WILEY-BLACKWELL

Cladistics (2014) 1–13

Cladistics

10.1111/cla.12078

Historical linguistics as a sequence optimization problem: the evolution and biogeography of Uto-Aztecan languages

Ward C. Wheeler[a,*] and Peter M. Whiteley[b]

[a]Division of Invertebrate Zoology, American Museum of Natural History, Central Park West @ 79th Street, New York, NY, 10024-5192, USA; [b]Division of Anthropology, American Museum of Natural History, Central Park West @ 79th Street, New York, NY, 10024-5192, USA

Accepted 18 March 2014

Abstract

Language origins and diversification are vital for mapping human history. Traditionally, the reconstruction of language trees has been based on cognate forms among related languages, with ancestral protolanguages inferred by individual investigators. Disagreement among competing authorities is typically extensive, without empirical grounds for resolving alternative hypotheses. Here, we apply analytical methods derived from DNA sequence optimization algorithms to Uto-Aztecan languages, treating words as sequences of sounds. Our analysis yields novel relationships and suggests a resolution to current conflicts about the Proto-Uto-Aztecan homeland. The techniques used for Uto-Aztecan are applicable to written and unwritten languages, and should enable more empirically robust hypotheses of language relationships, language histories, and linguistic evolution.

Introduction

How languages evolve has long been a central question for the human sciences. Linguistic elements may be transmitted horizontally ("borrowing") among neighbouring languages, but most language transmission obviously occurs via lineal descent with modification. Linguistic and biological evolution are thus analogous in important respects; constructing trees of languages "genetically" related in families is well established (e.g. Greenhill et al., 2009). Recently, phylogenetic models have enhanced both methodology and hypothesis-testing for language ancestry (e.g. Forster and Renfrew, 2006). Approaches now engage archaeology, anthropology, genetics, and computational science, as well as historical linguistics itself. Notwithstanding advances, disputes remain vigorous in both methods and results, including for well-studied language families such as Indo-European (see, for example, Forster and Renfrew, 2006; Campbell and Poser, 2008). Often, reconstructions are untestable—hence the vigour of disputation. The approach adopted here, by contrast, involves an inspectable set of procedures applied directly to empirical linguistic data. We use analytical methods derived from DNA sequence optimization algorithms, treating words as sequences of sounds. We demonstrate this with Uto-Aztecan (UA) languages of North and Middle America.

The basic approach articulated here is to remove the inferential overburden of hypothesized "proto-forms" (discussed below), and perform analysis solely using the observed sound content of words. In this way, the sequences of sounds that constitute all human languages form the empirical basis upon which language trees are built. To accomplish this, we have adapted techniques more usually applied to the analysis of DNA and protein sequence data, but are readily applied to sound sequences as well (as with other non-molecular sequence data; Schulmeister and Wheeler, 2004; Robillard et al., 2006). In moving from proto-forms to sound sequences, a transition occurs analogous to the advances forged in organismic systematic

*Corresponding author:
E-mail address: wheeler@amnh.org

analysis when hypothesized ancestors were rejected in favour of observation as primary evidence of relationship.

Historical linguistics

"Genetic" classification of languages is based on the comparative method, applied to languages known to be related. Sounds (phonemes), words or word elements (morphemes), and grammatical features are compared to identify regular correspondence patterns (Campbell, 2004). From such patterns analysts infer proto-forms, i.e. the sounds, words, and grammatical characteristics of a protolanguage. For example, English *foot*, Latin *pēs*-(*pedis*), Greek *poús*-(*podós*), and Sanskrit *pāt*-(*padáḥ*) are cognates, similar both in sound (phonologically) and in meaning (semantically), and their phonological differences reflect patterned sound correspondences. Regular sound correspondences and historical shifts are inferred from comparing such cognates, further leading to the reconstruction of a putative ancestral protolanguage. By such reasoning, Proto-Indo-European "foot" is inferred as **pōd*-, or **ped*- (Fortson, 2010). A reconstructed protolanguage is posited as the evolutionary ancestor of the observed descendant languages, and serves as the baseline from which historical shifts are inferred. A protolanguage is regarded as a real language once spoken by a delimited population in a particular time and place, or homeland. From that point of origin, all historically known languages of the family are held to descend. Establishing time and place for the protolanguage is the critical bridge from purely linguistic methods to the archaeological, ethnohistorical, and biological record of past communities. Differentially shared patterns of change from the protolanguage among descendant languages are used to argue for subgroups within the family. Phonologically, for example, medial -c- (a voiceless affricate) is held to be an ancestral UA form (Ramer, 1992a). It is now only present in the south, however; all northern UA languages instead feature medial -y- (a voiced palatal approximant) in cognate phonological environments. The -c-/-y- correspondence identified thus implies historical directionality, and is used as evidence for Northern Uto-Aztecan (NUA) as a genetic (rather than merely geographical) subgroup.

While much work in historical linguistics is rigorous, all of the procedures noted—establishing correspondences, reconstructing proto-forms, inferring directional changes, and identifying protolanguage homelands and dates—involve intuition, guesswork, and arguments from authority. No protolanguage has ever been (nor could be) observed, of course, and, notwithstanding strong claims by some linguists, reconstructions are ipso facto only hypotheses with limited testability. Conflicting analytical results are frequent and rarely subject to resolution. Typically, specialists disagree about both major and minor elements of a reconstruction, sometimes with diametrically opposing inferences of sound correspondence and directional shift (for some UA instances, see Hill, 2008, 2011a). Variation is compounded by the inherently patchy nature of the data: there are no records for many extinct languages (ca. 30 additional UA languages may have been present at European contact; Miller, 1983), which might have served to falsify proposed protolanguage reconstructions. Current hypotheses for Proto-Uto-Aztecan (PUA), for example, depict homeland alternatives at almost opposite ends of the UA spectrum: central Nevada (Merrill et al., 2009) vs. central Mexico (Hill, 2001a), with huge differences (> 7900 BP vs. 4500–3000 BP, respectively) in retrojected dates of initial branching.

The use of glottochronology to reconstruct ancestral language dates is perhaps the most controversial method, depending on an asserted constant rate of loss, 14% per millennium, in basic vocabulary across all languages. Dates of branch splitting on a language tree are calculated based on the percentage of shared basic vocabulary against that absolute rate. While most linguists formally reject glottochronology, dates for specific protolanguages and descendant branch-splitting are still regularly cited, and derived procedures continue to be developed (e.g. Holman et al., 2011)—a paradox. Dates for UA origins resting on such methods remain widely circulated (e.g. Merrill et al., 2009; Hill, 2012). Similarly, identifications of protolanguage homelands vary greatly and are hotly disputed, notably for Indo-European. Methods here are twofold: (i) correlating the greatest number of protolanguage terms for flora and fauna with the highest biogeographical concentrations of the species' ranges in question; or (ii) identifying the region of greatest concentrated diversity among daughter languages, and postulating this as the "centre of gravity" from which all descendant languages radiated outward, the more distal exhibiting lesser group-internal variation than the more proximate. "Linguistic palaeontology" (e.g. Hill, 2012) adds archaeological and biogeographical data to these techniques.

In short, while there are some established methodological conventions, these are often applied with great variation, and there are few settled methods for testing propositions, falsifying or corroborating hypotheses, or conclusively evaluating evidence. Protolanguages are akin to hypothetical ancestors in evolutionary biology, and hence are conclusions of analysis, not assumptions. To create a more evidence-based and objective approach, we have extended and applied techniques used in the phylogenetic analysis of molecu-

lar sequences, and treated the primary data—comparative word lists—as sequences of sounds, eschewing proto-forms entirely.

Classification of UA languages

The UA language family, comprising 40+ known languages, is perhaps the most intensively studied in the Americas. It embraces a panoply of cultural, demographic, and environmental variation. Our comparison specifically includes 37 contemporary, recent, or historical UA languages (Fig. 1). At first European contact, UA languages ranged from Idaho to Panama, including the Columbia Plateau, Great Basin, US Southwest, southern California, Sonoran desert, Sierra Madre Occidental, Valley of Mexico, and Central America (Miller, 1983; see Fig. 2). UA languages were surrounded and interspersed by unrelated families and language isolates (Goddard, 1999). While some languages have become extinct, others (e.g. Hopi, Tohono O'odham/Papago, Tepehuan, Mayo, Rarámuri/Tarahumara, Yaqui, Mayo, Cora, Huichol, and Nahuatl) remain widely spoken (Caballero, 2011). Range biogeography and climate are highly diverse, including semi-arid deserts, mountains, high plains, and marine coastline, along a temperate–neotropical continuum. Sociocultural adaptations have been equally varied, from small-scale foraging bands with minimal technology (such as the Southern Paiute), to the Mexica (Aztec) state—populous, stratified, and intensively agricultural, with advanced arts, architecture, and commerce. Between, variations include small-scale farmers (e.g. Hopi, Yaqui, and Huichol), marine-mammal hunters (Gabrielino, Luiseño), nomadic bison-hunters (Comanche), and (part-time) salmon fishers (Northern Paiute).

The history of UA classification is by no means consistently cumulative or substitutive (see, for example, Lamb, 1964; Hill, 2011a). Proposals set forth in one generation have been shot down in the next, only to re-emerge later. Brinton (1891) suggested three subdivisions: Shoshonean, Sonoran, and Aztecan. Kroeber (1907) divided Shoshonean into four branches: Plateau (=modern Numic—see Fig. 1), California (=Takic), Kern River (=Tübatulabal), and Pueblo (Hopi). Sonoran was divided into three or four subgroups, equivalent to modern Tepiman, Taracahitan, and Corachol (alternatively, Cora and Huichol were kept separate). Whorf (1935) dismissed Shoshonean, suggesting that only smaller subgroups were valid (equivalent to Central Numic etc.). Lamb (1964), also a splitter,

NORTHERN UTO-AZTECAN	NUMIC (Plateau)	WESTERN NUMIC	Northern Paiute
			Western Mono
		CENTRAL NUMIC	Tümpisa (Panamint, Koso)
			Big Smokey Valley Shoshone
			Western Shoshone
			Shoshone
			Comanche
		SOUTHERN NUMIC	Kawaiisu
			Chemehuevi
			Southern Paiute
			Southern Ute
	Tübatulabal (Kern)		Pahkannil (Tübatulabal)
	TAKIC (Southern California)	SERRANO-GABRIELINO	Kitanemuk
			Serrano
			Tongva (Gabrielino)
		CUPAN	Cahuilla
			Cupeño
			Luiseño
			Acjachemen (Juaneño)
	Hopi (Pueblo)		Orayvi Hopi
SOUTHERN UTO-AZTECAN	TEPIMAN (Pimic)		Tohono O'odham (Papago)
			Pima Bajo
			Northern Tepehuan
			Southern Tepehuan
	TARACAHITAN	TARAHUMARAN	Guarijío
			Rarámuri (Tarahumara)
		ÓPATAN	Eudeve
			Ópata
		CAHITAN	Arizona Yoeme (Yaqui)
			Mayo
	Tubar		Tubar
	CORACHOL-AZTECAN	Corachol	Cora
			Wixarika (Huichol)
		Nahua (Aztecan)	Pochutla Mexicano
			Classical Nahuatl
			Tetelcingo Mexicano
			Pipil

Fig. 1. Consensus-classification positions of Uto-Aztecan Languages used in the present analysis.

Fig. 2. Uto-Aztecan languages at first European contact (modified from Miller, 1983).

proposed eight family subgroups, equivalent to Numic, Tübatulabal, Takic, Hopi, Tepiman, Taracahitan, Corachol, and Aztecan; a ninth, Californian "Gemina," was later rejected.

In contrast, Hale (1958), a lumper, suggested Northern (NUA) and Southern (SUA) "sub-stocks," with SUA comprising Aztecan, and NUA all others. Voegelin et al. (1962) sought to reintroduce Brinton's tripartite division (Shoshonean, Sonoran, and Aztecan), and some subsequent scholars followed (see Miller, 1983). Heath (1977) proposed grouping Sonoran and Aztecan as SUA, and renaming Shoshonean as NUA. Campbell and Langacker (1978) agreed, further arguing for Corachol-Aztecan as an SUA subunit (i.e. breaching the earlier Sonoran–Aztecan boundary). Miller (1983, 1984) rejected NUA, recognizing five higher-order groups: Numic, Tübatulabal, Takic, Hopi, and SUA (the last subdivided into Sonoran and

Aztecan). Ramer (1992a) established (for many, at least) NUA as a meaningful genetic unit, but rejected SUA as such (Stubbs, 2011). Ramer's (1992b) NUA comprises three subgroups: Numic, Californian (uniting Tübatulabal with Takic), and Hopi.

A "Consensus Classification" emerged in the 1990s (Fig. 1, identifying individual languages by the names used in our comparison) from three sources: Goddard (1996), Campbell (1997), and Mithun (1999). However, substantive differences exist among the Consensus sources. While Mithun describes northern and southern UA groupings discursively, unlike Campbell, she does not include NUA and SUA as formal genetic groupings. Where Campbell unites Corachol–Aztecan, Mithun separates them; for Campbell, Tubar is a unit within Taracahitic, while Mithun keeps it apart (coordinate with Tepiman, Taracahitic, Corachol, and Aztecan). Also, Fig. 1 does not include some intermediate proposed groupings between columns 3 and 4. Notably, Mithun divides Southern Numic into two coordinate subgroups: Ute (comprising Chemehuevi, Southern Paiute, and Southern Ute) and Kawaiisu; by contrast, Campbell has three subgroups: Southern Paiute, Ute-Chemehuevi, and Kawaiisu. Within Nahua, both Campbell and Mithun separate Pochutec/Pochutla from the remainder as a distinct language. Merrill (2013) argues for the genetic unity of SUA.

Some remain reluctant to accept higher-order groupings. Caballero (2011), citing persistent controversy over subgroups, opts for only two ranks: individual languages, and one ascendant level into eight branches (as in Fig. 1 column 2, except for Tubar, placed within Taracahitan, and Corachol and Aztecan, kept separate). The *World Atlas of Language Structures* (WALS, 2013) divides only by "genus:" Aztecan, Cahita, Corachol, Hopi, Numic, Takic, Tarahumaran, Tepiman, Tubar, and Tübatulabal—an alphabetical, not geographical, listing—and individual languages within each. This listing does not recognize NUA, SUA, Taracahitan, or Corachol-Aztecan. A tree developed using the Automated Similarity Judgment Program (ASJP; Holman et al., 2011; Hill, 2011a) shows NUA comprising two subgroups (Hopi and "Numic-Californian") and SUA comprising Sonoran and Aztecan divisions. Hill (2011a), responding to the ASJP tree, instead identifies seven coordinate subgroups, based on traditional methods: NUA (subdivided into Numic, Hopi, and Californian), Tepiman, Cahitan, Ópata-Eudeve, Tarahumara-Guarijío, Tubar, and Corachol-Aztecan. She casts doubt on the validity of SUA, Sonoran, and Taracahitan.

With all this dissensus, the only UA tree accepted by most is a shallow bush, with little resolution (Davletshin, 2012). For such reasons, Miller (1983, 1984) and others have proposed a wave (rather than tree) model of dialect chains, subject to continuous inter-language horizontal change, in effect superseding descent with modification. We suggest that the multiple conflicting classifications and weak resolution of branches call for more precise methods.

Proposed geographical and temporal origins of UA languages

Most analyses favour a PUA homeland in the north among pre-horticultural foragers (for specific citations, see Hill, 2001a, 2012). Earlier suggestions include the Gila River highlands of southern New Mexico or the northern Sierra Madre of Sonora–Chihuahua, the Columbia River Plateau, the eastern border of California and Oregon, and a region centring on the four corners of New Mexico, Arizona, Sonora, and Chihuahua. The last rests on Fowler's (1983) collocation of animal and plant species terms in reconstructed PUA with the known biogeographical concentration of those species. However, the species' ranges in question also extend southward into central Mexico (Hill, 2001a). And Davletshin (2012) questions whether Fowler's analysis is falsifiable.

Recent hypotheses offer polar opposites for PUA homelands and trajectories of diversification, with opposing implications too for demographic and sociocultural history. The "farmer hypothesis" (in particular, Hill, 2001a, 2011a)—central also to recent debates over Indo-European—suggests the PUA homeland lies in the agricultural heart of Mesoamerica, and was contemporary with the domestication of maize, beans, and squash. From there, maize-bearing UA demes spread north, some losing agricultural words after shifting to a foraging adaptation (e.g. in the Great Basin and California). The principal proponent, Jane Hill, from a lexical comparison of terms for cultivars and agricultural techniques, argues for a PUA origin at ca. 5600–4500 BP (Hill, 2001a). Following demographic outspread, she holds, a dialect chain appeared ca. 4500–3500 BP, forming five distinct branches by 2500 BP: Proto-NUA, Proto-Tepiman, Proto-Taracahitan, Proto-Tubar, and Proto-Corachol-Aztecan [Proto-Taracahitan is subverted by her later questioning of Taracahitan (Hill, 2011a), however]. Hill's hypothesis upended all northern origin proposals, which identify agriculture as a later adoption by southward-spreading UA demes.

Hill's argument received support from some UA linguists, but strong opposition from others. Alternative reconstructions of PUA agricultural vocabulary (e.g. Campbell and Poser, 2008; Merrill et al., 2009) sharply differ from Hill's, albeit with the same basic linguistic data. Campbell and Poser (2008) re-affirmed Fowler's suggested PUA homeland among foragers in the southern Southwest/northern Mexico. Hill (2012) has

defended her hypothesis, however, offering new evidence of maize vocabulary from (non-agricultural) California UA languages and Comanche, and a new reconstruction of SUA pottery lexicon, suggesting PUA agricultural terms were originally borrowed from adjacent Otomanguean languages of central Mexico.

Hill's arguments have generated substantial interest among archaeologists and biological anthropologists seeking to explain agriculturally based transformations in late Archaic and early Formative archaeological cultures of the Greater Southwest (e.g. Kemp et al., 2010; Watson, 2010). Analysis of mitochondrial DNA data (Kemp et al., 2010) has not corroborated demic diffusion from Mesoamerica to the Southwest, however. Substantial areal variation in haplotype frequencies suggests genetic commonalities for southern UA-speakers with other Mesoamericans, and for northern UA-speakers with other Southwestern Native populations. Kemp et al. (2010) do not rule out northward migrations by UA agriculturalists, but suggest this may have been restricted to males, a pattern argued for migrations elsewhere in aboriginal America.

Directly challenging the farmer hypothesis, an argument by Merrill et al. (2009) (see also Merrill, 2012) focuses on archaeological and palaeoecological data to re-assert a northern PUA homeland, but in a new place, central Nevada. This argument also differs radically on dating PUA. While Hill (above) proposes a PUA date range of 5600–4500 BP, Merrill et al. (2009) suggest an origin at 8900 BP, breaking into Proto-NUA and Proto-SUA (PNUA and PSUA) ca. 7900 BP. Using ASJP (an automated statistical method), those dates have been rejected and an alternative proposed at 4018 BP (Holman et al., 2011). From the total UA historical range, ASJP has also been used to approximate the PUA homeland centre based on purely quantitative statistical grounds, placing it at 27.50°N by 110.25°W, in coastal Sonora south-east of Guaymas, modern Cahitan country (Wichmann et al., 2010).

In summary, after more than a century of sustained research, major disagreements remain in key aspects of UA historical linguistics: classification and internal branching patterns, origins in time and space, and correlations with demography and cultural adaptation.

Words as sound sequences

The use of basic lexicon for linguistic comparison has numerous precedents. The Swadesh list of 100 words generally most resistant to change across languages (or components of this list) has been used in several prior UA analyses (Hale, 1958; Miller, 1984; Cortina-Borja and Valiñas-Coalla, 1989; Holman et al., 2011). Our approach shares in part with these, but is distinguished by its direct attention to words as sound sequences, and by not reducing cognates into synthesized, pre-coded correspondences. Our method is agnostic on hitherto-identified sound correspondences, phonological shifts, or directional changes among languages. Neither do we here treat morphological or grammatical structures (we plan to address these in future). We focus rather on a discrete set of lexical data subjected to strict sequence-by-sequence comparison. We make no assumptions about protoforms at any genetic level (PUA, PNUA, P-Numic, etc.), and we explicitly exclude protolanguage terms found in the literature (variable in any event) from our analysis, which is restricted to empirically attested words in the described languages themselves. Note that our approach is based on the reconstruction of hypothetical ancestral sound sequences (i.e. hypothetical proto-words) at internal tree nodes, such that overall change though the tree is minimized. This is in opposition to similarity-based clustering and distance techniques (e.g. Cortina-Borja and Valiñas-Coalla, 1989), which maximize overall similarity and make no attempt at hypothetical ancestral word reconstruction.

Our basic data source is Miller's UA Cognate Sets, revised by Kenneth Hill (K. Hill, 2011b; version of May 2011).[1] Miller compiled UA cognates from ethnographic and linguistic records (Miller, 1967). Kenneth Hill transferred Miller's lists into word-processed format, adding extensive supplements. Of signal import, Hill, a long-term UA specialist, added renderings in IPA (the International Phonetic Alphabet). Linguistic comparison has frequently been hampered by lack of a standard orthography. UA languages have been transcribed with numerous different conventions, and sometimes re-transcribed in simplified form by later investigators, who thus add an unseen layer of phonological interpretation. IPA renderings of original linguistic records allow words to be rigorously compared as sequences of sounds.

Each word is treated as a separate sound sequence, homologous in all our study languages (including outgroups), in a manner analogous to the genes of a molecular biological study. The most important difference is the large sound alphabet size (148) as compared with that of DNA (4) or protein (20) sequences. For a given tree of language relationships, median

[1] Our data entry and preliminary analysis were conducted in 2011. Since then, an additional source on UA cognates, expanding Miller's lists substantially, has been published by Stubbs (2011). While Stubbs gives aggregated lists of cognates, Kenneth Hill's presentation of data by languages and sources remains key for our purpose, and his inclusion of International Phonetic Alphabet (IPA) transcriptions is indispensable. The ASJP data set (http://email.eva.mpg.de/~wichmann/languages.htm) used in the analysis of 40 UA words (Holman et al., 2011; Hill, 2011a) gives phonologically simplified representations. For our data set, see Supplementary Materials.

sound sequences were created at each tree vertex (representing hypothetical ancestral words) such that the overall weighted number of sound changes between ancestor and descendant sequences (changes in sounds, insertion and deletion of sounds) was minimized (Fig. 3). This is known in the computer science literature as the Generalized Tree Alignment Problem (GTAP; Sankoff, 1975) and has been shown to be an NP (non-deterministic polynomial-time)-hard optimization (Wang and Jiang, 1994). Given the intractability of identifying exact solutions, we used a series of heuristic approaches (see below; Wheeler, 1996, 1999, 2003; Varón and Wheeler, 2012, 2013) implemented in the program POY (Varón et al., 2010; Wheeler et al., 2013). From this analysis both language evolutionary trees and hypothetical proto-forms were created (see Supplementary Materials).

Data set

We used the Swadesh-100 word-list (Swadesh, 1971), i.e. of words found empirically to be most resistant to change over time. Miller (1984) used a similar list for 32 languages and dialects, but his comparison depended on symbolic reductions ("a," "b," "c," etc.) of words adjudged as cognate rather than systematic representations of the words as sound-strings themselves. Cortina-Borja and Valiñas-Coalla (1989) used a combination of statistical methods, including lexical distance matrices, principally on Miller's data set, as already precoded by Miller. The ASJP 40-word data set (http://email.eva.mpg.de/~wichmann/languages.htm) gives simplified representations.

Our word lists were compiled for 37 well-described UA languages, drawing principally on Miller's Uto-Aztecan Cognate Sets (Miller, 1967, 1988), as revised and expanded by Hill (2011b). To accommodate many of the lexical gaps in the Sets, we turned to the original sources (for specific sources, see Notes on Entries in Supplementary Materials). Approximately 75% of UA words in our 100 (102) word data set are taken directly from the Sets; 25% (ca. 1070 words) are from our own additions. Kenneth Hill's version of the Sets adds renderings in IPA. Our data additions were rendered into IPA, adhering strictly to Hill's conventions for the same sources.

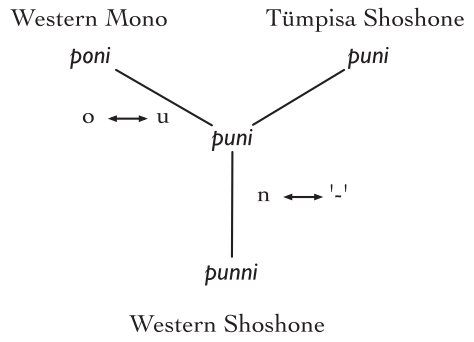

Fig. 3. Median optimization of "to see" (rendered in IPA) in three different languages. The median is created such that the overall cost of sound changes (here one substitution and one deletion) is minimized.

All words in the data set were rendered into LATEX TIPA 1.3 (Rei, 2004). Swadesh-100 word-lists were added for three non-UA outgroups: Ipai, Tewa, and Zuni (for sources, see Notes on Entries in Supplementary Materials). Outgroup words were rendered into IPA following Hill's conventions as far as possible. Ipai is one of the Cochimi–Yuman languages of southern California, which intrude geographically between NUA and SUA languages (Goddard, 1999). Tewa is a Kiowa–Tanoan Pueblo language; suggested supra-family connections between UA and Kiowa–Tanoan remain tentative (e.g. Davis, 1989; Hill, 2002). Zuni, a Western Pueblo language, is an isolate. Where alternative cognates are present in the data sources, the one that was semantically closest to the basic lexicon word of the Swadesh list was used. In total, 148 symbols were required for all UA and non-UA languages: 38 consonants, 99 vowels (nine base vowels with stress, tone, nasalization, etc., markers), 10 diacritics, and a syllable-break.

Phylogenetic analysis

All phylogenetic analyses were performed with POY5 (Wheeler et al., 2013). Given the unknowable nature of relative sound transformation cost, we examined several scenarios in a sensitivity analysis (Wheeler, 1995) context. Five sound graphs were constructed. In each case, the sounds (148 in our data set) constituted the vertex set with edges connecting them based on various notions of substitution propinquity. These were (i) all changes in sound were directly linked (equally costly; "1-1") including gain and loss of sounds (edit distance), (ii) distinction was made between vowel and consonant sounds such that all intra-vowel, intra-consonant, and gain–loss transformations were linked directly (equal cost), but transformations between vowels and consonants were twice as costly ("vcd"), (iii) same as the previous, but with gain–loss cost equal to vowel-consonant ("vcd2"), (iv) sounds were differentiated based on vowel/consonant, articulation, voicing (consonants), rounding (vowels), and stress/tone ("all"), and (v) sounds were linked by chains of single articulation changes ("graph"). In scenarios (iv) and (v), gain–loss cost was set to be equal to the maximum change between sounds (ensuring metricity; Wheeler, 1993).

Each of these transformation cost scenarios was subjected to 1000 replications of random addition-sequence Wagner builds+TBR branch swapping. Cladogram diagnosis and median sound sequence construction were accomplished initially using direct optimization (DO; Wheeler, 1996) and refined with Iterative-Pass optimization (IP; Wheeler, 2003). These five initial results were then pooled and used as starting points for repeated rounds of recombinations and TBR swapping in a genetical algorithm approach (Holland, 1975; Goloboff, 1999; Moilanen, 1999). Resulting trees were repeatedly cycled until results stabilized for all analytical parameters (as in Schuh et al., 2009).

To examine both the robustness of these results and their overall support, consensus trees were constructed for the runs. Individual and consensus topologies for the resulting trees are shown in Fig. 4. The overall "best" transformation cost-tree was identified using the RILD (Wheeler and Hayashi, 1998) and MRI (Wheeler et al., 2006) congruence measures. Both these measures point to scenario "1-1" as optimal (Table 1). Bremer support values (Goodman et al., 1982; Bremer, 1994) were calculated based on the TBR-neighbourhood of the "1-1" tree. Jackknife supports were calculated based on 256 replicated delete $1/e$ word samples (and hence, each replicate is based on approximately 65 words). In each replicate, a single Wagner tree was constructed and refined with TBR branch swapping (Fig. 5).

Language tree topology

The heuristically optimal scheme of UA languages based on sound-sequence analysis shows many groups commonly recognized as well as some novel patterns. Figure 5 is based on the all-equal ("1-1") sound transformation scenario. At the highest level, NUA is supported as a monophyletic group, but SUA is not: rather it is paraphyletic, branching along the old Sonoran and Aztecan (Nahua) lines (contra Merrill, 2013). The Nahua clade has strong support. Its appearance at the base of the tree is perhaps especially noteworthy.

Within NUA, Numic groups very strongly (and universally over transformation cost scenarios), with good support for the Central and Southern subdivisions, but not Western. Northern Paiute and Western Mono form a grade (in the best and two other cost scenarios); Northern Paiute appears as the sister taxon to Western Mono, Central Numic, and Southern Numic. Our analysis thus does not support an argument that Central Numic is transitional between Western and Southern Numic (Cortina-Borja and Valiñas-Coalla, 1989), and we do not affirm the ASJP tree that postulates a Northern Paiute–Southern Paiute clade (a linguistically implausible propinquity; see Hill, 2011a). The Takic subgroups have long been understood as quite distant from each other (Miller, 1984; Hill, 2012); this is confirmed by relative branch lengths on our tree. Within Takic, Cupan receives strong support, but unlike the Consensus Classification, Kitanemuk–Serrano shows a clear break from Tongva (Gabrielino).

The appearance of Tübatulabal and Hopi as sister to Takic and Numic is most interesting. Both languages cease to appear as so isolated in the total NUA array as in prior analyses. Hopi is the only NUA ethnolinguistic group with a long-term (prehistoric) dependence on maize–beans–squash agriculture. Hopi's position proximate to the SUA–NUA divide is also suggestive in this regard. The position of both Takic and Numic in more derived locations than Hopi may corroborate the idea that non-agricultural NUA demes developed later than ancestral agricultural demes migrating from the south. Given the equivalent proximity of Hopi–Tübatulabal to both Takic and Numic, the "Greater Takic" grouping (Hill, 2001b), comprising Takic with Tübatulabal and Hopi, is a basal NUA grade in our analysis. A "Californian" clade uniting Takic with Tübatulabal alone is unsupported.

Among the southern languages, Tepiman clusters and subdivides predictably on our tree, as do Taracahitan, Corachol, and Nahua. Tubar's relatively greater proximity to Corachol than to Taracahitan disconfirms Campbell (1997) and Caballero (2011) that Tubar belongs within Taracahitan, and supports the argument of Hill (2011a) that there are no grounds to re-establish a Sonoran (including Tubar) exclusive of Corachol–Aztecan. The relative proximity of the Tubar and Corachol branches on our tree calls for further investigation. The separation of Corachol from Nahua affirms the analysis of Holman et al. (2011), disconfirming Jane Hill's position (Hill, 2011a) (and that of earlier analysts) on their unity, and may confirm the conclusion of Kaufman (2001) that similarities between Corachol and Nahua are effects of relatively recent contact, rather than indications of long-term shared evolution (Kaufman also infers a long-term *in situ* presence for Cora, treating Nahua as a later arrival in central Mexico).

In our analysis, the basal position of Hopi within NUA suggests an intermediate position in certain respects, reflecting a long-pondered argument (Miller, 1984; Cortina-Borja and Valiñas-Coalla, 1989). Kroeber (1907) contended that Hopi was the most divergent of all NUA (his "Shoshonean") languages. Miller maintained that terms for maize cultivation were grafted onto Hopi from parts unknown (Hill, 2001a); Hill expands on this (Hill, 2001a), and Merrill (2012) offers some lexical confirmation from SUA. Available

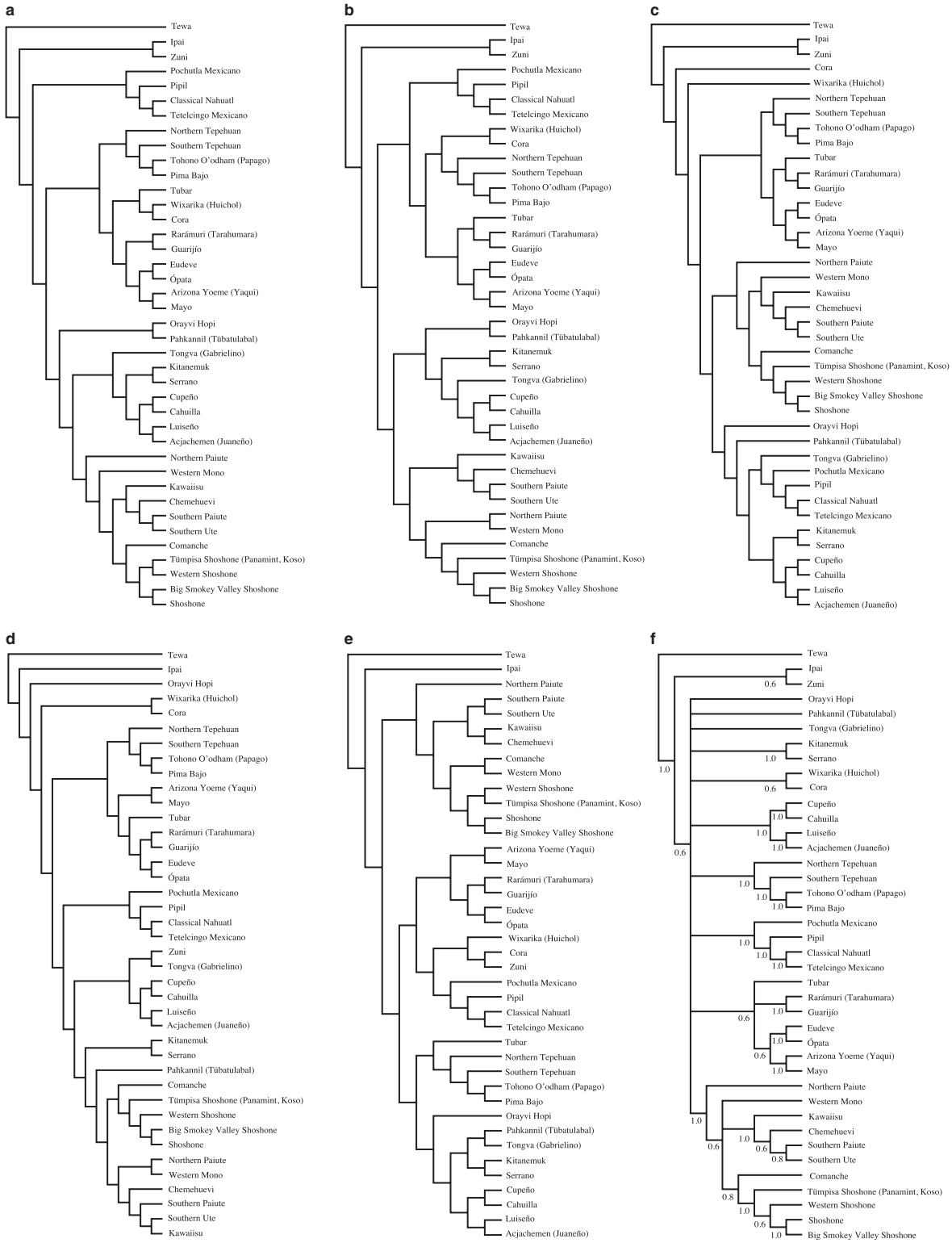

Fig. 4. Cladograms of analyses where: (a) all changes were equally costly (edit distance; "1-1"), (b) all intra-vowel, intra-consonant, and gain–loss transformation were equally costly, but transformations between vowels were twice as costly ("vcd"), (c) same as 'b' but with gain–loss cost equal to the vowel-consonant cost ("vcd2"), (d) costs were based on vowel/consonant, articulation, voicing (consonants), rounding (vowels), and stress/tone ("all"), (e) sound transformation costs were determined by chains of single articulation changes ("graph"), and (f) the fraction of analyses with each group (if ≥ 0.50) is shown on each branch.

Table 1
Sensitivity of analyses to transformational model

Model	Tree cost	Minimum cost	Maximum cost	RILD	MRI
1-1	10 677	8457	15 370	0.2079	0.3211
vcd	11 763	9135	16 426	0.2234	0.3604
vcd2	16 036	11 998	21 887	0.2518	0.4083
all	34 437	18 244	45 958	0.4702	0.5842
graph	61 218	41 426	86 741	0.3233	0.4368

ethnohistorical evidence is surely valuable here. Hopi accounts of their origins posit the in-migration of multiple clans from all directions to Tuuwanasavi, the "earth-center place" (e.g. Whiteley, 2011). These migrations, however, are apportioned among two major demic sources associated, respectively, with northern and southern UA geographical regions. Autochthonous Hopi clans site aboriginal emergence from the Grand Canyon at *sipàapuni*, an "earth-navel." Clans that migrated from the south rather trace their origin to Palatkwapi, a legendary settlement (perhaps in the Salt River valley, Whiteley, 2011; see also Ferguson and Colwell-Chanthaphonh, 2006; Merrill, 2012) abandoned after a flood. Tree propinquities in our analysis suggest Hopi has more southern features than other NUA languages. Hopi is exceptional

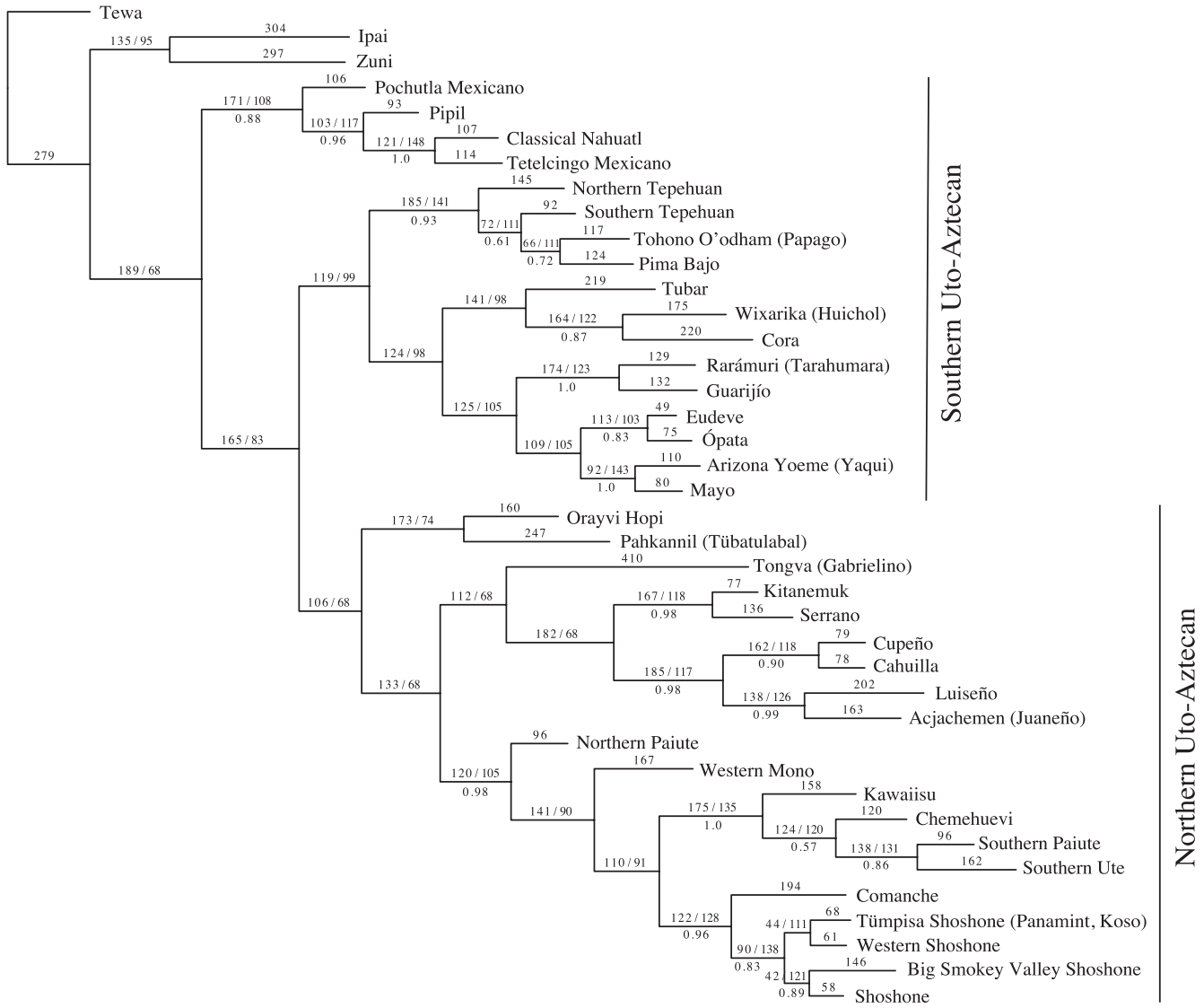

Fig. 5. Phylogram of Uto-Aztecan language relationships showing Southern (paraphyletic) and Northern (monophyletic) Uto-Aztecan groups. Branch lengths are shown (left above branches) and are proportional to (and labelled with) cost differences between single assignment word medians. Other branch lengths are possible based on alternative single assignments derived from potentially multiple word-medians. Bremer (right above internal branches) and jackknife supports (if ≥ 0.50, below branches) are shown. The total cost of the cladogram is 10 677 with all sound transformations (including insertion and deletion of sounds) costing one step (transformational cost matrix "1-1").

for its rich ethnographic and ethnohistorical record, and it would be optimal to include comparable data from other cases. Nonetheless, we believe the somewhat intermediate position of Hopi may be better explained by aggregate demic origins than by projecting long isolation in a putative PNUA homeland, with sporadic interspersion of agricultural vocabulary.

Maize cultivation and historical migration

The position of Nahua at the base of our tree is perhaps the most interesting of our results. Nahua is geographically near the PUA homeland nominated by Jane Hill (2001a), and close to early sites of crop domestication. Holocene domestication of maize (*Zea mays*), squash (*Cucurbita* spp.), and the common bean (*Phaseolus vulgaris*) has recently been sited in and near the Central Balsas River Valley of central Mexico (Ranere et al., 2009; Bitocchi et al., 2012), i.e. close by historical Nahua populations. Nahuatl myths depict Aztlan/Chicomoztoc, located in northern Mexico, as their homeland before migration to the Valley of Mexico (Smith, 2012). Most scholars have attributed some historical validity to the myths (e.g. Beekman and Christensen, 2003; Watson, 2010). Hill (2001a), however, warns against treating the myths as history, and thus sees this as no barrier to Nahua geographical proximity to her proposed PUA homeland (see also Hill, 2012). Our tree tends to support Hill's position, and may suggest Nahuatl migrations were not unidirectional, and followed an earlier proto-Nahua agricultural presence in Central Mexico.

The separation of geographically proximate Corachol and Nahua on our tree corresponds with the centre of gravity principle that identifies most current diversification as reflecting the area of protolanguage origin. A PUA homeland in or near the area occupied by historical Cora and some Nahua is suggested by our tree. Cora population history (Spicer, 1969) suggests long *in situ* presence from the western Sierra Madre to the Pacific coast (in modern Nayarit and Jalisco), with large communities (including Totorame) in the subtropical lowlands at the time of Spanish conquest. If Cora historical territory truly represents a deep occupation and Nahua speakers were later migrants from the north, PUA speakers may not have been the original domesticators of maize, beans, and squash. Rather, PUA speakers lived close by the sites of early domestication, and would have adopted agriculture as domesticates radiated northward into their area. Alternatively, if some Nahua demes were indeed present in central Mexico early on, they may have been directly associated with domestication, or indirectly, via proximity to Otomanguean speakers (following Hill, 2012). Figure 6 shows our tree mapped onto the landscape of UA language distribution.

Our analysis shows it is unlikely that the historical Numic area represents the PUA homeland, and suggests UA speakers migrated into this area from the south, either as foragers prior to the southerly domestication of crops, or as (male?) agriculturists who lost the art after the Great Basin became inhospitable to farming in the 13th century CE (Madsden, 1989) (Fig. 5). The intermediate status of Hopi, long-term agriculturalists, between SUA and NUA may corroborate

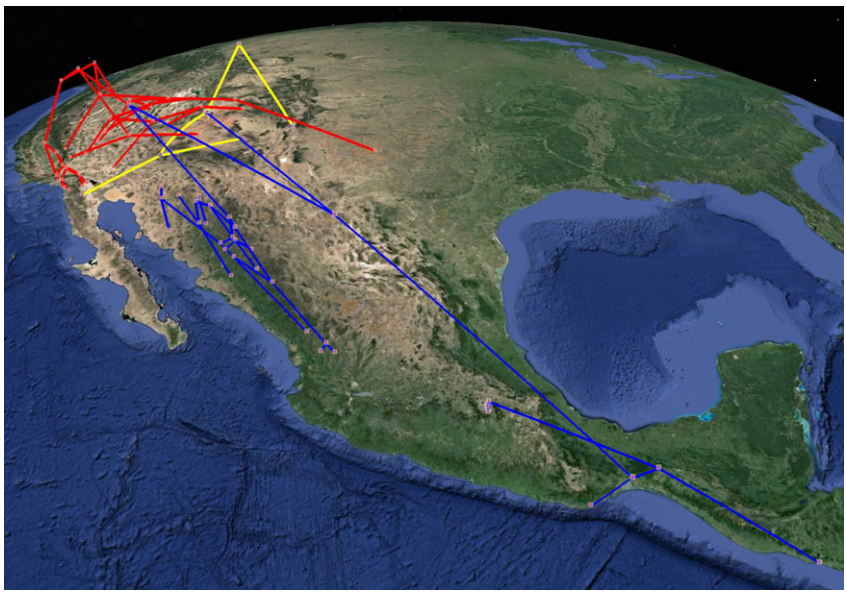

Fig. 6. Uto-Aztecan language tree showing southern origins of UA languages plotted using Supramap (Janies et al., 2007). Outgroup edges are yellow, SUA edges blue, and NUA edges red.

original northward migration of already agricultural UA demes. We remain agnostic about dates of origin or branching. Our analysis lends no support to retrodicted origin or branching-event dates via glottochronological methods that assume uniform rates of change in basic vocabulary over time, which would imply clearly absent ultrametric distances. Interestingly, individual branches exhibit low variance in length overall, perhaps due to the carefully selected nature of the Swadesh words.

Conclusions

Our method avoids the inference-laden approach used in most historical–linguistic analysis that treats a hypothetical form, PUA, as a fact from which subsequent conclusions are drawn. We offer finely measurable data for establishing relationships among languages. Although we do not include morphological or grammatical information that would enhance a total comparison, our analysis of basic lexicon cognates as sound sequences presents a directly empirical test of language groups and of the biogeographical and cultural changes their relationships imply. Our method provides a test of existing hypotheses for UA homeland origins, clearly favouring a southern origin model. The techniques we develop here are generally applicable to the evolution of written and unwritten languages, and, we predict, will result in more empirically robust hypotheses of language relationships and linguistic evolution.

Acknowledgements

We would like to acknowledge the manuscript comments of Ronald Clouse, Louise Crowley, John Denton, Daniel Janies, Prashant Sharma, and two anonymous reviewers. Steven Thurston aided greatly in the production of artwork. We thank Kenneth C. Hill for sharing his revised and expanded version of Miller's Uto-Aztecan Cognate Sets. This material is based upon work supported by, or in part by, the National Science Foundation (BCS-0925978), the US Army Research Laboratory and the US Army Research Office under contract/grant number W911NF-05-1-0271.

References

Beekman, C., Christensen, A., 2003. Controlling for doubt and uncertainty through multiple lines of evidence: a new look at the Mesoamerican Nahua migrations. J. Archaeol. Method Th. 10, 111–164.

Bitocchi, E., Nanni, L., Bellucci, E., Rossi, M., Giardini, A., Zeuli, P.S., Logozzo, G., Stougaard, J., McClean, P., Attene, G., Papa, R., 2012. Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. Proc. Natl Acad. Sci. USA 109, E788–E796.

Bremer, K., 1994. Branch support and tree stability. Cladistics 10, 295–304.

Brinton, D., 1891. The American Race: A Linguistic Classification and Ethnographic Description of the Native Tribes of North and South America. N.D.C. Hodges, New York, NY.

Caballero, G., 2011. Behind the Mexican mountains: recent developments and new directions in research on Uto-Aztecan languages. Lang. Linguist. Compass 5, 485–504.

Campbell, L., 1997. American Indian Languages: The Historical Linguistics of Native America. Oxford University Press, New York, NY.

Campbell, L., 2004. Historical Linguistics: An Introduction. MIT Press, Cambridge, MA.

Campbell, L., Langacker, R., 1978. Proto-Aztecan vowels. Int. J. Am. Linguist. 44, 85–102, 197–210, 262–279.

Campbell, L., Poser, W., 2008. Language Classification: History and Method. Cambridge University Press, New York, NY.

Cortina-Borja, M., Valiñas-Coalla, L., 1989. Some remarks on Uto-Aztecan classification. Int. J. Am. Linguist. 55, 214–239.

Davis, I., 1989. A new look at Aztec-Tanoan. In: Key, M., Hoenigswald, H. (Eds.), General and Amerindian Ethnolinguistics: In Remembrance of Stanley Newman. Mouton de Gruyter, Berlin, pp. 365–379.

Davletshin, A., 2012. Proto-Uto-Aztecans on their way to the Proto-Aztecan homeland: linguistic evidence. J. Lang. Relationship 8, 75–92.

Ferguson, T., Colwell-Chanthaphonh, C., 2006. History is in the Land: Multivocal Tribal Traditions in Arizona's San Pedro Valley. University of Arizona Press, Tucson.

Forster, P., Renfrew, C. (Eds.) 2006. Phylogenetic Methods and the Prehistory of Languages. MIT Press, Cambridge, MA.

Fortson, B., 2010. Indo-European Language and Culture: An Introduction, 2nd edn. Blackwell, Malden, MA.

Fowler, C., 1983. Lexical clues to Uto-Aztecan prehistory. Int. J. Am. Linguist. 49, 224–257.

Goddard, I., 1996. Handbook of North American Indians, Vol. 6, Languages. Smithsonian Institution, Washington, DC.

Goddard, I., 1999. Native Languages and Language Families of North America (Map). University of Nebraska Press, Lincoln, NE, revised and enlarged edition.

Goloboff, P.A., 1999. NONA (No Name) ver. 2. Published by the author, Tucumán, Argentina.

Goodman, M., Olson, C.B., Beeber, J.E., Czelusniak, J., 1982. New perspectives in the molecular biological analysis of mammalian phylogeny. Acta Zoologica Fennica 169, 19–35.

Greenhill, S., Currie, T., Gray, R., 2009. Does horizontal transmission invalidate cultural phylogenies? Proc. Biol. Sci. 276, 2299–2306.

Hale, K., 1958. Internal diversity in Uto-Aztecan. Int. J. Am. Linguist. 24, 101–107.

Heath, J., 1977. Uto-Aztecan morphophonemics. Int. J. Am. Linguist. 43, 27–36.

Hill, J., 2001a. Proto-Uto-Aztecan: a community of cultivators in Central Mexico? Am. Anthropol. 103, 913–934.

Hill, K.C., 2001b. Comments on Hopi and comparative Uto-Aztecan. In: Zamarrón, J., Hill, J. (Eds.), Avances y balances de lenguas yutoaztecas. Instituto Nacional de Antropología e Historia, Mexico City, pp. 313–343.

Hill, J., 2002. Toward a linguistic prehistory of the Southwest: "Azteco-Tanoan" and the arrival of maize cultivation. J. Anthropol. Res. 58, 457–475.

Hill, J., 2008. Northern Uto-Aztecan and Kiowa–Tanoan: evidence of contact between the proto-languages? Int. J. Am. Linguist. 74, 155–188.

Hill, J., 2011a. Subgrouping in Uto-Aztecan. Lang. Dyn. Change 1, 241–278.

Hill, K.C., 2011b. Wick Miller's Uto-Aztecan cognate sets, revised and expanded by Kenneth C. Hill.

Hill, J., 2012. Proto-Uto-Aztecan as a Mesoamerican language. Ancient Mesoamerica 23, 57–68.
Holland, J.H. (Ed.) 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI.
Holman, E., Brown, C., Wichmann, S., Müller, A., Velupillai, V., Hammarström, H., Sauppe, S., Jung, H., Bakker, D., Brown, P., Belyaev, O., Urban, M., Mailhammer, R., List, J., Egorov, D., 2011. Automated dating of the world's language families based on lexical similarity. Curr. Anthropol. 52, 841–875.
Janies, D., Hill, A.W., Guralnick, R., Habib, F., Waltari, E., Wheeler, W.C., 2007. Genomic analysis and geographic visualization of the spread of avian influenza (h5n1). Syst. Biol. 56, 321–329.
Kaufman, T., 2001. The History of the Nawa Language Group from the Earliest Times to the 16th Century: Some Initial Results. http://www.albany.edu/pdlma/Nawa.pdf.
Kemp, B., González-Oliver, A., Malhi, R., Monroe, C., Schroeder, K., McDonough, J., Rhett, G., Resendéz, A., Peñaloza-Espinosa, R., Buentello-Malo, L., Gorodesky, C., Smith, D., 2010. Ultraconserved words point to deep language ancestry across Eurasia. Proc. Natl Acad. Sci. 107, 6759–6764.
Kroeber, A.L., 1907. Shoshonean dialects of California. Univ. Calif. Publ. Am. Archaeol. Ethnol. 4, 66–165.
Lamb, S., 1964. The classification of the Uto-Aztecan languages: a historical survey. In: Bright, W. (Ed.), Studies in California Linguistics, University of California Publications in Linguistics 34. University of California Press, Berkeley, CA, pp. 106–125.
Madsden, D., 1989. Exploring the Fremont. Utah Museum of Natural History, Salt Lake City, UT.
Merrill, W., 2012. The historical linguistics of Uto-Aztecan agriculture. Anthropol. Linguist. 54, 203–260.
Merrill, W., 2013. The genetic unity of southern Uto-Aztecan. Lang. Dyn. Change 3, 68–104.
Merrill, W., Hard, R., Mabry, J., Fritz, G., Adams, K., Roney, J., MacWilliams, A., 2009. The diffusion of maize to the Southwestern United States and its impact. Proc. Natl Acad. Sci. 106, 21019–21026.
Miller, W.R., 1967. Uto-Aztecan Cognate Sets. Number 48 in University of California Publications in Linguistics. University of California Press, Berkeley, CA.
Miller, W., 1983. Uto-Aztecan languages. In: Ortiz, A. (Ed.), Handbook of North American Indians. Smithsonian Institution, Washington, DC, Vol. 10, pp. 113–124.
Miller, W., 1984. The classification of the Uto-Aztecan languages based on lexical evidence. Int. J. Am. Linguist. 50, 1–24.
Miller, W., 1988. Computerized Data Base for Uto-Aztecan Cognate Sets. Department of Linguistics, University of Utah, Salt Lake City, UT.
Mithun, M., 1999. The Languages of Native North America. Cambridge University Press, Cambridge.
Moilanen, A., 1999. Searching for most parsimonious trees with simulated evolutionary optimization. Cladistics 15, 39–50.
Ramer, A.M., 1992a. A Northern Uto-Aztecan sound law: *-c→-y-. Int. J. Am. Linguist. 58, 251–268.
Ramer, A.M., 1992b. Tubatulabal 'Man' and the subclassification of Uto-Aztecan. Calif. Linguist. Notes 23, 30–31.
Ranere, A., Piperno, D., Holst, I., Dickau, R., Iriarte, J., 2009. The cultural and chronological context of early Holocene maize and squash domestication in the Central Balsas River Valley, Mexico. Proc. Natl Acad. Sci. USA 106, 5014–5018.
Rei, F., 2004. Tipa Manual, Version 1.3. Graduate School of Humanities and Sociology, University of Tokyo, Tokyo.
Robillard, T., Legendre, F., Desutter-Grandcolas, L., Grandcolas, P., 2006. Phylogenetic analysis and alignment of behavioral sequences by direct optimization. Cladistics 22, 602–633.
Sankoff, D.M., 1975. Minimal mutation trees of sequences. SIAM J. Appl. Math. 28, 35–42.
Schuh, R.T., Weirauch, C., Wheeler, W., 2009. Phylogenetic analysis of family-group relationships in the Cimicomorpha (Hemiptera). Syst. Entomol. 34, 15–48.
Schulmeister, S., Wheeler, W.C., 2004. Comparative and phylogenetic analysis of developmental sequences. Evol. Dev. 6, 50–57.
Smith, M.E., 2012. The Aztecs, 3rd edn. Wiley-Blackwell, Hoboken, NJ.
Spicer, E.M., 1969. Northwest Mexico: introduction. In: Vogt, E.Z. (Ed.), Handbook of Middle American Indians. University of Texas Press, Austin, TX, Vol. 8, pp. 777–791.
Stubbs, B., 2011. Uto-Aztecan: A Comparative Vocabulary. Shumway Family History Services, Flower Mound, TX.
Swadesh, M., 1971. The Origin and Diversification of Language. Aldine, Chicago, IL.
Varón, A., Wheeler, W.C., 2012. The tree-alignment problem. BMC Bioinformatics 13, 293.
Varón, A., Wheeler, W.C., 2013. Local search for the generalized tree alignment problem. BMC Bioinformatics 14, 66.
Varón, A., Vinh, L.S., Wheeler, W.C., 2010. POY version 4: phylogenetic analysis using dynamic homologies. Cladistics 26, 72–85.
Voegelin, C., Voegelin, F., Hale, K., 1962. Typological and Comparative Grammar of Uto-Aztecan: I (Phonology). International Journal of American Linguistics Memoirs, 17, Bloomington, IN.
WALS, 2013. World Atlas of Language Structures. http://wals.info/.
Wang, L., Jiang, T., 1994. On the complexity of multiple sequence alignment. J. Comput. Biol. 1, 337–348.
Watson, J., 2010. The introduction of agriculture and the foundation of biological diversity in the southern Southwest. In: Auerbach, B. (Ed.), Human Variation in the Americas. Center for Archaeological Investigations, Southern Illinois University, Occasional Paper 38, Carbondale, IL, pp. 135–171.
Wheeler, W.C., 1993. The triangle inequality and character analysis. Mol. Biol. Evol. 10, 707–712.
Wheeler, W.C., 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst. Biol. 44, 321–331.
Wheeler, W.C., 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? Cladistics 12, 1–9.
Wheeler, W.C., 1999. Fixed character states and the optimization of molecular sequence data. Cladistics 15, 379–385.
Wheeler, W.C., 2003. Iterative pass optimization. Cladistics 19, 254–260.
Wheeler, W.C., Hayashi, C.Y., 1998. The phylogeny of the extant chelicerate orders. Cladistics 14, 173–192.
Wheeler, W.C., Ramírez, M.J., Aagesen, L., Schulmeister, S., 2006. Partition-free congruence analysis: implications for sensitivity analysis. Cladistics 22, 256–263.
Wheeler, W.C., Lucaroni, N., Hong, L., Crowley, L., Varón, A., 2013. Poy 5.0. American Museum of Natural History. http://research.amnh.org/scicomp/projects/poy.php.
Whiteley, P.M., 2011. Hopi place value: translating a landscape. In: Swann, B. (Ed.), Born in the Blood: On Native American Translation. University of Nebraska Press, Lincoln, NE, pp. 84–108.
Whorf, B., 1935. The comparative linguistics of Uto-Aztecan. Am. Anthropol. 37, 600–608.
Wichmann, S., Müller, A., Velupillai, V., 2010. Homelands of the world's language families: a quantitative approach. Diachronica 27, 247–276.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Supplementary Materials. Supplementary Data containing source data files and proto-language reconstructions.