5 What can we learn about the earliest human language by comparing languages known today?

Lyle Campbell
University of Canterbury, NZ

1 Introduction

Looking back from modern languages, what can we find out about the earliest human language? The goal of this paper is to determine what, if anything, can be learned about the earliest human language(s) from evidence extant in modern and older attested languages. It evaluates attempts arrive at the origins of language through such comparisons. The main finding is negative: because of so much change over such a long time, nothing of the original language(s) survives in modern languages in any form that could be usefully compared across-linguistically to give any indication of the lexical or structural content of the original language(s).

2 Methodological issues

A number of linguists have attempted to find deep genetic relationships, so-called 'macrofamilies', and some go even further, attempting to trace all human languages back to a single origin. 'Global etymologies' have been presented as evidence for 'Proto-World' (see Bengtson and Ruhlen 1994a, 1994b, Ruhlen 1987, 1994a, 1994b). Ruhlen's homepage, 'Proto-World' [1] is receiving considerable attention, and therefore it is important to scrutinize it carefully. I argue (see also Bender 1993, Hock 1993, Picard 1998, Rosenfelder 1999, Salmons 1992a, 1992b, Trask 1996:391–6, McWhorter 2001:287–303)

that such a scrutiny reveals that claims about global etymologies are mistaken and cannot teach us anything about the origins of human language.

Both friends and foes acknowledge that the principal method employed in global etymologies is 'mass [or multilateral] comparison' (cf. Ruhlen 1987:258). Ruhlen (1992:178) says, 'John Bengtson and I, operating in a Greenbergian tradition of multilateral comparison, have proposed some thirty etymologies connecting all the world's language families' (see Bengtson and Ruhlen 1994a, Ruhlen 1987:261, 1994a, 1994b). Aitchison (1996:172) calls this the ''lucky dip' approach: trawling through dictionaries, and coming across superficial resemblances between words in far-flung languages'. The criticisms of mass (multilateral) comparison are well-known and need not be repeated here (see Aitchison 1996:172–3, Campbell 1988, 1998a, 1998b, 1999, Campbell and Poser forthcoming, Matisoff 1990, McMahon and McMahon 1995, Rankin 1992, Ringe 1992, 1995, 1996, 1998, 1999, Trask 1996:376–403, etc.). Most linguists reject the global etymologies because they do not find the method used reliable (cf. McWhorter 2001:288). Aitchison (1996:173) summarizes the problems:

> Chance resemblances are easy to find among different languages if only vague likenesses among shortish words are selected sounds change radically over the centuries. Words which existed so long ago are unlikely to have survived in anything like their original state the 'lucky dip' approach does not make any attempt to eliminate accidental correspondences, nor does it control for phonetic probability or taboo meanings tend to be reduced to fairly simple, straightforward items, with a limited number of phonetic shapes. In these circumstances, chance similarities are likely to play a worryingly high role, and this 'mass comparison' method is unlikely to stand the test of time.

Is this correct date?

2.1 Global etymologies: the 'strong' cases

There is not space to evaluate each proposed global etymology; however, a few examples are sufficient to reveal the problems. The two strongest (most cited) have already been evaluated rigorously, **tik* 'finger' (Bender 1993, Salmons 1992a, 1992b) and **maliq'a* 'to suck(le), nurse, breast' (Hock 1993, Hock and Joseph 1996:498–502). As critics show, the data (Bengtson and Ruhlen 1994a:322–3) are much weaker than they at first appear, and the methods employed are unable to show that chance is not a more plausible explanation than genetic relationship.

In standard etymological criteria used among languages known to be related, purely accidental lexical matchings are constrained by the demands

of sound correspondences and semantic equivalence (see Goddard 1975:254–5, Salmons 1992a). The numerous non-cognate lexical similarities in closely related languages show why such criteria are necessary, e.g.: English *day*: Spanish *día* 'day' (these do not obey Grimm's law as true cognates do, English < Old English *dæg* < Germanic **dagaz* 'day' < Proto-Indo-European (PIE) **agh* 'day'; Spanish < Latin *dies* 'day' < PIE **dyē-*, **deiw-* 'to shine'); Spanish *mucho* 'much': English *much* (Spanish < Latin *multus* 'much, many' < PIE **mḷ-to-* 'great, strong', English < Old English *micel*, *mycel* 'great', 'much' < Gemanic **mik-ila* < PIE **meg-* 'great'); Hungarian *fiú* 'boy': Romanian *fiu* 'boy, son' (Hungarian < Proto-Finno-Ugric **poyi* 'boy, son'; Romanian < Latin *filius* 'son' < PIE **dhi:-lyo-* < **dhe(i)-* 'to suck, suckle').

Global etymology does not heed these constraints known to be necessary even in closely related languages, as seen from the example **kuna* 'woman'. This is one of the strongest cases (cf. Allman 1990, Bengtson 1991). Bengtson and Ruhlen (1994a:306) list for this words of the following shapes from various languages: *knw*, *eq*^{w}*en*, *xuonā*, *teknē*, *wanā*, *gerim*, *grua*, *ben*, *kin*, *žena*, *günü*, *arnaq*, *chana-da*, *k'uwi*, *hun*, *ʔunu*, *huini*, *kuyã*, *ekwaʔa*, *hanökö*, etc. While global etymologists do not spell out what criteria they follow to determine whether something fits, the target is CVC(V), where differences in the vowels are ignored. For **kuna* 'woman', the target is approximately *KVN(V)*, where '*K*' is any velar-like sound, '*N*' some *n*-like sound. However, matches are not tight, since for the '*K*', any of the following fits: *k, k', g, q, x, h, w, b, ž, ʔ, č*. For the final '*N*', any of the following count: *n, r, m, ã, w*ʔ*, ʔ*?, and Ø. Even '*KV*' seems to be accepted. As for the glosses accepted which allow a form of this vague phonological shape to be selected as a match, all of the following are encountered among the forms for the 'woman' global etymology: 'wife', 'woman', 'lady', 'mother', 'female' (any species), 'spirit of dead woman', 'girl', 'daughter', 'maiden', 'daughter-in-law', 'small girl', 'young woman', 'old woman'.

Salmons' (1992a:5) understanding of global etymologists' principles for whether something is a 'cognate' agree with mine:

> A. Ignore vowels entirely: Any vowel matches any other vowel …
>
> B. For consonants, roughly similar place of articulation suffices to establish cognates [though non-initial consonants are sometimes allowed drastic differences]. Minor place changes are acceptable: Velars match uvulars, palatals, etc. Other features play no role whatsoever, so that oral stops correspond to nasals, etc.

> C. Any differences in place which parallel widely attested sound changes such as lenition are acceptable, so that any consonant can be reflected by [h] …
>
> D. In semantics, any narrowing or any metaphorical extension is acceptable without further justification (such as cultural or historical arguments), so that 'dog' corresponds to 'fox, lynx, deer', etc. and 'arm' to 'elbow/hand, fingernail, foot, armpit, shoulder/arm' and so forth.[2]

So, how difficult can it be to find forms that fit the range of permitted sounds and meanings by accident for **kuna* 'woman'? Answer, easy. The following from Spanish illustrates how easy it is:

> *cónyuge* 'wife'
> *cuñada* 'sister-in-law'
> *china* 'girl, young woman' (and *chinita* 'Indian girl')
> *cana* 'old woman' (adjective)
> *canuda* 'old woman'

These are just accidentally similar to the forms in the **kuna* 'woman' global etymology, since we know their history and it shows the forms have etymologies where the sounds and meanings in question do not originally match the target of the global etymology. *Cónyuge* is from Latin *con-* 'with' + *jugum* 'yoke', where these pieces have nothing to do with 'woman'. Similarly, *cuñada* is from Latin *cognātus*/*cognāta* 'consanguineal relative' (*con-* 'with' + *nātus* 'born'), and so in origin has nothing to do with the sound-meaning equation of the global etymology for 'woman'. *China* is a loanword from Quechua *čina* 'female of animals', and thus cannot be a direct inheritance in Spanish from Proto-World. *Cana* is from Latin *canus*/*cana* 'white' (with the sense 'old' through 'grey hair'), with no connection originally with 'woman'.

2.2 Criticisms

Hock (1993) demonstrated the point that seeming fits are easy to find by accident for the various proposed global etymologies using such procedures; he showed in a comparison of Hindi and English (IE languages) that 65% of the items that would be identified as 'cognates' by the methods of global etymology are 'false friends,' i.e. non-genetic similarities. The excessive generosity in deciding what fits phonologically and semantically has frequently been criticized (cf. Aitchison 1996:173, Bender 1993, Trask 1996:395).

The exercise – as in the Spanish examples above – of finding various words with disparate known histories which nevertheless fit proposed global etymologies reveals the severest criticism, namely that global etymologies

cannot be tested. 'The methods of 'global etymology' remove all controls on accidental similarity' (Salmons :1). Without the constraints of standard methods, claims that something fits a putative global etymology are not falsifiable, since their fit cannot be checked against proposed sound correspondences or constraints on semantic shift (Salmons 1992a). Moreover, even if we eliminate all the forms whose history does not fit, using the same techniques, it is always possible to generate new examples to replace them. For example, even if we demonstrate that because of their history forms cannot be connected, as with the Spanish forms that fit **kuna* 'woman', this method can nevertheless just produce more examples of the same sort, whose histories may not be so well known (Spanish *cañenga* 'old woman', *changa* 'girl', unclear etymologies; see Salmons 1992a for many in the **tik* 'finger' set; Hock's 1993 criticism of the **maliq'a* 'suck(le), nurse, breast' set). Thus the supporters' allegation that 'you have to take our claims seriously, we have so many examples' (Bender 1993:192) is hardly compelling, given the nearly inexhaustible source of new examples which are accidentally similar but where there are no constraints on how to restrict such accidents.

In spite of the evidence to the contrary, Bengtson and Ruhlen (1994a:281) believe that 'the failure of our critics to appreciate the truly minuscule probability of accidental similarities is the chief impediment to their understanding why all the world's languages must derive from a common origin'. The longish lists of 'so many examples' at first strike the uninitiated as impressive. However, the fact is, given the looseness of the semantics and phonetics permitted for matches, large numbers of forms accidentally similar can easily be found. Thus, falsifiability is not possible (cf. Salmons 1992a:217). 'It is impossible to distinguish between significant and chance resemblances' (McWhorter 2001:297). 'How do we constrain our imagination and ingenuity if we lack explicit controls?' (Bender 1993:195).[3]

2.2.1 A test

Bengtson and Ruhlen (1994a:290) suggest tests which could falsify their claims, but which they believe will bear out their belief that their findings cannot be due to a mere assembly of accidentally similar forms. For the first test, in response to those who say 'one can find anything in linguistic data if one looks for it hard enough,' they say:

> 'Wanting' to find something is of very little help if it is not there that the Amerind family has two general words for females, TUNA 'girl' and KUNA 'woman' whereas KUNA is widely attested in the Old World we have found no trace of TUNA in the Old World. If it is so easy to

> find anything one looks for, why did we fail to find TUNA in roughly 4,500 Old World languages. That there is no trace of TUNA 'girl' in the Old World is because it never existed there. (Bengtson and Ruhlen 1994a:290)

So, if we do find words from Old World languages which fit the range of glosses and phonetic forms of the TUNA material presented for Amerind, Bengtson and Ruhlen would concede that it is possible to find accidentally similar forms. It is not difficult to meet this challenge. There is one of the weakest in Greenberg (1987), with examples presented from only four of Greenberg's eleven branches (Greenberg 1987:225, #125). The forms presented for the assumed Amerind etymology includes: *tun*, *tana*, *-tsan*, *šan*, *tsini*, *tu:ne*, *tele*, *suri-s*, *teŋ*, *tunna*, *t'an'a*, etc. The glosses covers: 'son, daughter, diminutive, small, child, be small, mother, daughter'. It is not difficult to find similar words in non-Amerind languages. A quick look at a few dictionaries in my office turns up: Finnish *tenava* 'kid, child', German *Tante* 'aunt'; Japanese *tyoonan* 'eldest son'; Malay *dayang* 'damsel'; Maori *teina* 'younger sister', 'younger brother', Somali *dállàan* 'child'; even English *son* fits.[4]

2.2.2 Another test

Bengtson and Ruhlen claim in several works that it would be impossible to take their list of proposed global etymologies and produce equally impressive lists of words if the meaning is shifted one number in each case, that is, where instead of their (1) AJA 'mother, older female relative', (2) BU(N)KA 'knee, to bend', etc., rather we should assemble sets similar to theirs but with (1) AJA 'knee', (2) BU(N)KA 'ashes, etc. In fact, Bender (1993) took up this challenge and demonstrated that such sets of similarities can be assembled easily, showing that accidental similarity is at stake in much of what they present.

2.2.3 'Reaching down'

Another criticism is 'reaching down' (Trask 1999), accepting forms as evidence of Proto-World which are found only in a single language or in a single branch of a family. This violates Meillet's (1925:38) heuristic that evidence is needed from more than one branch, and the more languages and branches represented the better as evidence of cognacy (Salmons 1992a, Trask 1996:394).

2.2.4 Unlikely semantics

Another criticism is that the semantics in well-studied families often reveal problems with forms selected as support for global etymologies (Salmon 1992a). For example, PIE **deik*- is listed as evidence of the **tik* 'finger, one' global etymology; however within IE languages, the meanings 'finger' and 'point' upon which Bengtson and Ruhlen focus are secondary, attested only in Latin and Sanskrit. The meaning supported by the other branches is 'to pronounce solemnly, to show', with 'derivatives referring to the directing of words and objects' (Watkins 1985:10; cf. Trask 1996:394). Lack of constraints on accident and semantic latitude 'leads to such absurdities such as accepting that Amerind Tikuna 'elbow' is genetically related to Latin 'to say'' (Bender 1993:196).

śr̥ṅga

2.2.5 Errors in data

Bengtson and Ruhlen's (1994a) global etymologies have been criticized for the many errors in the data. For example, Picard (1998:146) found in the 9 Algonquian forms listed in their 27 global etymologies, 3 were attributed to the wrong language, 4 were given with the wrong gloss, 4 had errors of morphological segmentation, 3 were transcribed wrongly, and all 9 have serious problems of this sort. In general, mistakes of these sorts are found throughout the words presented for the 27 proposed global etymologies.

2.2.6 Short forms

Another criticism is that short forms are not sufficiently long to eliminate chance as a possible explanation for similarities perceived. The length of proposed cognates and the number of matched segments within them are important, since the greater the number of matched segments in a proposed cognate set, the less likely it is that accident may account for the similarity (cf. Meillet 1958:89–90); as Greenberg (1996:134) put it, 'the longer an item, the greater its weight' (cf. Ringe 1992, 1995, 1996, 1999, Nichols 1996). Unfortunately short form examples are common in proposed global etymologies. Ryan's (2001) 90 monosyllabic words of 'Proto-language' are all CV or similar. Several of Bengtson and Ruhlen's (1994a) 27 proposed global etymologies are short. Only one is longer than two syllables, **maliq'a* 'suck(le), nurse, breast', which has been thoroughly discredited (Hock 1993, Hock and Joseph 1996:498–502). Most are intended to be bisyllabic (19), though occasional CV words are cited in (e.g. Korean *ka* 'dog' for global **kuan* 'dog'); 4 are monosyllabic CVC

shape (with occasional CV examples, e.g. Proto-Yao *(w)i 'two' and Mak *wa* 'twin' in support of global **pal* 'two'); and 2 are CV(C).

For **ku(n)* 'who?' (Bengtson and Ruhlen 1994a:303–5) we find: *xa, ka, kí/ká, k(w)/q(w), gin, ka:na, k^wo/k^wi, ke/ki, ku/ko, hu, kua, kutte, kun, qun, kon, ken, gi, gæ, xaj, aj, udu, i:, adi, ono, o:n(i), k'e, mik/mek, ajkia, qa-, kjei, gyis-oto, gùsú, gigi, gunuga, kamu, o-ko-e, ku'a(ʻ), gu-, jus, kek^w, ka-n, a:č'is, kwanu, go:š, xaŋ, key, ki:, kia, k'owa, kin, kai, karea, karo, kejaito, go:si, kate, kia, koide, katsik, kona, gaga, kepia*, etc. Clearly it is possible by chance to find similar forms in many languages; the -n is not necessary for a match, and any vowel counts; for initial **k*, it appears a wide range of consonants qualifies. As for the meanings, anything vaguely interrogative seems accepted – '*who, what, when, which, where, why, how, how much, how many, interrogative particles, whither, whence, someone, either or, anything*', etc. In short, if anything from *i:, udu*, and *aj* to *qanangun*, *kiš-to*, and *ekkwarijawa* meaning anything from 'who' to 'anything' can be seen as evidence in favour of this set, then it is indeed difficult to imagine how chance as a possible explanation for forms such as these could be denied.

The treatment of **mi(n)* 'what?' (Bengtson and Ruhlen 1994a:313–5) is similar. Among the forms presented are: *kama, ma, m(j), mann, mi, mah-ma:, mi:t, miya, mena, -ma, maj, mo-, ma/mo, -u:, mida:, wi-/we-, amin, minh/minya, amae, mu, a:mai, m'as, matswɛ, mi:š, maua, manti, mato, may, mano, muski, makaya, maap, mukat, muda, manpat, miki, muru, mba'e, mukoka, mi, muena, ampô-ny, matuni, mašika*. The glosses include: '*if, when, where, who, which, what, how much, when, what kind of, sentence interrogative, thing, this, something*', etc. To find a match by accident, one need only find some form in any language which means something interrogative or 'if, something, thing, whether', with *m*, although the *m* is not strictly required, since some forms listed lack it. Again, chance is surely a major factor behind the grouping of many of these examples.

In the often-cited **tik* 'finger, one' case, the forms demonstrate that a match need have little in common with the final *-k*: *tsiho, dě, dɛ̂ʔ, ti, tu, (s-)tʲa, tay, (tu-)diŋ, (pi-)tˢi*, etc. The assumption that sound changes produced these forms gives the investigator excessive power to imagine matches where chance is probable.

Please confirm this is correct date

The failure of the methods to distinguish chance from real history as the explanations of the sets of compared words offered as global etymologies is a devastating criticism. Much work has shown such methods incapable, even remotely, of exceeding chance as the probable explanation for the forms cited: Nichols 1996, Ringe 1992, 1999, Salmons 1992a (cf. McWhorter 2001:292–303).

3 Some things that are not reliable evidence

3.1 Nursery words

It has been recognized for centuries that nursery formations (the *mama-nana-papa-dada-caca* sort) should be avoided as evidence of genetic relationship, since they typically share a high degree of cross-linguistic similarity which is not due to common ancestry (cf. Greenberg 1957:36). Nevertheless, such words are frequent in the evidence put forward for hypotheses of distant genetic relationship, including Proto-World (cf. Bengtson and Ruhlen 1994a:292–3, Ruhlen 1994b:122–4, 2000).

Murdock (1959) investigated 531 terms for 'mother' and 541 for 'father' in different languages and concluded that the data 'confirm the hypothesis [of] a striking convergence in the structure of these parental kin terms throughout historically unrelated languages' (Jakobson 1962[1960]:538). Jakobson explained the non-genetic similarity cross-linguistically among such terms as nursery forms which enter common adult vocabulary:

> Often the sucking activities of a child are accompanied by a slight nasal murmur, the only phonation which can be produced when the lips are pressed to mother's breast or to feeding bottle and the mouth is full. Later, this phonatory reaction to nursing is reproduced as an anticipatory signal at the mere sight of food and finally as a manifestation of a desire to eat, or more generally, as an expression of discontent and impatient longing for missing food or absent nurser, and any ungranted wish. Since the mother is *la grande dispensatrice*, most of the infant's longings are addressed to her, and children gradually turn the nasal interjection into a parental term, and adapt its expressive make-up to their regular phonemic pattern. (Jakobson 1962[1960]:542–3.)

The forms with nasals are found more frequently in terms for females, stops for males, but not exclusively so. Because these kinship terms are often found to be phonetically similar across genetically unrelated languages, and because this non-genetic similarity has plausible explanations, such nursery words are not considered viable evidence in proposals of distant genetic relationship (see Campbell 1998a:227–9). The cases put forward as evidence of Proto-World are not reliable evidence; the following from Bengtson and Ruhlen's (1994a, Ruhlen 1994b, 2000) global etymologies are challenged: **aya* 'mother', older female relative', **mama* 'mother', **papa* 'father', and **kaka* 'older brother', 4 of 27 sets.

3.2 Onomatopoeia

Onomatopoetic forms may be similar because the words in different languages have independently approximated sounds of nature; such cases must be eliminated from proposals of genetic relationship. As Swadesh (1954:313) advised, 'a simple way to reduce the sound-imitative factor to a negligible minimum is to omit from consideration all such words as 'blow, breathe, suck, laugh' and the like, that is all words which are known to lean toward sound imitation'. Judgements of what is onomatopoetic may be subjective; however, forms whose meaning plausibly lends itself to mimicking the sounds of nature are often found in proposals of distant genetic relationship, e.g. comparisons among languages of words meaning 'blow/wind' which approximate *p(h)u(h/x/w/f)* phonetically, and of 'breast/suckle, nurse/suck' *(V)m/nVm/n*, *s/s/tś'č/Vp/b/k*, or *s/s/tś'č/Vs/s/tś'č*, as seen in numerous forms presented as putative cognate sets in proposed but controversial 'macrofamily' hypotheses (cf. Rosenfelder 1999). Some words which frequently are similar across languages due to onomatopoeia are: 'break/cut/chop/split', 'baby', 'breathe', 'choke', 'cough', 'cry', 'cricket', 'crow' (and many bird names in general), 'frog/toad', 'lungs', 'beat/hit/pound', 'call/shout', 'drip', 'hiccough', 'kiss', 'nose/smell', 'shoot', 'sneeze', 'snore', 'spit', 'whistle'.

Proposed global etymologies must contend with the question of possible onomatopoeia (and of affective, expressive, or sound symbolic forms) among the words from various languages listed. Here, I list some of examples from the proposed global etymologies together with a brief indication of why some scholars see onomatopoeia or affective forms in these cases. Some of these will be more persuasive than others, though all warrant serious consideration. To the extent that onomatopoeia and affective formation are involved, the similarities seen in cross-linguistic comparisons owe their origin to later developments, not to inheritance from 'Proto-World'. These include the following (from Bengtson and Ruhlen 1994a:277–336, Ruhlen 1994b:101–24).

'Breast/suck(le)/nurse' **maliq'a* (see Hock 1993), illustrated by: *maal-, melu-t, mellu, mekku, umlix, mik'-is, murgi, mallaqa*, etc. Similarities among these words across various languages are generally thought to be due to imitation of the noises children make when nursing, sucking. In this case it is complicated by the fact that many of the words given (see Bengtson and Ruhlen 1994a:308–9) mean 'swallow', 'food', 'chew', 'eat', 'throat', 'neck', and 'chest', and thus have no particular motivation to mimic sucking/nursing noises, but, then, this only means that onomatopoeia and accidental similarities both are involved.

'Dog' **kuan*, with forms: *!gwaī, gwí, kwon, ka, x^{w}oʔi, kawun, kwi*, etc. Some linguists believe similarities such as these are imitative of dogs 'hauling'

and 'barking' and 'growling', perhaps with a nursery component, since dogs since children have affective associations with household pets. As Hock and Joseph (1996:498) point out, 'in a number of Indo-European languages, the original word for 'dog' was replaced by words with initial *ku-* such as Sanskrit *kurkura-* 'the one that snarls, growls, or barks, i.e. makes a sound [kurkur]'''; they cite as further examples English *cur*, German *Köter*, Modern Hindi *kutta*:, Tamil *kurai* 'to bark' / *ku:ɻan* 'dog'; many other 'dog' examples could be added, e.g. Finnish *koira*, Māori *kuri*:, etc.

'Fly' (verb) **par*, illustrated with: *pil, far, ferfir, par, -biri, phur, aphir, bin, ʔbil, pen, pau, pal, parpal, purupuru, piropir*, etc. Many see in such words (which include in Bengtson and Ruhlen's lists also 'wing', 'butterfly', 'flee', 'moth', 'bird') both onomatopoetic and affective, sound symbolic aspects. Such words for 'fly' and 'wing' suggest the imitation of the sounds of 'flapping', 'fluttering', 'flying' made by birds' wings, thus explaining (1) the similarity found among unrelated languages, and (2) why the same language can have multiple non-inherited words of this sort (compare for example English's: *fly, flap, flutter, flit, flicker, whoosh*, etc.).

That affective sound play is involved in some cases cited as evidence is especially evident cross-linguistically in words for 'butterfly' (cf. folk-etymological *flutterby*) (some examples are in Bengtson and Ruhlen's 1994a:317–8 global etymology for 'to fly'):

> 'butterfly': Albanian *flutur*, Arabic (Moroccan) *fertattu*, Bunabun *piropir*, Dravidian (Kolami *gu:ge*, Gondi *gu:ge*, Parji *gogava:la*; Tamil *pa:ppa:tti*, Malayalam *pa:ppa:tti*, Kodagu *pa:pïli*, Gonid *pa:pe*:, *pipri*:, Kurux *papla*:), Estonian *liblikas*, Finnish *perhonen*, French *papillon* (< Latin *pāpillō*), Guarao *guaroguaro*, Hindi *tiitri/titli*, Indonesian *kupu-kupu*, *rama-rama*, Italian *farfalla*, Māori *pu:rerehua*, *pe:pe*, Mískito *pulpul, dildil*, Nahuatl *papalo:-tl*, Paya *waruwaru*, Portuguese *borboleta*, Proto-Austronesian **qaLi-baŋbaŋ*, Proto-Lezghian **pa(r)pal-*, Proto-Mayan **pehpen*, Proto-Zoquean **me:meʔ*, Quechua *pilʸpintu*, Sumu *saisai*, Swedish *fjäril*, Tequistlatec *-bobolóh*, Totonac *špiʔpiʔle:ʔqa*, Ulwa (Sumu) *kublamhlamh*, Welsh *pilipala*.

'Smell/nose' **čuna/*čunga,* with: *sun, sina, snā, čona, sányuu, sinqa, tsinyu*, etc. These comparisons suggest imitation of the sounds of 'sniffing', 'snuffling', and 'smelling', which in many languages have affective and nursery-word connections from the runny noses associated with children and their numerous childhood illnesses. (cf. English phonaesthetic forms with no regular etymologies: *sneer, sneeze, sniff, sniffle, snivel, snot, snotty, snort, snuff, snuffle*).

'Water' **aq'wa* (with forms such as *akwa, okho, gugu, k'a*, etc.). The similarity of sound suggests to many the imitation of the sound of swallowing water, a nursery form, or of gurgling running water.

4 The futility of modern lexical comparisons as evidence of Proto-World

Can lexical comparisons across known languages offer any insight into Proto-World or the origin of human language(s)? Lexical comparisons have seldom been considered convincing proof of genetic relationship without additional support, e.g. from sound correspondences and shared irregularities in morphological elements. It is easy to see why this should be even more the case with global etymologies. By glottochronology, after about 14,000 years, nearly all of a language's basic vocabulary will be replaced, so in related languages which split up before 15,000 years ago, we will not find recognizable cognates.[5] Glottochronology may not be supported, but this illustrates the point that over vocabulary is replaced and the lexical comparisons of global etymologies must expect cognate vocabulary to survive in modern languages for tens of thousands of years unreplaced and in recognizable form – extremely unlikely given the amount of normal lexical replacement and phonological change that take place in far shorter lengths of time (see below).[6]

Given the extremely long time since the origin of human language, absolutely all lexical items from that period will have been replaced or changed beyond recognition in all languages. Others make the same point about so much change over such a long time leaving no residue in modern languages or leaving whatever survivals could be imagined too garbled through the regular workings of linguistic changes to be recognizable (cf. Trask 1996:392, McWhorter 2001:292, Hock 1993:218).

The extent of this problem can be appreciated from Hindi and English, languages known to be related.[7] I mark the forms compared in the Swadesh 100-word list with the following codes before the numbers:[8]

+: true cognate which would be recognized by the methods utilized by global etymologists

+?: true cognate which might be accepted by global etymologists, though are by no means obvious

-: non-cognate form which would nevertheless be accepted by the methods

-?: non-cognate form which perhaps would be accepted by the methods, though it should not be

#: true cognates which would be missed by the methods of global etymology

#?: true cognates which very likely would be missed by the methods, though perhaps not.

Equivalents from Māori (an Austronesian language) are also compared. I code English-Māori similarities with <-> for cases which would be accepted by the method, though they are not cognates, and <-?> for weaker cases that perhaps would be accepted. For Hindi-Māori similarities, in order to distinguish them form those with English, the symbol <%> is used for those the method would accept, and <%?> for those it might accept.[9]

	English	**Hindi**	**Māori**
1	I	maĩ) (but see *me*)	ahau
2	you	a:p (polite), tum, tu: (informal)	koe (singular)
#3	we	ham (cf. Sanskrit *vayam* 'we')	ma:tou (exclusive several), ta:tou (inclusive several)
4	this	yah	-? te:nei
5	that	vah, voh	- ? te:ra: (that away), te:na: (that near)
+? 6	what	kya:	%? he aha?
+? 7	who	kaun	- wai
+8	not	nahĩ)	ka:hore, ka:o
9	all	sab	katoa
10	many	bahut	- maha
#11	one	e:k	tahi
+12	two	do:	-? rua
-? 13	big	baṛa:	- pi:ki, nui
-? 14	long	lamba:	-? roa
15	small	choṭa:	iti
16	woman	stri:, aurat	- wa:hine
17	man	a:dmi:, puruṣ	ta:ne
18	person	vyakti:, log, insan	tangata
19	fish	machli:	ika
20	bird	pakśi:	manu
21	dog	kutta:	% kuri: (cf. English *cur*)
22	louse	jũ:	%? kutu
23	tree	pe:ṛ	ra:kau (cf. to:tara 'tree' (podocarpus)

	English	Hindi	Māori
24	seed	bi:j	%? pua
25	leaf	patta:	wha:rangi
26	root	mu:l	pakiaka (cf. rauruhe 'fern root')
27	bark	chha:l(f.)/kha:l	-? pa:pa:kiri, kiripaka, kiri, hiako
29	flesh *	mã:s	- mi:ti (English loan)
30	blood	xu:n, lahu, rekt	toto
31	bone	haḍḍi:	-? poroiwi, wheua, iwi
#32	egg	aṇḍa:10	– he:ki, hua manu
33	grease	charbi/chikna:'i	hinu
#34	horn	sĩ:g	- haona, maire, pi:hi
35	tail	dum/pũ:chh	- te:ra (English loan), waero, whiore
+?36	feather	par	piki
37	hair	ba:l	- huruhuru
38	head	sir	ma:tenga
39	ear	kan	taringa
#?40	eye	ã:kh	%? kanohi
+?41	nose	na:k	ihu
-42	mouth	mũh	-? ma:ngai
+?43	tooth	dã:t	niho
#44	tongue	ji:bh, zaba:n[11]	arero
45	claw	chã:gul/na:xun/pã:jah	maikuku
#?46	foot	pã:v, pair	-?/% pu:, waewae, take
-?47	knee	ghuṭna	turi, pona
-48	hand	ha:th	ringa
-?49	belly	peːṭ	-? puku, %? ho:para
50	neck	gardan	kaki:
51	breast	chha:ti:	uma,poho
52	heart	dil[12]	nga:kau
53	liver	jigar/kaleyja	ate
54	drink	pi:-	inu, unu
55	eat	kha:-	% kai
56	bite	ka:ṭ-	% kakati, ngau
57	see	de:kh-	kite
58	hear	sun-	rongo
#59	know	ja:n-	mo:hio
60	sleep	so:-	moe
61	die	mar-	% mate
62	kill	ma:r-/ma:r ḍa:l-na:	% whakamate, -mate
63	swim	tair-	kaukau
64	fly	uṛ-	rere, tere

	English	**Hindi**	**Māori**
65	walk	chal- 'walk', ja:- 'walk, go'	- wa:ke (English loan), %? haere
66	come	a:-	heke, kuhu
-67	lie	leṭ-	takoto
68	sit	baiṭh-	noho
69	stand	khaṛa + ho- 'standing'	tu:, tu:tu:
70	give	de:-	%? tapae
71	say	kah-	%? ko:rero
+72	sun	su:raj, su:rya	%? ra:
73	moon	chã:d	- marama
+/-74	star	ta:r, sita:ra:[13]	whetu:
75	water	pa:ni:	- wai
76	rain	ba:riś	ua
77	stone	patthar	% po:hatu, ko:hatu
78	sand	ba:lu	onepu:
79	earth	zami:n, prithvi:, mitti	oneone, paru
80	cloud	ba:dal	kapua, ao
81	smoke	dhuã:[14]	paoa
82	fire	a:g	% ahi, -? ka:pura
83	ash	ra:kh	%? pungarehu
84	burn	jal-	%? ka:, ngiha, tahu, wera
-85	path	pagḍaṇḍi:, pa:th	huanui
86	mountain	paha:ṛ	- maunga
87	red	la:l	%? whero
-?88	green	hara:	- kiri:ni (English loan), % karera, ka:riki
89	yellow	pi:la:	%? Punga, ko:whai
90	white	safe:d	ma:, tea
91	black	ka:la:	pango, mangu
92	night	ra:t	po:
93	hot	garm (gerem)	%? wera (cf. *warm*)
94	cold	ṭhaṇḍa:	makariri
+?95	full	pu:ra:	-/% puhapuha
96	good	accha:	pai, tika
+?97	new	naya:	ho:u
98	round	go:l	porotaka
99	dry	su:kha:	maroke
+100	name	na:m	ingoa

* Note: if 'meat' could be substituted, one gets a <-> for the English-Hindi comparison)

The ancestor of English and Hindi did not begin to diversify into separate languages until some 5,000 or 6,000 years ago, but we find only some five clear cognates on the Swadesh list (those marked <+>), only some 13 by generous criteria (marked <+?>), several of which would only be chosen by someone utilizing liberal notions of phonetic similarity. If the impact on the vocabulary of clearly related languages is so great after only a few millennia, surely there is no hope for comparisons at the level of Proto-World, comparisons in which the languages involved are assumed to have separated from one another some 100,000 years ago or more (see below). It is clear that the English-Hindi comparison – with only 9 cases clearly selected by the method (those marked <+> or <->) – fares worse than the English-Māori comparisons (15 cases accepted, marked <->; not counting the 'maybes', marked <-?>) and than the Hindi-Māori comparisons (10 cases, marked <%>). The differences between English-Hindi and Māori with the other two languages are so striking that a shift in coding for a few items would not greatly alter the outcome that looks as similar to English and Hindi, to which it is not related, as related English and Hindi do to one another. Clearly if unrelated Māori exhibits more matchings with both English and Hindi of the sort the method accepts than these two IE languages do with each other, then there is something alarmingly wrong with this method. This comparison demonstrates that it cannot perform better on related languages than on unrelated ones and therefore sheer accident must be the explanation for many of the matchings accepted as global etymologies.

Taking into account what is known of IE would reveal more English-Hindi cognates, but also would expose additional similarities known not to be cognates. As Hock (1993) pointed out, often the cognates are changed so much by sound changes that they would not be recognized by search for superficial similarity followed by global etymologists. For example, the following English-Hindi cognates are not phonetically similar enough to be selected by such methods (from Hock 1993:218): horn: *sĩ:g* (< Sanskrit *śr̥ṅga* - 'horn'), *sister: bahan* (< Sanskrit *svasar*, cf. Old English *sweostor*), *be*: *ho*:- (< Sanskrit *bhavati* < PIE **bhu*:), *we*: *ham* (< Sanskrit *vayam*), etc. Hock (1993) and Hock and Joseph (1996:469, 491–3) list several others.15 Hock and Joseph (1996:492–3) report that in an open-ended search, some 55 genuine cognates turn up which are still similar enough phonetically and semantically to appear related, plus some 30 other cognates so altered by linguistic change that they would probably not be recognized without historical knowledge. However, this is complicated by:

1 more than 45 loanwords in Hindi from Sanskrit which have English cognates, but are not direct inheritances in Hindi;

2 5 loans from Persian into Hindi;

3 10 or more loans from other sources;

4 60 cases of phonetically and semantically similar forms known from their history to be purely accidentally similar. As Hock and Joseph (1996:493) show, no matter how the genuine cognates are balanced against accidental similarities and loans, there is less than a 50:50 chance that similarities that would be selected by the method used to identify global etymologies would select genuine cognates.

The argument of too much garbling having taken place since Proto-World for anything to survive or be recognizable depends in part on the date assigned to Proto-World. Clearly if human language is 100,000 (coeval with anatomically modern humans) or older, as some claim, then the amount of garbling and replacement are surely far too much to imagine the survival of anything like a recognizable cognate. There is, however, an interesting twist on views of the date. Researchers of the Santa Fe Institute reason in reverse. Since they believe that real evidence of Proto-World survives in today's languages, they argue that the date of human language must be much later than commonly thought in order to accommodate these assumed linguistic survivals: here are serious indications that all existing human languages are descended from a single ancestor, 'Proto-World', which would have been spoken some tens of thousands of years ago. (It seems that an age of one or two hundred thousand years can be ruled out: there would not be any significant amount of evidence remaining. http://www.santafe.edu/sfi/organization/annualReport/00/activities/evolution.html.)

Bengtson and Ruhlen (:4, 57) also suggest that the date involved is not so early, that 'the origin of modern linguistic diversity is to be traced only to the advent of *behaviorally*-modern humans, who appear in the archaeological record between 50,000 and 40,000 years ago.' They ask, might not linguists 'be able to perceive similarities going back 40,000 years?'

The answer is almost certainly 'no' (seen in the Hindi-English-Māori comparisons). However, this dating is also too recent. Australian aboriginal peoples reached Australia by 50,000 years ago. This means that human language must be at least as old as their arrival, since no one imagines they arrived and then developed language subsequently. This probably took place considerably before the rock painting, venus and animal figurines, and burial rites of the European Upper Paleolithic, from ca. 35,000 years ago (mentioned in the Santa Fe Institute's report) sometimes associated with early human language. Even if human language were as young as 40,000 years, this length of time would be more than sufficient to produce the same result, so much lexical replacement and change that nothing reliable could be inferred for Proto-World from lexical comparisons. The extent of the English-Hindi differences after

only a fraction of that time, some 5,000 years, should be sobering for anyone who expects recognizable lexical survivals some 35,000 years or more further into the past.

5 Structural speculations

What would the structure of 'Proto-World' ('proto-language') look like? Can we get an idea looking back from structural traits of modern languages? Would 'Proto-World' be simple or complex? Both views have been favored, though the simple-to-complex view has dominated. A third view imagines that whatever in today's language has functional or typologically motivation would also have characterized early human language. I consider each, beginning with the last.

5.1 Functional-typological accounts

To illustrate this sort of argument, consider the claim that Proto-World had SOV word order (cf. Newmeyer 2000). One reason for suspecting this has to do with the claim that changes from OV > VO are more common and natural than VO > OV. A more extreme form of the claim is that languages can only acquire SOV order through language contact, that SOV does not arise through internal developments (cf. Faarlund 1990:84, Tai 1976). However, this claim is incorrect (Campbell, Bubenik, and Saxon 1988, Harris and Campbell 1995:405). While borrowing is a prominent path for the development of new SOV languages, there are other pathways. Another reason is that 'SOV order predominates among the world's languages today' (Newmeyer 2000:372; see Song 2001:49–137). Nevertheless, the following are relevant: (1) some languages have shifted their word order thoroughly even more than once, meaning it is difficult to project their histories from their current state of affairs. (2) There are strong functional typological motivations for why some languages will prefer SOV over the other logically possible word orders (see Song 2001), meaning that regardless of the word order it starts out with, a language may have changed to SOV for good reasons. (3) The number of logically possible orders available is small (only two, OV or VO, in some interpretations), constrained further by the typological tendencies mentioned. Taken together, these considerations make it clear that word-order in Proto-World need not have been SOV. That is, from what we know of possible word-order changes and typological motivation, and given the time depth, human language could have started with any word order and we could easily get to today's distribution of word orders in the world's languages.

Another example involves Nichols' 'stable features'. Nichols (1992, 1995, 1998) argues that certain typological traits are 'relatively persistent in language families, of relatively low frequency worldwide, not readily diffused, and not likely to arise spontaneously' (Nichols 1998:143–4). These include: head/dependent marking, typological alignment (nominative-accusative, ergative, active), morphological complexity, inclusive/exclusive, alienable/inalienable, noun classes, numeral classifiers, etc. Some have speculated that Proto-World would have been characterized by these 'stable' traits, either because these traits represent retentions in modern languages, or because, given their stability and utility, languages of the remote past as now would tend to have such traits, even if those in today's languages do not reflect direct survivals. This does not represent Nichols' own view; she rather concludes that 'nongenealogical comparison [among these 'stable' traits] can tell us a good deal about when and where modern language arose and about the proximate and ultimate major geographical contributors to large populations of languages' (Nichols 1998:165). Nevertheless, there is an implication in her 'nongenealogical comparison' that many of these will be traits of early human language, in Africa, which after initial spread tended to persist with subsequent change delivering the geographical distributions of the traits; she focuses on this distribution, though the origin is implied.

A serious problem for relating the 'stable traits' to Proto-World is that there is nothing particularly stable about most of them. For example, the inclusive/exclusive first person pronoun contrast is not stable, but can develop or be lost easily. The same language can differ in that some dialects have the contrast and others lack it, where the change is very recent. For example, some Mam (Mayan) dialects have the contrast, 'exclusive' clitic *-a*/*-ya*, 'inclusive' Ø; other Mam dialects lack it. The inclusive/exclusive contrast is typically superficial, not deeply integrated in the fabric of the grammar; there is nothing about it which would lead us to expect long-term 'stability' (see Jacobsen 1980:204, Foley 2000:392 for other examples).

Notwithstanding, Nichols 'turns this one example [inclusive/exclusive opposition as a global cline] into a more general model of the history of diversity' (Nichols 1992:215). However, given the apparent general instability of this feature, the conclusion is not warranted.

The claim of stability for a number of other traits is also unsupported (see Campbell b, Campbell and Poser forthcoming).

If these traits turn out not to be stable, then the speculation that they provide some insight into the structural contents of early human language is without foundation.[16]

5.2 Simple-to-complex

Views common in the 19th century and resurrected in grammaticalization see language as formerly simple, made more complex through time as morpho-syntactic elements were created through grammaticalization (see Heine and Kuteva 2002; Comrie 1992). Heine and Kuteva (2002:394) do not insist overtly on the simple-to-complex trajectory in language evolution, but do argue on the basis of 'grammaticalization theory' for a concrete-to-abstract direction in language evolution and believe that 'at the earliest conceivable stage human language(s) might have lacked grammatical forms such as case inflections, agreement, voice markers, etc., so that there may have existed only two types of linguistic entities: one denoting thing-like, time-stable entities (i.e. nouns), and another one for non-time-stable concepts such as events (i.e. verbs)' – that is, simple-to-complex via grammaticalization.

While it is reasonable to suspect that human language may have began as something more simple that evolved to something more elaborate, it is by no means a necessary assumption, as observed in the complex-to-simple views held by some (below). Speculation along this line sometimes reasons that anything not common in today's languages, or not needed for effective communication, would not yet have emerged in early human language. Thus, for example, it has been supposed that Proto-World would have lacked morphophonemic alternations, tones, vowel nasalization, clicks and various other complex sounds, and affixes (see Comrie 1992); it would have had no tense markers, no aspect markers, definitely no evidential markers, no future markers; it would probably have had only main clauses, or conjunction/subordination only by juxtaposition; no overt copula; etc. While this is not an unreasonable possibility, there is no compelling reason why it had to be the case. For example, for those who believe human emotion played a role in the emergence of language, perhaps early tonal contrasts would not seem unlikely, if they evolved from emotion-laden intonational differences. Evidential markers, for example visible vs. non-visible, could be extremely useful to a hunting society.

Simplicity for ease of production makes a good story, but more complexity for ease of understanding is also reasonable. In the end we shall never know! Would a very simple Proto-World have been mangled beyond recognition by massive later accretions and changes, or would structurally more elaborate language in its early stages have been distorted far beyond recognition because of loss, replacement, and normal analogical and phonological change? Either way, too much change has taken place since the origins of human language ever to know where the truth may lie.

6 Society and language complexity

That early stages of human language(s) may have been complex in structure has also been alleged by some. This view takes encouragement from the often repeated opinion that language becomes more complex in isolated communities or in small-scale societies where most members interact with one another face to face (see Andersen 1988, Hymes 1974, Nettle 1999, Nettle and Romaine 2000, Ross 1996, 1997, Trudgill 1989). Hymes (1974:50) asserted 'the surface structures of languages spoken in small, cheek-by-jowl communities so often are markedly complex, and the surface structures of languages spoken over wide ranges less so.' The earliest speakers of human language(s) probably were members of such small-scale isolated communities, and consequently, according to this claim, may have had complex language(a). But does this view have merit?

The view is often attributed to Jakobson (1929[1962]:82): 'dialects which serve as vehicles of communication in large areas and gravitate towards the role of koiné tend to develop simpler systems than dialects that serve purely local purposes' (Andersen 1988:37), to which Andersen (1988:60) adds, 'dialects that serve predominantly local functions are more prone to elaborate phonetic detail rules.' Later versions of the claim lean not to the tendency towards conservativism but to develop complexity. Andersen speaks of 'relatively open' and 'relatively closed communities,' arguing that 'the greater potential for variablility of usage in open communities favors a more active leveling of irregularities in these, and the lesser variablility a more faithful transmission of morphological irregularity in closed communities' (Andersen 1988:61). He asserts that 'the conservativism of relatively closed dialects is common knowledge' but argues that 'phonetic norm elaboration' is also common in closed dialects (Andersen 1988:62), including ''exorbitant' phonetic changes' (Andersen 1988:73–4). Trudgill (1989:227), speaking of 'high- and low-contact varieties,' extends Andersen's dichotomy to include different languages. For Trudgill, high contact leads to simpler systems: 'dialects which serve a relatively wide socio-spatial function tend to have simpler systems than dialects with a more restricted function' (p.228), that 'in low-contact situations we know that the speed of linguistic change will typically be slow' (p.229), and that 'many of the changes that take place in this sort of situation [low-contact] are of the type that move in the opposite direction complication as opposed to simplification' (p.229). In this way Trudgill sees how the relative greater isolation of Faroese over Danish could perhaps explain the seeming less conservativeness of Danish (p.231), but, how, then, are we to understand the fact that higher-contact Danish is linguistically more conservative in some regards than some of its lower-contact Scandinavian sisters, e.g. in Danish

/k/ is preserved before front vowels, where Swedish and Norwegian have changed it to a fricative.

This notion of isolated, low-contact varieties being conservative is just to opposite of Nettle and Romaine's view (below). Nevertheless, Trudgill (1989:234) does suggest some kinds of 'changes typical of low-contact social contexts.' One is the development of grammatical agreement; however, for example, case and number agreement on adjectives in Finnic languages is generally understood to be do to contact with IE languages. That is, due not a low-contact, but high-contact phenomenon. Another is the 'proliferation of clicks in the Khoisan languages,' but, then, the many clicks in southern Bantu languages are due to language contact, with Koisan languages. The Northwest Coast Linguistic Area of North America is characterized by extensive language contact and extremely elaborates phonemic inventories. Clearly there is no easy correlation of the sort envisaged by Trudgill between relative contact or isolation and structural complexity.

Nettle (1999:138) also argues for 'community size' as a cultural or social variable which may correlate with language structure:

> If a group consists of just a few hundred people, the idiosyncracies of one very influential individual can spread through it very easily. This is not the case if the group consists of thousands or tens of thousands of people. In general, the smaller the community, the greater the probability that a given variant that has no functional advantage at all, but is neutral or slightly disadvantageous, can replace the existing item and become the norm. (Nettle 1999:139)

Nettle and Romaine (2000:12) add 'languages which are used only for in-group communication in small groups can afford complexity.' 'In small language groups innovations and new usages can quickly spread throughout a whole village.' The basic idea in this literature is that such communities, isolated or characterized by face-to-face communication, tolerate eccentricities, and so complexity can grow and highly unusual linguistic traits can become part of the structure of the language.

A problem is that there are many counterexamples, many simple but relatively isolated small languages and many large and non-isolated but complex languages. For example, looking at phonological complexity (from which some of the proponents take their inspiration), we see counterexamples in numerous small and isolated languages such as Rotokas, Pirahã, Hawai'ian, Māori, etc. which have extremely limited phonemic inventories. Rotokas (a 'Papuan' language of Bougainville, 4,000 speakers), has only 7 segments, only 6 consonants; Pirahã (of the small Muran family in Brazil, spoken by only about 150 speakers) has only 8 consonants and 3 vowels (cf. Maddieson

1984). Hawai'ian has only 8 consonants. Isolated South Island Māori, instead of becoming more complex, reduced its 10 consonants, merging /N/ with /k/, leaving 9: /p, t, k, ɸ, h, m, n, r, w/. On the other hand, there are plenty of large non-isolated languages which are complex or exhibit unusual traits, some having become more complex over time. For example, of the Quechua languages, the one spread by the Inca Empire, spoken by several millions, is phonologically very complex, 3 series of obstruents, plain, glottalized and aspirated, at 5 points of articulation (6 in some varieties). Zulu, not small (6,000,000 speakers) nor isolated, with 35 consonants, acquired an extremely elaborate system of click consonants. Eastern Armenian added glottalized stops (under influence from Caucasian languages), now with 29 consonants, which include the 3 series, plain, aspirated, and glottalized stops and affricates, both dental and palato-alveolar affricates and fricatives, etc. Georgian (4,000,000 speakers) is complex (29 consonants), with 3 series of stops and affricates, plain, voiced, and glottalized; uvular stops, etc., and, in morphosyntax, a rich case system, exceptionally complex verb morphology, etc. Even English, probably one of the least isolated languages, has unusual phonological traits, e.g. interdentals, /θ/, /ð/, and 'r' rare in other languages. Arabic, with many millions of speakers, a language of civilization and empire for centuries, not only has interdentals, /θ/, /ð/, it has them and the other coronal fricatives and stops in plain and 'emphatic' (pharyngealized) versison, plus it has pharyngeal fricatives (/ʕ/ and /ħ/). There is far from an easy correlation between size/isolation and complexity. Nothing follows for the structure of Proto-World.

7 What of the structure of the earliest human language(s)?

As just seen, there is good reason to be skeptical about many claims about structural properties of the earliest human language. So, is there anything we can know or reasonably infer about the structure of the earliest human language(s), looking back from modern and attested older languages? The answer is a qualified 'yes,' limited by both logic and content in ways language evolution enthusiasts may not find exciting. This has to do with the design features of human languages. It is argued that the earliest human language will have had the design features of human language and this gives us some clues to its nature.

The logical limitation has to do with definitions and the problem of 'emergence.' Uniformitarianism holds that things about language that are possible today were not impossible in the past and that things impossible today were not possible in the past. This means that whatever is diagnostic of human language today would also have been among the properties of the

earliest human language(s) and that the earliest language(s) would not be characterized by either the presence of things not known in modern languages nor by the absence of things present in all modern languages. So, logically, the earliest language(s) must have exhibited the design features characteristic of human languages today. However, abiding by uniformitarianism means we cannot address 'emergence.' That is, it is generally assumed that there was some earlier non-language communication system (perhaps like other primates' call and display systems) which did not have all the design features of human language but which evolved so that it emerged, as new biological species emerge, crossing the line from non-language to language. However likely it is that such emergence took place, by the uniformitarian principle the point of inquiry is cut off as we go back in time at the point where any form of communication ceases to have the requisite design features that qualify it as language as known today. Thus, while we can speculate about the nature of the earliest human language, looking back from what is known from known languages, we cannot go beyond the logical boundary defined by uniformitarianism without losing empirical constraints and being left in the realm of speculation. We can assume that the earliest language(s) did meet the design feature requirements of human language, but this is in a sense a definitional demarcation which says anything else is not human language, which cuts off access before emergence, leaving unaddressed the question most fascinating to many, of how human language originated and evolved from something that was not (yet) human language.

Accepting the uniformitarian constraint, that anything lacking the design features of human language is not human language, imposes the limitation that the earliest human language that qualifies as such will not have been different in design from languages known in modern times, and that therefore we can assume, though only by default, that the earliest human language which qualifies as language was characterized by these same design features. Let us look at some of the design features that have been proposed and consider what they might mean for the structure of the earliest language(s) (cf. Hockett 1960):

Duality of patterning (double articulation) (recombination of sounds in association with meaning to allow an open-ended number of linguistic signs)

Grammar (fixed or preferred sequences of linguistic elements)

Open-ended word classes (probably at least noun or noun-like and verb or verb-like categories)

Verbal channel (with consonant or consonant-like and vowel or vowel-like segments)

Discourse function of categories (e.g. subject vs. object, agent vs. patient, predicate, etc.)

Multimodality (statements, questions, commands, negation; narrative, conversation)

Synonymy (rephrasability)

Recursion (clauses embedded in other clauses)

Productivity (ability to produce utterly new utterances)

Pantopicality (unlimited by context or topic)

Displacement (reference to the imperceptible things, not in the here and now)

Metalanguage (ability to talk about talking)

Prevarication (verbal deception)

This constitutes my guess as to what 'Proto-World' must have been like: it must have had design features such as these. However, since these features are broad, they do not constrain the form of the earliest human language(s) very much with respect to specific structural traits. They do not help us select the most likely earliest structures from among the variants/parameters known in human languages today. For example, from the design feature of a verbal channel with consonants and vowels, we may infer that probably the earliest human language had consonants and vowels, but whether it had a simple or complex phonemic inventory cannot to be known from this. In the design trait of multimodality, we can infer that the earliest language presumably had means for forming questions, but whether this was with intonation, question particles, inversion of elements, or something else, we cannot know. In the discourse function of categories, presumably the earliest language had means for hearers to distinguish agents from patients, but from this we cannot know whether this may have involved ergative-absolutive, nominative-accusative, or active-stative alignment, whether it involved word order, case marking, cross-referencing, or context and semantic clues. In short, the design features give us some ideas of the nature of the first language(s), but nothing specific, and even relying on them for our guesses about the nature of early language is strained, since by definition, anything not (yet) fitting these conditions is eliminated from consideration. Surely for language evolution, it is precisely those pre-language developments which led to language(s) with all these design features which are most interesting, but about which we can know next to nothing.

8 Conclusion

So, what can we find out or reasonably hypothesize about the earliest human language(s) from looking back from evidence in modern and attested older languages? We can speculate, perhaps even reasonably in some cases, but we can 'know' extremely little. What can we find out from lexical comparisons? Answer: essentially nothing, though we can learn object lessons from the many problems found in the methods which have been utilized to attempt to get at 'global etymologies.' Perhaps because of the assumption that all the world's languages are genetically related, descendants of 'Proto-World,' global etymologists are disposed to believe in etymological connections among words in contemporary languages, and this will to believe permits them to accept as related forms which do not exceed sheer accidental similarity as a more plausible explanation. I conclude with Bender (1993:203), ''global etymologies' are an illusion. They are an artifact of too much freedom of choice and the loss of control.' The global etymologists have not met their burden of proof. In the long time since the origin of human language(s), so much vocabulary replacement has taken place that in effect no forms once found in 'Proto-World' could have survived. Moreover, if some form had survived (and I assert it did not), after so much change it could not be recognized, and, if it should preserve a recognizable shape (and again I assert it could not), there would be so few such surviving forms that it would be impossible to distinguish successful survivors from forms similar by sheer accident. In short, the search for global etymologies is at best a waste of time, at worst an embarrassment to linguistics as a discipline, confusing and misleading those who might look to linguistics for understanding in this area.

What can we find out Proto-World from structural comparisons? Answer: nothing especially useful, though functional typological and structural considerations may provide broad guidelines to what even the earliest human language would have to have in order to qualify as a human language. Again, though, we learn object lessons from the problems encountered in such structural comparisons. In particular, we learn that there is no correlation to be found between size of speech community or social organization and structural aspects of languages. We can speculate that the design features of human language give us a small handle on the necessary nature of the earliest human language(s), but these are so broad that essentially any linguistic structure known in any language today would qualify as possible.

Notes

1 http://members.aol.com/_ht_a/yahyam/page24/protoworld.htm

2 Similar points are made by Rosenfelder 1999. As he explains, based on the **maliq'a* 'suck(le), nurse, breast' example:

> Take a closer look at the list; the rules for this game are evidently quite lax. The vowels are completely ignored. The middle consonant varies from l to ly to lh to n to r to zero. The end consonant ranges from g to j to d to k to q to q' to kh to k' to X to zero. Switching around medial consonants seems to be allowed; extra consonants and syllables can appear where needed. Observe the semantic variation as well: body parts ranging from neck to nape to throat to breast to cheek; actions including swallowing, milking, drinking, chewing, and sucking. Some defenders of Ruhlen and Greenberg make much of the probability of finding such lists among given numbers of families; but notice that one can pretty much pick and choose what languages from a family to include. If Greek doesn't do it for you, try Latin; if Hebrew doesn't work, use Arabic. (Rosenfelder 1999)

3 It might be asked, does not the case become stronger when so many words from so many languages are piled onto a particular putative global etymology? The answer is no: an error does not become a truth through the addition of many more errors of the same sort. 'A bad methodology doesn't become more respectable just by repeating it' (Rosenfelder 1999a). This has been demonstrated often in critiques of mass or multilateral comparison (see Ringe 1992, 1999, for example).

4 Even English *daughter* (Old English *dohtor*, PIE **dhughə̂ter*) fits in view of such forms as *tsuh-ki* and *u-tse-kwa* in the list. Note, incidentally, the considerable overlap between this and Ruhlan's (1994a:192–206) proposed Amerind **t'a?na* 'child, sibling'. Note also, incidentally, that it does not mean 'girl' in any of the languages Greenberg cited, though 'girl' is the gloss assigned to the overall set.

5 Nichols (1998:128) points out that, according to the method, 'after 6,000 years of separation, two languages are expected to exhibit only 7% shared cognates; and 7% represents the lowest number of resemblant items that can safely be considered distinct from chance.'

6 Moreover, given that languages have some vocabulary similarities due to chance, any word that did manage to persist unreplaced since the dawn of human language so many long millennia ago could not be reliably distinguished from sheer accidental similarities. That is, given the extremely small number of such putative survivals, it would be impossible to determine whether they are due to accidental similarity or to inheritance from the very distant past.

7 Baxter and Manaster-Ramer (2000) also compare English and Hindi vocabulary, but their purpose is different form mine. They argue that it is possible to detect the genetic relationship between English and Hindi based on modern data; they compare Hindi and English in a list of 33 'especially basic word-meanings' (p.174) utilizing probabilistic techniques. My point is rather, how little recognizable cognate material remains in these two languages known to be related and how it fares on the methods of global etymology when compared with unrelated languages.

8 I do not have access to Hindi etymological materials, and therefore make judgements about cognacy based on limited knowledge of Indo-European and Sanskrit; I may have missed some true cognates or perhaps misassigned a form as a cognate which is only accidentally similar; I believe, however, not many such errors occur.

9 I thank Miriam Butt, Stephen Fennell, Mate Kapovic, David Nash, Roger Lass, Robert Rankin, and Larry Trask for helpful comments and information with the Hindi forms and their history. Errors are mine.

10 Baxter and Manaster-Ramer (2000:177) identify this set as true cognates, though it is by no means obvious. The PIE form from which English egg comes is **əyo-*, from **ōwyo-*, not an obvious source for the Hindi form, but possible.

11 Hindi *ji:bh* comes from Sanskrit *jihva*:, from pie **dn̥ghū̄*, from whence English *tongue*.

12 Hindi has *he*ŗ*day* 'heart', which is cognate, but *dil* is the common form in use.

13 Hindi *ta:r* may be cognate with *star*, though it is not certain; but *sita:ra*: is a persian loanword, not a direct cognate.

14 The Hindi form is cognate with English *fume*, but this is a loanword in English.

15 For example, if we do not rely strictly on the forms that appear on the Swadesh list, but on what we know from other facts about the history of English and of Hindi, we could extend the list of cognates somewhat, for example:

1	I / maĩ (cf. *me*)
61	die / mar- (cf. *murder*)
69	stand / khaṛa 'standing' (cf. Hindi *tha:* 'was', the true cognate of English *stand*)
93	hot / garm (cf. *warm*)

However, historical facts such as these are not known in the vast majority of comparisons undertaken in attempts to establish global etymologies, and so these forms could not legitimately be used to increase the apparent similarity between English and Hindi for this test. Also, when known historical facts

are taken into account, some cases that might have seemed likely drop out, for example, Hindi *hath*: English *hand*, when we see that Hindi comes from Sanskrit *hásta*, cf. Hittite *kkessar* < PIE **ghesor.*

16 Moreover, even if any did prove stable in Nichols' sense (though the evidence is against this), it could still well be the case that the modern distribution of these traits reflects changes much later in time, recent acquisitions or losses of the traits, much after the advent of Proto-World. Indeed there is historical linguistic documentation to this end for many of these traits in numerous languages (e.g changes to ergativity, development of inclusive/exclusive contrasts, of numeral classifiers, etc.; see Campbell and Poser [forthcoming] for details).

References

Aitchison, J. (1996) *The Seeds of Speech: Language Origin and Evolution*. Cambridge: Cambridge University Press.

Allman, W. F. (1990) *The Mother Tongue*. U.S. News and World Report, Nov. 5, 1990. 109.18:60–70.

Andersen, H. (1988) *Center and Periphery: Adoption, Diffusion and Spread. Historical Dialectology,* ed. by Fisiak, J. 39–83. Berlin: Mouton de Gruyter.

Baxter, W. H. and Ramer, A. M. (2000) Beyond lumping and splitting: probabilistic issues in historical linguistics. *Time Depth in Historical Linguistics*, ed. by Renfrew, C., McMahon, A. and Trask, L., 167–88. Cambridge: The McDonald Institute for Archaeological Research.

Bender, M. L. (1993) Are global etymologies valid? *General Linguistics* 33.191–219.

Bengtson, J. D. (1991) Paleolexicology: a tool towards language origins. *Studies in Language Origins*, vol. 2, ed. by von Raffler-Engel, W., Wind, J. and Jonker, A., 175–86. Amsterdam: Benjamin, J.

Bengtson, J. D. and Ruhlen, M. (1994)a. Global etymologies. *On the Origin of Languages: Studies in Linguistic Taxonomy*, ed. by Ruhlen, M., 277–336. Stanford: Stanford University Press.

Bengtson, J. D. and Ruhlen, M (1994)b. Another look at *TIK 'Finger, One'. *California Linguistics Newsletter* 24.2:9–11.

Bengtson, J. D. and Ruhlen, M (1997) In defense of multilateral comparison. *California Linguistics Newsletter* 25.1:3–4, 57

Campbell, L. (1988) Review of Language in the Americas, by Joseph Greenberg. *Language* 64, 591–615.

Campbell, L. (1997) *American Indian Languages: The Historical Linguistics of Native America.* Oxford: Oxford University Press.

Campbell, L. (1998a). *Historical Linguistics*. Edinburgh: Edinburgh University Press. (1999, American edition: Historical linguistics. Cambridge, MA: MIT Press.)

Campbell, L. (1998b). Nostratic: a personal assessment. *Nostratic: Sifting the Evidence*, ed. by Joseph B. and Salmons J., 107–152. Amsterdam: John Benjamins.

Campbell, L. (1999) Nostratic and linguistic palaeontology in methodological perspective. *Nostratic: Evaluating a Linguistic Macrofamily*, ed. by Renfrew, C. and Nettle, D.179–230. Cambridge: The McDonald Institute for Archaeological Research.

Campbell, L. (In press) How to show languages are related: methods for distant genetic relationship. *Handbook of Historical Linguistics*, ed. by Janda, R. D. and Joseph, B. D. London: Blackwell.

Campbell, L. and Poser, W. (Forthcoming) How to show languages are related.

Campbell, L, Bubenik, V. and Saxon L. (1988) Word order universals: refinements and clarifications. *Canadian Journal of Linguistics* 33:209–230.

Comrie, B. (1992) Before complexity. *The Evolution of Human Languages*, ed. by Hawkins, J. A. and Murray Gell-Mann, 193–210. Redwood City, CA: Addison-Wesley Publishing Co.

Faarlund, J. T. (1990) Syntactic and pragmatic principles as arguments in the interpretation of runic inscriptions. *Historical Linguistics and Philology*, ed. by Fisiak, J. 165–86. Berlin: Mouton de Gruyter.

Foley, W. A. (2000) The languages of New Guinea. *Annual Review of Anthropology* 29.357–404.

Goddard, I. (1975) Algonquian, Wiyot, and Yurok: proving a distant genetic relationship. *Linguistics and Anthropology in Honor of C. F. Voegelin*, ed. by Kinkade, M. D., Hale, K. L. and Werner, O. 249–62. Lisse: de Ridder, P., Press.

Greenberg, J. H. (1957) Genetic relationship among languages. *Essays in Linguistics*, Chapter 3, 35–45. Chicago: University of Chicago Press.

Greenberg, J. H. (1987) *Language in the Americas*. Stanford: Stanford University Press.

Greenberg, J. H. (1996) In defense of Amerind. *International Journal of American Linguistics* 62.131–64.

Harris, A. C. and Campbell, L. (1995) *Historical Syntax in Cross-Linguistic Perspective.* Cambridge: Cambridge University Press.

Heine, B. and Kuteva, T. (2002) On the evolution of grammatical forms. *The Transition to Language*, ed. by Wray, A., 376–97. Oxford: Oxford University Press.

Hock, H. H. (1993) Swallow tales: chance and the 'World Etymology' MALIQ'A 'swallow, throat'. *Chicago Linguistic Society* 29.215–9.

Hock, H. H. and Joseph, B. (1996) *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics*. Berlin: Mouton de Gruyter.

Hockett, C. F. (1960) The origin of speech. *Scientific American* 203.88–96.

Hymes, D. (1974) Speech and language: on the origins and foundations of inequality among speakers. *Language as a Human Problem*, ed. by Morton Bloomfield and Einar Haugen, 45–71. New York: W. W. Norton & Co.

Jacobsen, W. R., Jr. (1980) Inclusive/exclusive: a diffused pronominal category in native Western North America. *Papers from the Parasession on Pronouns and Anaphora,* ed. by Kreiman, J. and. Ojeda, A. E., 204–27. Chicago: Chicago Linguistic Society.

Jakobson, R. (1929) Remarques sur l'évolution phonologique du russe comparée à celle des autres langues slaves. (*Travaux du Cercle Linguistique de Prague,* 2) [Reprinted 1962 *Selected writings of Roman Jakobson*, I: Phonological studies. The Hague: Mouton.]

Jakobson, R. (1960) Why 'mama' and 'papa'? *Perspectives in Psychological Theory*, ed. by Kaplan, B. and Wapner, S., 21–9. New York: International Universities Press. [Reprinted 1962 *Selected writings of Roman Jakobson*, I: Phonological studies, 538–45. The Hague: Mouton.]

Maddieson, I. (1984) *Patterns of Sounds*. Cambridge: Cambridge University Press.

Matisoff, J. A. (1990) On megalo-comparison: a discussion note. *Language* 66.106–20.

McMahon, A. and McMahon S. (1995) Linguistics, genetics and archaeology: internal and external evidence in the Amerind controversy. *Transactions of the Philological Society* 93.125–225.

McWhorter, J. (2001) *The Power of Babel: A Natural History of Language*. New York: Time Books.

Meillet, A. (1925) *La Méthode Comparative en Linguistique Historique*. Paris: Champion. (Translation 1967: *The Comparative Method in Historical Linguistics*. Paris: Champion.)

Meillet, A. (1958) Linguistique historique et linguistique générale. (*Société Linguistique de Paris, Collection Linguistique*, 8.) Paris: Champion.

Murdock, G. P. (1959) Cross-language parallels in parental kin terms. *Anthropological Linguistics* 1.9:1–5.

Nettle, D. (1999)a. *Linguistic Diversity*. Oxford: Oxford University Press.

Nettle, D. and Romaine S. (2000) *Vanishing Voices: the Extinction of the World's Languages*. Oxford: Oxford University Press.

Newmeyer, F. J. (2000) On the reconstruction of 'Proto-World' word order. *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, ed. by Knight, C., Studdert-Kennedy, M., and Hurford, J. R. 372–90. Cambridge: Cambridge University Press.

Nichols, J. (1992) *Linguistic Diversity in Time and Space*. Chicago: University of Chicago Press.

Nichols, J. (1995) Diachronically stable structural features. *Historical Linguistics 1993: Selected Papers from the 11th International Conference on Historical Linguistics*, ed. by Andersen H., 337–56. Amsterdam: John Benjamins.

Nichols, J. (1996) The comparative method as heuristic. *The Comparative Method Revised*, ed. by Durie, M. and Ross, M., 39–71. Oxford: Oxford University Press.

Nichols, J. (1998) The origins and dispersal of languages: linguistic evidence. *The Origin and Diversification of Language*, ed. by Jablonski, N. and Aiello, L., 127–70. San Francisco, CA: California Academy of Sciences.

Picard, M. (1998) The case against global etymologies: evidence from Algonquian. *International Journal of American Linguistics* 64:141–7. [Published also 1995, On the nature of the Algonquian evidence for global etymologies. *Mother Tongue* 24:50–4.]

Rankin, R. L. (1992) Review of language in the Americas, ed. by Greenberg, J. H. *International Journal of American Linguistics* 58:324–51.

Ringe, D.A., Jr. (1992) On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society* 82.1:1–110.

Ringe, D.A., Jr. (1995) 'Nostratic' and the factor of change. *Diachronica* 12.55–74.

Ringe, D.A., Jr. (1996) The mathematics of 'Amerind'. *Diachronica* 13.135–54.

Ringe, D.A., Jr. (1998) Probabilistic evidence for Indo-Uralic. *Nostratic: Sifting the Evidence*, ed. by Joseph, B. and Salmons, J., 153–97. Amsterdam: John Benjamins.

Ringe, D.A., Jr. (1999) How hard is it to match CVC-roots? *Transactions of the Philological Society* 97.213–44.

Rosenfelder, M. (1999) Deriving Proto-World with tools you probably have at home *http://www.zompist.com/proto.html*

Ross, M. (1996) Contact-induced change and the comparative method: cases from Papua New Guinea. *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*, ed. by Mark Durie and Malcolm Ross, 180–217. Oxford: Oxford University Press.

Ross, M. (1997) Social networks and kinds of speech community events. *Archaeology and Language I: Theoretical and Methodological Orientations*, ed. by Blench, R. and Spriggs, M., 209–61. 82–90. London: Routledge.

Ruhlen, M. (1987)a. *A Guide to the World's Languages, volume 1: Classification*. Stanford: Stanford University Press.

Ruhlen, M. (1987)b. Voices from the past. *Natural History* 96(3):6–10.

Ruhlen, M. (1992) An overview of genetic classification. *The Evolution of Human Languages*, ed. by Hawkins, J. A. and Gell-Mann, M., 159–89. Redwood City, CA: Addison-Wesley Publishing Co.

Ruhlen, M. (1994)a. On the Origin of Languages: Studies in Linguistic Taxonomy. Stanford: Stanford University Press.

Ruhlen, M. (1994)b. *The Origin of Language: Tracing the Evolution of the Mother Tongue*. New York: John Wiley & Sons.

Ruhlen, M. (2000) Why kaka and aya? Functional Approaches to Language, *Culture and Cognition: Papers in Honor of Sydney B. Lamb*, ed. by Lockwood, D. G., Fries, P. H., and Copeland, J. E, 521–5. Amsterdam: John Benjamins.

Ryan, P. (2001) Proto-language monosyllables: with their principal meanings. http://www.geocities.com/Athens/Forum/2803/ProtoLanguage-Monosyllables.htm

Salmons, J. (1992)a. A look at the data for a global etymology: *tik 'finger'. *Explanation in Historical Linguistics*, ed. by Davis, G. W and Iverson, G. K., 208–28. Amesterdam: John Benjamins.

Salmons, J. (1992)b. Theory and practice of global etymology. *Proceedings of the 15th International Congress of Linguists* 1.153–5.

Salmons, J. (1997) Global etymology as pre-Copernican linguistics. *California Linguistic Notes* 25.1, 5–7.

Santa Fe Institute (2000) *Annual Research Report*.http: //www.santafe.edu/sfi/organization/annualReport/00/activities/evolution.html.

Song, J. J. (2001) *Linguistic Typology: Morphology and Syntax*. Harlow, England: Longman.

Swadesh, M. (1954) Perspectives and problems of Amerindian comparative linguistics. *Word* 10.306–32.

Tai, J. H.-Y. (1976) On the change from SVO to SOV in Chinese. *Papers from the Parasession on Diachronic Syntax*, ed. by Steever, S. B. Walker, C. A. and Mufwene, S. S. 291–304. Chicago: Chicago Linguistic Society.

Trask, L. (1996) *Historical Linguistics*. London: Arnold.

Trask, L. (1999) *Nostratic: Examining a Linguistic Macrofamily*, ed. by Renfrew, C. and Nettle, D., 157–76. Cambridge: The McDonald Institute for Achaeological Research.

Trudgill, P. (1989) Contact and isolation in linguistic change. *Language Change: Contributions to the Study of its Causes*, ed. by Breivik, L. E. and Jahr, E. H., 227–38. Berlin: Mouton de Gruyter.

Watkins, C. (1985) *The American Heritage Dictionary of Indo-European Roots*. Boston: Houghton Mifflin.