and *tʂura- 'to put' have the most basic semantics we would expect without the directional suffixes. Fourth, positing *qati- and *tʂura- instead of *qa- and *tʂu- relieves us of having to explain the residual phonological material /ti/ and /ra/ in those terms. For these reasons, we reconstruct *qati- 'to herd animals, pursue' and *tʂura- 'to put, place'

in Pre-Proto-Quechuan, while acknowledging the possibility that *qa- and *tʂu- were the Pre-Proto-Quechuan forms.[5]

Proceeding with this argument, if we find the same fossilized directional morphemes in other Proto-Quechuan clusters identified by CCNC, we can treat this as corroborating evidence for genuine lexicalized Pre-Proto-Quechuan morphemes. First, the algorithm identified three Proto-Quechuan roots beginning in /wa/ within a cluster connecting the concepts TIE, ROPE, and BUNDLE: *wata- 'to tie, repair', *waska 'rope' and *wanku- 'to wrap, bundle, bandage' (Table 4). However, there are several other roots in the Proto-Quechuan lexicon involving hanging, tying, and cord that we would add to these, which are listed in (1) in the introduction of this paper. These include *warku- 'to hang up' (boldfaced in Table 4), whose adjoining phonological material is identical to the Proto-Quechuan upward directional morpheme *-rku, and shares the same semantics.

Table 4. Some /wa/-initial Proto-Quechuan roots

Proto-Quechuan root	CLICS^2 concept	Identified by CCNC?
*wata- 'to tie, repair'	TIE	YES
*waska 'rope'	ROPE	YES
*wanku- 'to wrap, bundle, bandage'	BUNDLE	YES
*waya- 'loose, to loosen'	LOOSE	NO
*wayu- 'hanging fruit, to hang, to mature (fruit)'	HANG	NO
***warku- 'to hang up'**	**HANG UP**	**NO**
*watu 'strap, cord, belt'	BELT	NO
*waʎqa 'pendant'	NECKLACE	NO

It is possible that *wata- 'to tie, repair' was the base form here, because the deletion of /t/ would be expected in most of these environments in Table 4 (and in (2) above) based on the phonotactic patterns in Figure 6. However, *wata- 'to tie, repair' does not seem to have the most basic semantics on the list. In the absence

5. Two other likely Pre-Proto-Quechuan roots cited by Adelaar 2006:130, which were not identified by CCNC, are *su- 'to take' (cf. *surqu- 'to remove, take out, extract') and *ya- 'to go' (cf. Proto-Quechuan *yayku- 'to enter', and a full directional paradigm alongside /ya/ in some Central Peruvian varieties of Quechua; see Parker 1973:22–23). Other Proto-Quechuan roots including possible fossilized directional suffixes include *hirpu- 'to pour liquid or grains into a container, to stuff into' (i.e., downward); *tarpu- 'to sow seeds' (i.e., downward, and see other /ta/-initial planting and digging terms); and *parqu- 'to irrigate' (i.e., 'to move water out'; see other /pa/-initial water terms). These are left out of the present analysis because of their more speculative character, and because they were not identified by CCNC.

of further information, we reconstruct Pre-Proto-Quechuan *wa- 'cord; to hang, tie'. One problem with this argument is that the remaining phonetic material in the roots besides *warku- 'to hang up' remains unexplained. This is a necessary next step, but in our view, the fact that this remains unexplained does not invalidate the strength of the evidence.

Moving beyond roots that include fossilized directional suffixes, there is one more case in which known Proto-Quechuan morphology can help us confirm a cluster identified by CCNC. This is a CLICS² semantic cluster linking the concepts EAT, FOOD and SWALLOW and beginning in /mi/ in Table 5 (a). To this cluster, we would add *miški 'sweet' and *miči- 'to pasture, feed', the latter of which appears to comprise an archaic Pre-Proto-Quechuan verb root *mi- 'to eat' and the causative marker *-či that is still found across the Quechuan family – i.e., 'to cause to eat' (a point made by Parker 1969a; Adelaar 1986; Cerrón-Palomino 1987: 191; Muysken 2012).

Table 5. Some /mi/-initial Quechuan roots

a.	**Proto-Quechuan root**	**CLICS² concept**	**Identified by CCNC?**
	*miku- 'to eat'	EAT	YES
	*mirkapa 'snack, provisions'	FOOD	YES
	*miʎpu- 'to swallow'***	SWALLOW	YES
	*miči- 'to pasture, feed'	FEED	NO
	*miški 'sweet'	SWEET	NO
b.	**Modern Quechuan languages**		
	milq'uti 'esophagus' (CUS)		
	miʎkapu 'glutton' (CUS)		

*** Cerrón-Palomino (1987: 191) argues that the sequence /ʎpu/ in the root *miʎpu- 'to swallow' is a variant of the downward directional morpheme *-rpu, giving the meaning 'to eat downwards'. This is indeed convincing, and it would strengthen this analysis, but we have chosen not to adopt that analysis here until the phonetic difference in the directional morpheme can be explained.

Although *miku- is the most semantically basic term in this list, it is not likely that it is the genuine base form. /kč/ sequences are common in the Proto-Quechuan phonotactic patterns (Figure 6), so there would be no reason to expect a reduction from miku-či > mikči > *miči. Furthermore, *miku- 'to eat' appears to contain the Proto-Quechuan mediopassive morpheme *-ku, which as Cerrón-Palomino (1987: 191) observes, retains that morpheme's alternation (/u/ > /a/ before a syllable containing /u/) even in this lexicalized environment (i.e., /mikamu-/). For these reasons, we construct the monosyllabic Pre-Proto-Quechuan verb root *mi- 'to eat'.

7.3 Proto-Quechuan roots without known lexicalized morphology

A larger number of roots in clusters identified by CCNC appear alongside yet unidentified phonetic material. There is little to guide us here beyond the kinds of analytical judgments that historical linguists often have to make regarding their reconstructions. In this section, we examine the rest of the clusters of three or more roots identified by CCNC, and consider them based on our own evaluations of their plausibility. We begin with clusters of four, and then move on to clusters of three, leaving aside clusters of only two elements for reasons of space and to focus on the clusters that present the strongest evidence.

To begin with, /wi/-initial terms involving twisting and physical deformity appear in a cluster mapped to the concept TWIST, shown in Table 6 (a). Note that although some terms from particular Quechuan languages in Table 6 (b) are very similar to reconstructed Proto-Quechuan forms in (a) (e.g., PQ *wiksu 'twisted, cross-eyed, bow-legged' and Ancash Quechua *wikšu* 'having a twisted mouth'), they are in fact separate forms.

Table 6. Some /wi/-initial Quechuan roots

a.	**Proto-Quechuan root**	**CLICS^2 concept**	**Identified by CCNC?**
	*wikʎu 'twisted, deformed'	TWIST	YES
	*wiksu 'twisted, cross-eyed, bow-legged'	TWIST	YES
	*wiqru 'lame, having an injured foot, bow-legged, twisted'	TWIST	YES
	*wištu- 'twisted, crippled, to hobble, limp'	TWIST	YES
b.	**Modern Quechuan languages**		
	wiqa- 'twisted yarn, to twist' (YAU)		
	winqu 'crooked, curved, bent' (YAU)		
	winku- 'to become deformed, twisted by heat' (ANC)		
	wikru 'bent in the form of an arch' (ANC)		
	wikšu 'having a twisted mouth' (ANC)		
	wišpa 'having a twisted mouth' (JUN)		
	wipla 'crippled, limping' (ANC, JUN)		
	wiqu- 'twisted, serpentine, zig-zag' (CUS); 'to zig-zag' (ANC)		
	winkuʎu 'small twisted cord' (CUS)		

We believe that the semantic coherence of these terms is clear. The question that remains is whether we should reconstruct a Pre-Proto-Quechuan root (e.g., *wi- or *wi(kV)- 'to twist, deform'), or whether it is better explained as a phonestheme. We believe phonesthesia is a better explanation for three reasons: first, the semantics are plausibly iconic (see the comments by Firth 1930 on *twist, twirl, tweak*, etc. mentioned in the introduction to this article). Second, there is no basic verb root in the list from which the others appear to be formed (though this may be because the basic root in question simply does not appear in the Proto-Quechua reconstruction). Finally, given these factors, positing a Pre-Proto-Quechuan form *wi- or *wi(kV)- would leave more phonetic residue than seems justifiable on balance.

The next cluster of four, shown in Table 7, identifies Proto-Quechuan roots beginning with /ka/ and involving the CLICS² concepts BITE and CHEW. There are several similar terms in particular Quechuan languages, but some contain glottalized and aspirated velar stops in the relevant Quechuan varieties. In light of the contentious debate regarding the historical status of these features (e.g., Torero 1964; Parker 1969b; Stark 1975; Mannheim 1991; Landerman 1994; Campbell 1995, etc.), they are omitted here.

Table 7. Some /ka/-initial Quechuan roots

Proto-Quechuan root	CLICS2 concept	Identified by CCNC?
*katʂu- 'to bite, chew'	CHEW	YES
*katʂka- 'to gnaw, chew'	CHEW	YES
*kaštu- 'to chew'	CHEW	YES
*kani- 'to bite'	BITE	YES

It seems likely that *katʂu- 'to bite, chew', rather than *ka-, is the base form in all of these roots. For instance, *katʂka- 'to gnaw, chew' includes a final phonetic sequence /ka/, which appears to have triggered the deletion of the prior vowel in the manner described in §7.1. The same is true for *kaštu- 'to chew', with the *tʂ > š / _t lenition proposed in that section. It is possible that *kani- 'to bite' was formed this way too, i.e., /katʂu-ni-/ > /katʂni-/ > *kani- 'to bite'; this is consistent with the phonotactic patterns in Figure 6, and affricates are rarely followed by a voiced consonant in any Quechuan language. Thus, there are four advantages of reconstructing *katʂu- 'to bite, chew' as the base form of the terms in Table 7: (1) it is consistent with Proto-Quechuan phonotactics, (2) it already appears in the Proto-Quechuan list (making phonesthesia an unlikely explanation), (3) it has the most basic semantics and 4) the amount of residual phonetic material that remains unexplained is relatively small (compared to what we would have if we

reconstructed *ka-). If this argument is correct, it means that among the roots in Table 7, only *katʂu- 'to bite, chew' was an independent morpheme in Pre-Proto-Quechuan, and the rest are historically polymorphemic.

We now turn to the clusters of three Proto-Quechuan roots identified by CCNC. An interesting place to start is with three clusters that each begin with /ʎu/, with the meanings 'to peel, strip, pluck' in Table 8 (a–b), 'mud; to smear' in Table 9 (a–b) and 'slippery, to slip' in Table 10 (a–b). It is possible that those are all, in fact, a single group. However, while the semantics within each of those three clusters are quite narrow, the broader grouping is less convincing. Thus, since the CCNC method identified them separately, we have chosen to treat them separately as well. While this is the most conservative choice, it creates some curious problems – note, for instance, that the Cajamarca Quechua term *ʎučka* 'mud' (Table 9) and the Proto-Quechuan term *ʎučka- 'slippery, to slip, slide' (Table 10) appear in different categories. There is no simple solution to this problem.

The first of the /ʎu/-initial clusters, involving the connected CLICS² concepts SKIN, PULL OFF (SKIN) and PEEL, are shown in Table 8. These refer to peeling the skin from animals and plucking hair from the skin.

Table 8. A first set of /ʎu/-initial Quechuan roots

a.	**Proto-Quechuan root**	**CLICS² concept**	**Identified by CCNC?**
	*ʎupi- 'to pluck feathers or hair from the skin'	SKIN	YES
	*ʎuču- 'to take off, strip, skin, slip off, remove'	PULL OFF (SKIN)	YES
	*ʎušti- 'to peel, strip, denude'	PEEL	YES
b.	**Modern Quechuan languages**		
	ʎuč'i- 'to skin, peel, exfoliate' (CUS)		
	ʎuštu- 'peel, strip' (CAJ, ANC, JUN)		

*ʎuču- 'to take off, strip, skin, slip off, remove' is the best candidate for a base form here, for three reasons. First, that form is consistent with Proto-Quechuan phonotactics (i.e., /ʎuču-pi-/ > /ʎučpi-/ > *ʎupi-; note that /čp/ sequences are significantly infrequent in Figure 6, though it is not clear why the affricate didn't simply undergo lenition as in §7.2). Note too that the term *ʎušti- 'to peel, strip, denude' would be consistent with the lenition process described in §7.2 and statistically supported in Figure 6 (i.e., /ʎuču-ti-/ > /ʎučti-/ > *ʎušti-). Second, *ʎuču- also appears in the Proto-Quechuan list already. Third, the amount of residual phonetic material that must be explained in this interpretation is minimal. Thus, on the basis of this cluster, we reconstruct the Pre-Proto-Quechuan verb root *ʎuču-, 'to peel, strip, pluck', and argue that the rest of the Proto-Quechuan terms in Table 8 (a) are historically derived from it.

The next of the /ʎu/-initial clusters, involving the CLICS² concept MUD, are shown in Table 9:

Table 9. A second set of /ʎu/-initial Quechuan roots

a.	**Proto-Quechuan root**	**CLICS² concept**	**Identified by CCNC?**
	*ʎuta- 'to smear, plug with mud'	MUD	YES
	*ʎuši- 'to smear with mud or other substance'	MUD	YES
	*ʎuqʎa 'flood, avalanche, mudslide'	MUD	YES
b.	**Modern Quechuan languages**		
	ʎučka, lučka 'mud' (CAJ)		
	ʎusma- 'to paint' (JUN)		
	ʎunq^{h}i- 'to paint, anoint face' (CUS)		
	ʎuša- 'to paint one's face' (JUN)		
	ʎunč'i- 'to anoint, daub' (CUS)		
	ʎukʎu- 'gelatinous substance, fat; to float (fat on the surface of soup)' (CUS)		
	ʎuqmi 'porridge' (JUN)		

These roots show a clear semantic affinity. One possibility would be to reconstruct a Pre-Proto-Quechuan root *ʎu- or *ʎuču- 'mud; to smear'. However, given the possibly iconic nature of these terms (note the Cusco Quechua ideophone *ʎuq* 'sound of viscous substance'; Academia Mayor de la Lengua Quechua 2005: 101), and given the lack of a plausible base term in the list, we believe these might be formed through phonesthesia.

A third cluster of /ʎu/-initial roots, in both Proto-Quechuan and in modern Quechuan languages, are organized around the CLICS² concept SLIP. These roots are shown in Table 10. They have to do with slipping, falling and smooth or slippery surfaces.

Table 10. A third set of /ʎu/-initial Quechuan roots

a.	**Proto-Quechuan root**	**CLICS² concept**	**Identified by CCNC?**
	*ʎupti- ~ *ʎutpi- 'to slip off, come loose'	SLIP	YES
	*ʎušpi- 'to slip, to leak out, to lick a plate or pot clean'	SLIP	YES
	*ʎučka- 'slippery, to slip, slide'	SLIP	YES

Table 10. *(continued)*

b.	**Modern Quechuan languages**
	ʎutspi- 'to slip, fall down' (ANC)
	ʎuqpi- 'to slip' (ANC)
	ʎust'a- 'to slip, slip off, slip away' (BOL)****
	ʎusp'a 'worn down, polished, smooth' (CUS)
	ʎunk'u- 'lick food leftovers with index finger' (CUS, BOL)
	ʎusk'a ~ *ʎusk^{h}a* 'polished, slippery, burnished' (CUS)
	ʎunk'i- 'to smooth, burnish' (BOL)
	ʎusq'u- 'to smooth' (BOL)

**** The Bolivian and Southern Peruvian forms in (10b) are all glottalized, which in those varieties is a semi-regular outcome of affricate lenition in syllable codas (Landerman 1998: 40).

These roots are also clearly related. The phonotactic patterns in Figure 6 suggest that these Proto-Quechuan and modern Quechuan roots may contain an archaic root *ʎuč(V) 'slippery, to slip', whose final vowel is unclear. However, given that there is no plausible base form in the list, and given the possibly iconic semantics of the terms, we believe these might be formed through phonesthesia.

The next cluster comprises a group of /ma/-initial roots linked by the CLICS^{2} concept SPREAD OUT, shown in Table 11. We have not identified any other /ma/-initial roots with these semantics in modern Quechuan languages – that is, all of the cases we have found in modern Quechuan languages are reflexes of the Proto-Quechuan roots in Table 11.

Table 11. Some /ma/-initial Quechuan roots

(11)	**Proto-Quechuan root**	**CLICS^{2} concept**	**Identified by CCNC?**
	*masa- 'to spread out in the sun'	SPREAD OUT	YES
	*mašta- 'to spread out fabric'	SPREAD OUT	YES
	*manta- 'to spread out fabric'	SPREAD OUT	YES

One possibility is to reconstruct Proto-Quechuan *ma- 'to spread out' here. However, it is also possible that the Pre-Proto-Quechuan root was *masa- 'to spread out in the sun', and that *mašta- 'to spread out fabric' was formed by the addition of /ta/ and the deletion of the preceding vowel. Indeed, the resulting *s > š / _t change is plausible, because the sequence /st/ is a statistically significant gap in Proto-Quechuan phonotactics ($p < 0.001$), while /št/ is correspondingly over-

abundant ($p < 0.05$) (see Figure 6).[6] We can see the same change in pairs such as Proto-Quechuan *rasu- 'to snow, sleet, hail' and *rašta-* 'snow, slush hail; to snow' (ANC, PAC; *lašta-* in Yauyos and Junín). *masa- is also a plausible base form for *manta- 'to spread out fabric', in which case the sibilant would be deleted before the consonant cluster /nt/ (masa-nta- > masnta- > *manta-). This would leave relatively little unexplained phonetic residue (by contrast to *ma-), and these terms' semantics do not appear particularly prone to phonesthesia. Thus, we believe that *masa- 'to spread out' is the most parsimonious proposal for a Pre-Proto-Quechuan root in this case.

The next cluster includes verbs beginning with /ka/ with the CLICS² concepts ROAST OR FRY and BURN (SOMETHING), shown in Table 12.

Table 12. Some /ka/-initial Quechuan roots

a.	**Proto-Quechuan root**	**CLICS² concept**	**ID'd by CCNC?**
	*kamča- 'toasted corn, to toast'	ROAST OR FRY	YES
	*kanka- 'roasted, to roast, grill'	ROAST OR FRY	YES
	*kana- ~ *kaña- 'to burn'	BURN (SOMETHING)	YES
b.	**Modern Quechuan languages**		
	kaya- 'to light on fire, burn' (ANC)		

The semantics of these terms are narrow enough to reconstruct a root 'to burn' in Pre-Proto-Quechuan. The basic semantics of these terms do not appear consistent with a phonesthetic interpretation. Given the recurrence of a nasal consonant in each of the Proto-Quechuan terms in Table 12 (a), it is plausible that they include a Pre-Proto-Quechuan root with a nasal – that it was not *ka-, but rather *kana- or *kaña- 'to burn'. However, if this were the case, it would be difficult to explain the origin of the labial nasal in *kamča- 'toasted corn, to toast'. For this reason, we reconstruct *ka- 'to burn', while acknowledging that there may be something more going on.

Finally, the last cluster considered in this section is among the most interesting. CCNC identified a set of /ru/-initial terms linked in a CLICS² cluster involving EGG and STONE, shown in Table 13.

The first two terms in Table 13, *runtu and *ruru, have a broad and overlapping set of meanings (for instance, both refer to 'testicle' and 'egg' in various modern Quechuan varieties). The semantics of these terms allow us to reconstruct *ru

6. Intriguingly, an equivalent term *mantṣa-* 'to spread out' is attested in Central Aymaran. Since *tṣ > t is a well-known change in the Aymaran family, the Proto-Quechuan form *manta- 'to spread out fabric' may be explainable through Aymaran contact.

Table 13. Some /ru/-initial Quechuan roots

Proto-Quechuan root	CLICS² concept	ID'd by CCNC?
*runtu 'egg, hailstone, testicle'	EGG	YES
*ruru 'round thing, pit, egg, testicle, kidney'	EGG	YES
*rumi 'stone'	STONE	YES

'small, round thing'. The other terms in the list, *rumi 'stone' and *ruru 'round thing, pit, egg, testicle, kidney', are not likely base forms, because the sequences /mr/ and /rm/ are common in the Proto-Quechuan lexicon. However, there is more to this story. When we look back to the root-FINAL CLICS² clusters identified by CCNC in §6.2, we also find roots ending in /ru/ with the same meaning, including the same term *ruru listed in Table 14. (It is arguable whether 'tooth' indeed refers to a small, round thing, but this is the cluster that the CLICS² database gives us.)

Table 14. Some /ru/-final Quechuan roots

Proto-Quechuan root	CLICS² concept	ID'd by CCNC?
*kiru 'tooth'	TOOTH	YES
*muru 'seed, pit'	SEED	YES
*ruru 'round thing, pit, egg, testicle, kidney'	EGG	YES

This pattern suggests that *ru 'small, round thing' is not limited to initial position, and thus may have had a different morphosyntactic status than the other roots identified above. It may have had the quality of a shape classifier, a category of morphological elements that are ubiquitous in Amazonian languages (e.g., Aikhenvald 2012: 279–303), but are not found in the Quechuan and Aymaran languages today.[7] This is an intriguing possibility that might be explored with a different kind of semantic similarity framework.

7. Indeed, there are many other such recurrent shape-based elements in the Proto-Quechuan lexicon: /pu/-terms regarding swollen shapes (e.g., *punki- 'to swell, inflate', *pukuču 'bladder', *pušuʎu ~ pušʎu ~ šupuʎu 'blister', *puyñu 'pitcher, jug' *puru 'gourd, vessel made from gourd', as well as dozens of terms from particular Quechuan languages; see also Urban 2018); /ti/-terms regarding pointy things (e.g., *timpi- 'to stick, nail', *tipa- 'pin, to fasten or prick with a pin', *tipki 'metal pin'); /tu/-terms regarding poking (e.g., *tuqu- 'hole, to make a hole', *tukši- 'to stab, prick, puncture', *tupši- 'to peck, point with finger'); /qu/-terms regarding bodily bulges and protrusions (e.g., *qunqur 'knee', *qutu 'tumor', *quruta 'testicle'); etc. If indeed these involve shape-based classifiers, Quechuan may have been more typologically similar to Amazonian languages early in the lineage's history than at the Proto-Quechuan stage (Mannheim 2018: 512).

7.4 Rejected clusters

The CCNC method turned up three clusters of three or more Proto-Quechuan lexemes that we have dismissed because we deem their semantics to be insufficiently narrow. Some number of such spurious similarities between phonetic strings for similar concepts are to be expected in any dataset of this size (as the results on pseudo-lexicons have demonstrated).

One cluster identified by CCNC includes the Proto-Quechuan verbs *muča- 'to kiss', *musya- 'to divine, sense, realize, perceive', *muki- 'to choke, asphyxiate, suffocate because of a strong odor; to rot' and *mutki 'to smell, perceive odor'.[8] Similarly, CCNC identified another cluster including *wiši- 'to pour, collect, transfer liquid or grains', *wišču- 'to discard, throw away, toss out' and *wika- 'to throw into the air'. We are not convinced of the semantic unity of these terms either. However, we noted earlier that *wiši- 'to pour, collect, transfer liquid or grains' is likely related to *wišʎa 'ladle', though this connection was not identified by CCNC.

Three /ču/-initial terms are mapped to the CLICS² term MAIZE: *čučuqa 'corn-based dish', *čuʎpi 'corn variety' and *čuqʎu 'ear of corn'. However, the Quechuan languages have vast and rich inventories of corn terms, most of which do not include /ču/. Therefore, we find the evidence insufficient to support this cluster.

7.5 Limitations of the semantic framework: An example

We developed and used CCNC in this paper to provide strong, objective statistical proof of the correspondence between sub-root phonetic strings and semantic similarity – as represented by colexification – in Proto-Quechuan. However, colexification is a rather restrictive model of semantic relatedness, and it is not optimal for detecting the kinds of semantic connections that we would expect to be expressed by derivational morphology (which may be the most relevant in the Proto-Quechuan case). For example, Quechuan terms for seeds and grains tend to include the string /mu/. However, CCNC did not detect the semantic connection between SEED and THRESH (shown in Table 15) because no languages in the CLICS² sample use the same term for those concepts. This is unsurprising, given the nature of the lexicalization as a proxy for semantic similarity; in order to detect such a connection, we would need to employ a different kind of external semantic standard.

8. We think it is more likely that three /mus/-initial verbs are related: *musqu- 'to dream', *muspa- 'to daydream, be delirious, rave' and *musya- 'to guess, divine, sense, realize, perceive'. However, these verbs were not identified by CCNC, so we leave them out of this analysis.

Table 15. Some /mu/-initial Quechuan roots

a.	**Proto-Quechuan root**	**CLICS² concept**	**ID'd by CCNC?**
	*muru 'seed, pit'	SEED	NO
	*murka/i- 'to thresh grains'	THRESH	NO
	*muhu 'seed'	SEED	NO
	*muti 'boiled corn kernels'	MAIZE	NO
b.	**Modern Quechuan languages**		
	murmiy ~ murmuy 'small grain' (CUS)		
	muc^{h}a- 'to remove the grains from an ear of corn' (CUS)		
	mukʎu 'seed (coca)' (CUS)		

It is likely that the underlying Pre-Proto-Quechuan form here is *muru 'seed, pit', because (1) it is already in the list in Table 15 (a); (2) it has the most basic semantics of that list; (3) the phonotactic patterns in Figure 6 suggest that /r/ might be altered or deleted in all of the relevant contexts. However, there is a much larger group of /mu/-initial terms that refer to small, round things more generally, which may suggest that *muru 'seed, pit' itself contains an archaic component morpheme *mu (affixed with the same *ru 'small, round thing' in Table 14). Since CLICS² did not make this connection available to CCNC, we simply mention it here as a likelihood, and do not include it in the final list in our conclusion.

Thus, the CLICS² database has been a helpful external measure for establishing the plausibility of our argument, but a fuller account of fossilized Pre-Proto-Quechuan morphology will require a semantic framework better suited to the specific kind of semantic relatedness at stake in this case. One possibility is searching for shape-based regularities, as mentioned above, to explore whether these are indeed archaic shape classifiers. Some candidates include *ru 'small round thing', *mu 'small round thing', *ti 'pointy thing', *tu 'poking thing', *pu 'swollen shape' (see also Urban 2018) and *qu 'bodily bulge or protrusion' (see footnote 7).

8. Conclusion and next steps

In this paper, we offered evidence that some Proto-Quechuan roots are historically polymorphemic. We did this by developing the Crosslinguistic Colexification Network Clustering (CCNC) algorithm (described in §3), and applying it (§5) to a list of 809 reconstructed Proto-Quechuan lexical items and 259 reconstructed Proto-Aymaran lexical items (as well as an English list). In a first sta-

tistical test (§6.1), the Proto-Quechuan list showed clearly non-random phonetic similarity among semantically related roots. By contrast, the Proto-Aymaran and English lists did not show non-random phonetic similarity among semantically related roots. In a second statistical test (§6.2), we found that these non-random phonetic sequences appear largely at the beginning of Proto-Quechuan roots. We then evaluated the clusters of Proto-Quechuan roots identified by the CCNC method (§§7.2–7.4), one by one, on the qualitative basis of semantic plausibility; some of these were found to contain phonesthemes. We drew on our analysis of Proto-Quechuan phonotactics (§7.1) as we evaluated the evidence for particular Pre-Proto-Quechuan roots.

The nine Pre-Proto-Quechuan roots we proposed in this article are listed in (3).

(3) *Reconstructed Pre-Proto-Quechuan roots*
*tʂura- 'to put'
*katʂu- 'to bite, chew'
*ka- 'to burn'
*ʎuču- 'to peel, strip, pluck'
*masa- 'to spread out'
*mi- 'to eat'
*qati- 'to herd, move'
*ru 'small, round thing'
*wa- 'cord; to hang, tie'

According to our analysis, roots in the list above (like *katʂu- 'to bite, chew') go back to Pre-Proto-Quechuan, while other Proto-Quechuan roots (like *katʂka- 'to gnaw, chew') were later built up from them. In many cases, it is not yet clear what the adjoining phonetic material (e.g., final /ka/ in *katʂka-) might have been, though in some cases this phonetic material clearly corresponds to known Proto-Quechuan morphology.

On the basis of our phonotactic analysis, we also posited several sound changes that appear to have affected these Pre-Proto-Quechuan morphemes as they became lexicalized within Proto-Quechuan roots. Most notably, this includes a process by which suffixation triggers the deletion of an intervening vowel, similar to the irregular process of morphophonemic vowel deletion that we find in the Aymaran languages. The statistical analysis presented in Figure 6 suggests that some phonological changes then affected the resulting consonant clusters, for example those in Table 16.

Interestingly, the processes of root formation described in this article do not seem to have stopped entirely at the Proto-Quechuan stage. Some semantically related roots sharing initial phonetic substrings are found only in individual

Table 16. Some proposed sound changes affecting Proto-Quechuan roots

	Pre-Proto-Quechuan (or Aymaran)		Proto-Quechuan
tʂ > š / _t	*katʂu- 'to bite, chew'	>	*kaštu- 'to chew'
č > š / _t	*ʎuču- 'to take off, strip, skin, slip off, remove'	>	*ʎušti- 'to peel, strip, denude'
tʂ > š / _p	*atʂ'i- 'to dig, scratch' (PA)	>	*atʂpi- ~ <u>ašpi-</u> ~ aspi- 'to dig, scratch'
s > š / _t	*masa- 'to spread out'	>	*mašta- 'to spread out fabric'

Quechuan languages or branches, and seem to have emerged after the diversification of the family. For instance, Southern Peruvian and Bolivian Quechua have made great phonesthetic use of /ʎu/-initial terms for slipperiness, as in Table 10 (b), while /wi/-initial terms for twisting and bodily deformity have proliferated in the Quechuan varieties of Central Peru, as in Table 6 (b). The apparent exceptions to these regional embellishments are the Quechuan varieties of Ecuador and Northern Peru, which do not exhibit many such form/meaning correspondence beyond those inherited from Proto-Quechuan. This process seems to have halted during the Quechuan expansion into the Northern Andes. (Note that this may present possibilities for establishing a relative chronology of the Northern Quechuan expansion.)

One recurrent question in our analysis is the extent to which this phenomenon can be explained as the product of phonesthesia. We argued that a few terms can indeed be explained that way. However, this is an implausible explanation for the entire pattern, because (1) many of our recurrent phonetic substrings do, in fact, co-occur with known Quechuan morphology; (2) the semantics of many roots identified by CCNC are not likely to be involved in phonesthesia; (3) these recurrent elements are simply too widespread in the lexicon.

A related question is how to account for the yet unexplained phonetic residues adjoining our recurrent phonetic substrings. One next step, certainly, is to search for semantic regularities there. For instance, many Quechuan roots ending in /tʂi/ have to do with yanking: *rutʂi-* 'to yank out grass' (CAJ) (cf. PQ *rutu- 'to shear, cut hair'); *ʎaptʂi-* 'to pull the top of a stalk off a plant' (YAU) (cf. PQ *ʎapi- 'to squeeze, crush, smoosh'); *latʂi-* 'to yank a string' (JUN) (cf. *lapu-* 'to pull hair', JUN) and many others besides. To investigate this systematically, we would need to develop a different standard of semantic similarity that covers the Proto-Quechuan data better than the CLICS² database (for instance, a network that captures derivational connections). Once this is possible, we expect that many more archaic morphemes will become clear within Proto-Quechuan roots, both initially and finally.

The findings presented here have implications for South American linguistic prehistory. First, they offer a view on the initial convergence between Pre-Proto-Quechuan and Pre-Proto-Aymaran. As outlined in the introduction to this paper, our findings lend empirical support to Adelaar's hypothesis, discussed in the introduction, that "Quechuan may have adopted an Aymaran model by reassigning elements from its own original morphemic inventory to borrowed functions" (2012b: 463). The Proto-Quechuan lexicon exhibits a strong correspondence between sub-root phonetic strings and semantic similarity, which is what we would expect if some Proto-Quechuan roots contain Pre-Proto-Quechuan morphemes that were lexicalized within (mostly) minimally bisyllabic roots (as in Aymara). We also showed that Proto-Aymaran does not exhibit a correspondence between sub-root phonetic strings and semantic similarity, which suggests that Aymaran may indeed have been the model on which this aspect of Pre-Proto-Quechuan structure was reformatted.

Comparing the phonotactics and morphophonemics of Proto-Quechuan and Proto-Aymaran also represents a promising way forward. To take just one example, the thorny issue of vowel deletion is particularly important for this discussion. In all attested Aymaran languages, particular suffixes delete the preceding vowel (Hardman 1983; Cerrón-Palomino 2000; Coler 2014: 55–59; Coler et al. 2020), a system that goes back to Proto-Aymaran. For instance, when the Proto-Aymaran verb *hala- 'to run' is suffixed with the outward directional suffix *-šu, the intervening vowel is deleted (/hala-šu/ > /halšu/) (Cerrón-Palomino 2000: 247). This is what has been hypothesized in the Quechuan examples throughout this paper, though it appears more predictable in this Quechuan case than in Aymaran.

It is also notable that one of the major phonotactic distinctions that has been identified between the Quechuan and Aymaran languages is the presence of voiceless consonant codas in Quechuan roots, and their absence in Aymaran roots (Adelaar 1986; Emlen 2017). If our analysis is correct, this tendency was made more acute by this vowel deletion process. If so, the Quechuan and Aymaran lineages might have been more similar in this respect before the initial convergence than after.

Examining the historically polymorphemic nature of some Proto-Quechuan roots allows us to approach the early Quechuan-Aymaran interaction with greater nuance. For instance, consider the suggestive connection between the Proto-Aymaran and Proto-Quechuan roots in Table 17. None of the Proto-Aymaran terms is shared with any Quechuan variety, and none of the Proto-Quechuan terms is shared with any Aymaran variety.

It appears that the Proto-Quechuan roots in the right column of Table 17 are built up from the Proto-Aymaran roots in the left column. These might have been early Aymaran loans in Pre-Proto-Quechuan – or, more provocatively, they may be cognates inherited from a common Quechumaran ancestor language. Explain-

Table 17. Some Proto-Aymaran and Proto-Quechuan verbal roots

Proto-Aymaran	Proto-Quechuan
*atʂ'i- 'to dig, scratch'	*atʂpi- ~ ašpi- ~ aspi- 'to dig, scratch'
*aya- 'to carry (long objects)'	*aysa- 'to pull, drag, haul with rope'
*č'iʎa- 'to cut, pull apart, peel'	*čiʎpi- 'to split into pieces'

ing them will require a more nuanced relative chronology of contact between the lineages, a willingness to approach Proto-Quechuan forms below the level of the root (as demonstrated in this article), and quantitative analysis of both languages' phonotactic patterns.

A second implication of our argument for South American linguistic prehistory has to do with the external genetic relations of Pre-Proto-Quechuan. If a language related to Quechuan does indeed survive somewhere in the region – which might be difficult to recognize, especially if it remained outside of the transformative influence of Aymaran contact – then it will be important to keep the archaic elements presented in this paper in mind. If our account is correct, only Proto-Quechuan terms like *katʂu- 'to bite, chew', and not historically polymorphemic Proto-Quechuan terms like *katʂka- 'to gnaw, chew', would be the most appropriate comparanda. It remains to be seen whether enough Quechuan material can be reconstructed to support such a comparison, but this approach would certainly be most consistent with what we now know about the early history of the Quechuan lexicon.

Our findings move us somewhat further away from a Quechumaran genealogical grouping, at least in the two ways it has been articulated so far. First, before the 1960s, this grouping was proposed to explain the families' many identical or nearly identical lexical items and strong structural resemblances. Then, once most Andeanists since the 1960s took those obvious resemblances to be the result of contact, the Quechumaran hypothesis came to refer to something new: a genealogical explanation for the less obvious resemblances that might remain in the lexicons or in the grammatical structures once those superficial similarities had been accounted for (Campbell 1995; Cerrón-Palomino 2000: 311–312). However, our analysis suggests that a large part of the Proto-Quechuan lexicon is made up of fossilized polymorphemic roots, while the Proto-Aymaran lexicon is not. If this is true, then a sensible way forward is to account for as much archaic morphology within Proto-Quechuan roots as possible, as part of a broader comparative project regarding the two lineages. As we have argued throughout this paper, we believe that lexicalized archaic morphemes are a widespread property of the Proto-Quechuan lexicon, and that we have only scratched the surface. (Quechuan

specialists reading this article surely will have thought of many more promising candidates.)

This third, new iteration of the Quechumaran hypothesis is thus something rather different from the first two, and is certainly worth pursuing. However, the further we walk down this path, the further we move away from what motivated centuries of scholars to assert a Quechuan-Aymaran genealogical grouping in the first place: the striking formal resemblances between the Quechuan and Aymaran languages. For this reason, it may be more fruitful at this point to instead search more widely for the external relations of Pre-Proto-Quechuan and Pre-Proto-Aymaran (Adelaar 1986: 380; 2013a), with the patterns presented here in hand.

Funding

The research leading to these results received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement no. 295918; Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number UR 310/1–1; and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 818854 – SAPPHIRE).

Acknowledgements

The authors thank many colleagues for their helpful comments and guidance over the course of this article's long journey to publication: Willem Adelaar, Claire Bowern, Rodolfo Cerrón-Palomino, Simeon Floyd, Harald Hammarström, Bruce Mannheim, Matthias Pache, Michael Proctor, Joe Salmons, Hermann Sonntag and Matthias Urban. Thanks also to four anonymous reviewers for their careful and constructive feedback. Work was carried out at the Leiden University Centre for Linguistics, the John Carter Brown Library, and at the DFG Center for Advanced Studies "Words, Bones, Genes, Tools" at the University of Tübingen.

Abbreviations

ANC Ancash Quechua
BOL Bolivian Quechua
CAJ Cajamarca Quechua
CUS Cusco Quechua
ECU Ecuadorian Quichua
JAQ Jaqaru (Central Aymaran)
JUN Junín-Huanca Quechua
PAC Pacaraos Quechua
TAR Tarma Quechua
YAU Yauyos Quechua
PQ Proto-Quechuan
PA Proto-Aymara

References

Academia Mayor de La Lengua Quechua. 2005. *Diccionario: Quechua-Español-Quechua, Qheswa-Español-Qheswa: Simi Taqe*. Cuzco: Academia Mayor de La Lengua Quechua.

Adelaar, Willem F. H. 1986. La relación Quechua-Aru: Perspectivas para la separación del léxico. *Revista Andina* 4(2). 379–426.

Adelaar, Willem F. H. 1987. Comentarios a Martha Hardman y Bruce Mannheim. *Revista Andina* 5(2). 83–91.

Adelaar, Willem F. H. 2006. The vicissitudes of directional affixes in Tarma (Northern Junin) Quechua. In Grażyna J. Rowicka & Eithne B. Carlin (eds.), *What's in a verb? Studies in the verbal morphology of the languages of the Americas*, 121–141. Utrecht: LOT.

Adelaar, Willem F. H. 2009. Inverse markers in Andean languages: A comparative view. In Leo Wetzels (ed.), *The linguistics of endangered languages. Contributions to morphology and morpho-syntax*, 171–185. Utrecht: Netherlands Graduate School of Linguistics (LOT).

Adelaar, Willem F. H. 2010. Trayectoria histórica de la familia lingüística quechua y sus relaciones con la familia lingüística aimara. *Boletín de Arqueología PUCP* 14. 239–254.

Adelaar, Willem F. H. 2011. Reconstruyendo el paradigma verbal quechua: El caso de la transición de primera a segunda persona. In Willem F. H. Adelaar, Valenzuela Pilar & Roberto Zariquiey (eds.), *Estudios sobre lenguas andinas y amazónicas. Homenaje a Rodolfo Cerrón-Palomino*, 21–31. Lima: Fondo Editorial de la Pontificia Universidad Católica del Perú.

Adelaar, Willem F. H. 2012a. Languages of the Middle Andes in areal-typological perspective: Emphasis on Quechuan and Aymaran. In Lyle Campbell & Grondona, Verónica (eds.), *The indigenous languages of South America: A comprehensive guide*, 575–624. Berlin: Walter de Gruyter. https://doi.org/10.1515/97831102580 35.575

Adelaar, Willem F. H. 2012b. Modeling convergence: Towards a reconstruction of the history of Quechuan-Aymaran interaction. *Lingua* 122(5). 461–469. https://doi.org/10.1016/j.lingua.2011.10.001

Adelaar, Willem F. H. 2013a. Quechua I y Quechua II: En defensa de una distinción establecida. *Revista Brasileira de Linguística Antropológica* 5(1). 45–65. https://doi.org/10.26512/rbla.v5i1.16542

Adelaar, Willem F. H. 2013b. Searching for Undetected Genetic Links between the Languages of South America. In Ritsuko Kikusawa & Reid, Lawrence A. (eds.), *Historical linguistics 2011. Selected papers from the 20th International Conference on Historical Linguistics, Osaka, 25–30 July 2011*, 115–128. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/cilt.326.10ade

Adelaar, Willem F. H. 2017. A typological overview of Aymaran and Quechuan language structure. In Alexandra Y. Aikhenvald and Dixon, R. M. W. (eds.), *The Cambridge handbook of linguistic typology*, 651–682. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316135716.021

Adelaar, Willem F. H. & Pieter Muysken. 2004. *The languages of the Andes*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486852

Aikhenvald, Alexandra Y. 2012. *Languages of the Amazon*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199593569.001.0001

Blake, Barry J. 2017. Sound symbolism in English: Weighing the evidence. *Australian Journal of Linguistics* 37(3). 286–313. https://doi.org/10.1080/07268602.2017.1298394

Blasi, Damián E., Søren Wichmann, Harald Hammarström, Peter F. Stadler & Morten H. Christiansen. 2016. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences* 113(39). 10818–10823. https://doi.org/10.1073/pnas.1605782113

Campbell, Lyle. 1995. The Quechumaran hypothesis and lessons for distant genetic comparison. *Diachronica* 12(2). 157–200. https://doi.org/10.1075/dia.12.2.02cam

Campbell, Lyle & William J. Poser. 2008. *Language classification: History and method*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486906

Cerrón-Palomino, Rodolfo. 1987. *Lingüística Quechua*. Cuzco: Centro de Estudios Rurales Andinos Bartolome de las Casas.

Cerrón-Palomino, Rodolfo. 1989. *Lengua y sociedad en el valle del Mantaro*. Lima: Instituto de Estudios Peruanos.

Cerrón-Palomino, Rodolfo. 1994. *Quechumara. Estructuras paralelas de las lenguas Quechua y Aymara*. La Paz: CIPCA.

Cerrón-Palomino, Rodolfo. 2000. *Lingüística Aimara*. Cuzco: Centro de Estudios Regionales Andinos Bartolomé de Las Casas.

Cerrón-Palomino, Rodolfo. Forthcoming. Lingüística histórica y filología en el área andina: Encuentros y desencuentros. *Revista Andina* 56. 101–128.

Cobo, Bernabé. 1890 [1653]. *Historia del nuevo mundo*. Sevilla: Imp. de E. Rasco.

Coler, Matt. 2014. *A grammar of Muylaq' Aymara: Aymara as spoken in Southern Peru*. Leiden: Brill. https://doi.org/10.1163/9789004284005

Coler, Matt, Nicholas Q. Emlen & Edwin Banegas-Flores. 2020. Vowel deletion in two Aymara varieties. *Italian Journal of Linguistics* 32(1).

Dunn, Michael, & Angela Terrill. 2012. Assessing the lexical evidence for a Central Solomons Papuan family using the Oswalt Monte Carlo Test. *Diachronica* 29(1). 1–27. https://doi.org/10.1075/dia.29.1.01dun

Emlen, Nicholas Q. 2017. Perspectives on the Quechua-Aymara contact relationship and the lexicon and phonology of Pre-Proto-Aymara. *International Journal of American Linguistics* 83(2). 307–340. https://doi.org/10.1086/689911

Emlen, Nicholas Q. & Willem F. H. Adelaar. 2017. Proto-Quechua and Proto-Aymara agropastoral terms: Reconstruction and contact patterns. In Martine Robbeets & Alexander Savelyev (eds.), *Language dispersal beyond farming*, 25–45. Amsterdam: John Benjamins. https://doi.org/10.1075/z.215.02eml

Firth, John R. 1930. *Speech*. London: Ernest Benn.

François, Alexandre. 2008. Semantic maps and the typology of colexification. In Vanhove Martine (ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 163–215. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.106.09fra

Gentner, Dedre. 1981. Some interesting differences between nouns and verbs. *Cognition and Brain Theory* 4(2). 161–178.

González Holguín, Diego. 1607. *Gramática y arte nueva de la lengua general de todo el Perú llamada lengua qquichua o lengua del Inca*. Lima: Francisco del Canto (1842 edition, Genoa: Pagano).

Hardman, Martha J. 1983. *Jaqaru: Compendio de estructura fonológica y morfológica*. Lima: Instituto de Estudios Peruanos.

Haynie, Hannah J. 2014. Deep relationships among California languages. *Diachronica* 31(3). 407–447. https://doi.org/10.1075/dia.31.3.04hay

Heggarty, Paul & David Beresford-Jones. 2010. Agriculture and language dispersals: Limitations, refinements, and an Andean exception? *Current Anthropology* 51(2). 163–191. https://doi.org/10.1086/650533

Kessler, Brett. 2001. *The significance of word lists: Statistical tests for investigating historical connections between languages*. Stanford: CSLI Publications.

Kwon, Nahyun & Erich R. Round. 2015. Phonaesthemes in morphological theory. *Morphology* 25(1). 1–27. https://doi.org/10.1007/s11525-014-9250-z

Laime Ajacopa, Teofilo, Efraín Cazazola, Félix Layme Pairumani & Pedro Plaza Martínez. 2007. *Diccionario bilingüe, Iskay Simipi Yuyayk'ancha: Quechua-Castellano, Castellano-Quechua*. Unpublished manuscript. La Paz, Bolivia.

Landerman, Peter. 1994. Glottalization and aspiration in Quechua and Aymara reconsidered. In Peter Cole, Gabriella Hermon & Mario Daniel Martín (eds.), *Language in the Andes*, 332–378. Newark: Latin American Studies Program, University of Delaware.

Landerman, Peter. 1998. Internal reconstruction in Aymara and Quechua. In Jane H. Hill, P. J. Mistry, Lyle Campbell (eds.), *The life of language: Papers in linguistics in honor of William Bright*, 35–57. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110811155.35

List, Johann-Mattis, Simon J. Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi & Robert Forkel. 2018. CLICS^2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology* 22(2). 277–306. https://doi.org/10.1515/lingty-2018-0010

List, Johann-Mattis, Anselm Terhalle & Matthias Urban. 2013. Using network approaches to enhance the analysis of cross-linguistic polysemies. In *Proceedings of the 10th International Conference on Computational Semantics – Short Papers*, 347–353. Potsdam: Association for Computational Linguistics.

Mannheim, Bruce. 1991. *The Language of the Inka since the European invasion*. Austin: University of Texas Press.

Mannheim, Bruce. 2018. Three axes of variability in Quechua: Regional diversification, contact with other indigenous languages, and social enregisterment. In Linda J. Seligmann & Kathleen Fine-Dare (eds.), *The Andean world*, 507–523. Milton Park: Routledge. https://doi.org/10.4324/9781315621715-33

Mason, John Alden. 1950. The languages of South American Indians. In Julian Steward (ed.), *Handbook of South American Indians, volume 6*, 157–317. Washington, D.C.: United States Government Printing Office.

Mossel, Arjan, Nicholas Q. Emlen, Simon van de Kerke & Willem F. H. Adelaar. 2020. Puquina kin terms. In Astrid Alexander Bakkerus, Rebeca Fernández Rodríguez, Liesbeth Zack & Otto Zwartjes (eds.), *Missionary linguistic studies from Mesoamerica to Patagonia*, 277–298. Leiden: Brill.

Münch, Alla & Johannes Dellert. 2015. Evaluating the potential of a large-scale polysemy network as a model of plausible semantic shifts. In Wahle, Johannes, Marisa Köllner, Harald Baayen, Gerhard Jäger & Ttineke Baayen-Oudshoorn (eds.), *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*. [http://www.sfs.uni-tuebingen.de/~jdellert/pubs/amuench-jdellert-2015-polysemy-semantic-shifts.pdf, accessed 25 April, 2020].

Muysken, Pieter. 2012. Modelling the Quechua-Aymara relationship: Structural features, sociolinguistic scenarios, and possible archeological evidence. In Paul Heggarty & David Beresford-Jones (eds.), *Archaeology and language in the Andes: A cross-disciplinary exploration of prehistory*, 85–110. Oxford: Oxford University Press. https://doi.org/10.5871/bacad/9780197265031.003.0004

Oré, Luís Jerónimo de. 1607. *Rituale seu Manuale Peruanum et forma brevis administrandi apud Indos sacramenta per Ludovicum Hieronymum Orerium*. Napoli: Apud Io. Iacobum Carlinum & Constantinum Vitalem.

Orr, Carolyn & Robert E. Longacre. 1968. Proto-Quechumaran. *Language* 44(3). 528–555. https://doi.org/10.2307/411720

Parker, Gary J. 1963. La clasificación genética de los dialectos Quechuas. *Revista del Museo Nacional* 32. 241–252.

Parker, Gary J. 1969a. Comparative Quechua phonology and grammar I: Classification. *University of Hawaii Working Papers in Linguistics* 1(1). 65–87.

Parker, Gary J. 1969b. Comparative Quechua phonology and grammar III: Proto-Quechua lexicon. *University of Hawaii Working Papers in Linguistics* 1(4). 1–61.

Parker, Gary J. 1973. *Derivacion verbal en el quechua de Ancash*. Lima: Universidad Nacional de San Marcos, Centro de Investigación de Lingüística Aplicada.

Ringe, Donald A. 1992. On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society* 82(1). 1–110. https://doi.org/10.2307/1006563

Salmons, Joseph C. & Brian D. Joseph. 1998. *Nostratic: Sifting the evidence*. Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/cilt.142

Stark, Louisa R. 1975. A reconsideration of Proto-Quechua phonology. *Lingüística e indigenismo moderno de América. Trabajos presentados al XXXIX CIA*, 209–219. Lima: IEP.

Steiner, Lydia, Michael Cysouw & Peter Stadler. 2011. A pipeline for computational historical linguistics. *Language Dynamics and Change* 1(1). 89–127. https://doi.org/10.1163/221058211X570358

Szabó, Lili & Çağrı Çöltekin. 2013. A linear model for exploring types of vowel harmony. *Computational Linguistics in the Netherlands Journal* 3. 174–192.

Torero, Alfredo. 1964. Los dialectos Quechuas. *Anales Científicos* 2(4). 446–478.

Uhle, Max. [1910] 1969. *Estudios sobre historia incaica*. Lima: Universidad Nacional Mayor de San Marcos.

Urban, Matthias. 2018. Quechuan terms for internal organs of the torso: Synchronic, diachronic, and typological perspectives. *Studies in Language*. 42(3). 505–528. https://doi.org/10.1075/sl.16081.urb

Weber, David John. 1987. *Estudios quechua: Planificacion, historia y gramática*. Lima: Ministerio de Educación, Instituto Lingüístico de Verano.

Weber, David John. 1996. *Una gramática del quechua del Huallaga (Huánuco)*. Lima, Peru: Instituto Lingüístico de Verano.

Zalizniak, Anna A., Maria Bulakh, Dmitrij Ganenkov, Ilya Gruntov, Timur Maisak & Maxim Russo. 2012. The catalogue of semantic shifts as a database for lexical semantic typology. *Linguistics*. 50(3). 633–669. https://doi.org/10.1515/ling-2012-0020

Abstract

In the Proto-Quechuan lexicon, many two-segment phonetic substrings recur in semantically related roots, even though they are not independent morphemes. Such elements may have been morphemes before the Proto-Quechuan stage. On the other hand, this may simply be due to chance, or to phonesthesia. In this paper, we introduce a methodology which allows us to evaluate our claims against a neutral standard of semantic relatedness. We obtain very strong statistical evidence that there are hitherto unexplained recurrent elements within Proto-Quechuan roots, but not within Proto-Aymaran roots. Most appear to reflect archaic Quechuan morphology, which has implications for the early Quechuan-Aymaran relationship.

Résumé

Dans le lexique proto-quechua, de nombreuses sous-chaînes phonétiques à deux segments se répètent dans des racines de sens similaire, même si synchroniquement elles ne peuvent pas être analysées comme des morphèmes indépendants. D'un côté, ces éléments peuvent avoir été des morphèmes avant le stade proto-quechua. De l'autre, les similitudes pourraient simplement être dues au hasard ou à la phonesthésie. Dans cet article, nous introduisons une méthodologie qui permet de tester les deux hypothèses par rapport à un standard neutre de similarité sémantique. Nous obtenons des preuves statistiques très solides qu'il existe des éléments récurrents jusque-là inexpliqués dans les racines proto-quechuas, mais pas dans les racines proto-aymaras. Nous montrons également la plupart semblent refléter de la morphologie du quechua archaïque. Ces résultats alimentent les théories sur les premières relations entre les langues quechuas et l'aymara.

Zusammenfassung

Im rekonstruierten Wortschatz des Proto-Quechua lassen sich viele aus zwei Segmenten bestehende Lautfolgen in semantisch verwandten Wurzeln entdecken, die sich aber auf dieser Sprachstufe nicht mehr auf unabhängige Morpheme zurückführen lassen. Solche Lautfolgen können einerseits ererbte Morpheme aus einer älteren Sprachstufe darstellen, andererseits aber auch einfach auf zufällige Ähnlichkeiten oder auf Phonästhesie zurückzuführen sein. In diesem Artikel beschreiben wir eine neuartige Methode, mit der wir diese Frage auf der Grundlage eines unabhängigen Standards semantischer Ähnlichkeit statistisch bewerten können. Wir erhalten sehr starke Evidenz dafür, dass die wiederkehrenden Elemente in Proto-Quechua-Wurzeln nicht auf Zufall beruhen können, während eine parallele Analyse des Proto-Aymara keine signifikanten Muster entdeckt. Nach einer genauen Analyse der so gefundenen Lautfolgen stellen wir fest, dass die meisten in der Tat zu plausiblen archaischen Quechua-Morphemen führen, die zur weiteren Klärung des Verhältnisses zwischen frühem Quechua und Aymara beitragen können.

Address for correspondence

Nicholas Q. Emlen
Leiden University Centre for Linguistics
Postbus 9515
2300 RA Leiden
The Netherlands
n.q.emlen@hum.leidenuniv.nl
https://orcid.org/0000-0003-0702-1982

Co-author information

Johannes Dellert
University of Tübingen
DFG Center for Advanced Studies "Words, Bones, Genes, Tools"
Department of Linguistics
jdellert@sfs.uni-tuebingen.de

Publication history

Date received: 2 December 2016
Date accepted: 12 May 2020
Published online: 5 August 2020