Research Note

The Austronesian Comparative Dictionary: A Work in Progress

Robert Blust and Stephen Trussel

UNIVERSITY OF HAWAIʻI AT MĀNOA AND TRUSSEL SOFTWARE DEVELOPMENT

The *Austronesian comparative dictionary* (ACD) is an open-access online resource that currently (June 2013) includes 4,837 sets of reconstructions for nine hierarchically ordered protolanguages. Of these, 3,805 sets consist of single bases, and the remaining 1,032 sets contain 1,032 bases plus 1,781 derivatives, including affixed forms, reduplications, and compounds. Historical inferences are based on material drawn from more than 700 attested languages, some of which are cited only sparingly, while others appear in over 1,500 entries. In addition to its main features, the ACD contains supplementary sections on widely distributed loanwords that could potentially lead to erroneous protoforms, submorphemic "roots," and "noise" (in the information-theoretic sense of random lexical similarity that arises from historically independent processes). Although the matter is difficult to judge, the ACD, which prints out to somewhat over 3,000 single-spaced pages, now appears to be about half complete.

1. INTRODUCTION.[1] The December 2011 issue of this journal carried a Research Note that described the history and present status of POLLEX, the Polynesian Lexicon project initiated by the late Bruce Biggs in 1965, which over time has grown into one of the premier comparative dictionaries available for any language family or major subgroup (Greenhill and Clark 2011). A theme that runs through this piece is the remarkable growth over the 46 years of its life (at that time), not just in the content of the dictionary, but in the technological medium in which the material is embedded. As Greenhill and Clark put it (2011:553) "POLLEX has followed ever-advancing technology—from typewriter and edge-punched cards, through microfiche and mainframe computers, to wide dispersal on personal computers. Progression to an online database is the next natural step." This important step to an online, openly accessible database was taken in 2011, and as a result the material in it has become increasingly internationalized.

[1] The second author, who has produced the online dictionaries at http://trussel2.com, is responsible in this paper for section 4 (except for the definitions of "Roots," "Loans," "Noise," and "Formosan"), and for all statistics, including the appendix. The first author is responsible for the rest. We would like to thank Larry Hyman for information on the current status of Bantu lexical reconstruction, Doug Cooper for information on the Mon-Khmer Languages Project (MKLP), Daniel W. Bruhn for information on the Sino-Tibetan Etymological Dictionary and Thesaurus (STEDT), and Stefan Georg for information on the Proto-Turkic dictionary that forms part of the Russian Etymological Project.

Oceanic Linguistics, Volume 52, no. 2 (December 2013)

The purpose of this research note is to report on the history and current status of the *Austronesian comparative dictionary* (ACD), a project with some features that parallel POLLEX, but others that distinguish it from POLLEX, from other comparative projects working with Austronesian (AN) language data, and from comparative dictionaries for other language families. We hope, by providing an up-to-date account of where this project now stands, to sketch a broader picture of etymological research on AN languages, and in the process to situate comparative work in this language family within the wider frame of historical linguistics as a discipline.

With around 1,260 languages, Austronesian is second to Niger-Congo in size, and its geographical spread of 206 degrees from Madagascar to Easter Island, and 72 degrees from Taiwan to New Zealand, gave it the largest territorial range of any language family prior to the European colonial expansions of the past five centuries (Lewis, Simons, and Fennig 2013; Blust 2009). Although some of these languages are large (Javanese, with about 90 million first-language speakers, and Malay/Indonesian, with over 200 million first- and second-language speakers, are among the world's ten largest languages), most are quite small. In 2005, the population of Vanuatu, for example, with 105 recognized languages, was 205,754, for an average of about 1,923 speakers per language. However, since this census figure represents total population, not just the indigenous people of the islands, and since a growing (though unknown) proportion of urban indigenous people do not speak any vernacular, this estimate of average language size is higher than would be the case if a finer discrimination were made, and many individual languages, both in Vanuatu and elsewhere, have no more than two or three hundred speakers

It is important to note that the distribution of phylogenetic diversity in the AN language family is highly skewed. Within the past 350 years, the island of Taiwan, with an area roughly that of Holland, was home to at least 24 indigenous languages, of which 14 now survive. However, these 14 languages represent nine of the ten generally recognized primary branches of the family—all others languages, including Yami of Botel Tobago island within the political domain of Taiwan, falling into the enormous Malayo-Polynesian subgroup that contains around 1,235 languages (Blust 1999, 2009). The second very large innovation-defined group in AN is Oceanic, a collection of about 460 languages in Melanesia, most of Micronesia, and Polynesia, which has been the subject of intensive comparative lexical study over the past half century, beginning with key contributions by Milke (1961, 1968) and Grace (1969), and continuing with the landmark publications of Ross, Pawley, and Osmond (1998, 2003, 2008, 2011, to appear).

1.1 DOCUMENTATION. In general, the attention given to language description is strongly correlated with its political importance, which in turn is strongly correlated with its size. Given the number of languages and their typically small size, the first issue encountered in the study of AN languages is, thus, one of documentation: publicly available information for many of the smaller languages is limited to vocabularies of a few hundred words, for example, Reid (1971) or McFarland (1977) for the minor languages of the Philippines, Tryon (1976) for the languages of Vanuatu, and Tryon and Hackman (1983) for the languages of the Solomon Islands. In some cases, these wordlists represent older sources that are not always phonetically reliable (for example, Ray 1913 for the languages of Borneo).

Despite this problem, there are many excellent descriptions of AN languages representing all geographical regions and most major subgroups of the family. Some of the larger languages are represented by extensive dictionaries, in a few cases (as for Malagasy, or various lowland languages of the Philippines) dating from the seventeenth century. Among very large dictionaries (none of which has a reverse index), English (1986) is a Tagalog–English dictionary of 1,583 pages, each 8.25 x 5 in. in 10-point type and double columns; Hardeland (1859) is a Ngaju Dayak–German dictionary of 638 pages, each 8.5 x 5.5 in. in 8-point type and double columns; Wilkinson (1959) is a Malay–English dictionary of 1,291 pages, each 9.5 x 7 in. in 10-point type and double columns; Zoetmulder (1982) is an Old Javanese–English dictionary of 2,368 pages, each 9.5 x 6 in. in 10-point type and double columns; Beaujard (1998) is a Malagasy–French dictionary (Tañala dialect) of 891 pages, each 9.5 x 6 in. in 10-point type and double columns; and Verheijen (1967) is a Manggarai–English dictionary of 772 pages, each 9.25 x 6.25 in. in 10-point type and double columns. Few other dictionaries match these in size, but there are many excellent dictionaries that are nonetheless very substantial, and there is also a theoretically and qualitatively diverse collection of grammars for languages scattered throughout the family. As noted in Blust (2009:xxiii), the problem of whether the documentation of AN languages should be considered sufficient to justify the kinds of historical inferences that are based on them is reminiscent of the philosophical conundrum of whether a glass that contains water up to the mid-point is half-full or half-empty: many languages remain poorly described, but very often we have good descriptions of other languages that are not very distantly related. All things considered, then, the resources available for comparative work in AN can be considered excellent.

In the most critical area, Taiwan (where nine of the proposed ten primary branches of the family are represented), the descriptive resources for lexical comparison have improved dramatically in recent years. Of the 14 languages that are still spoken or that have only recently become extinct, reasonably good published dictionaries exist for six: Paiwan (Ferrell 1982), Amis (Fey 1986), Pazeh/Pazih (Li and Tsuchida 2001), Thao (Blust 2003), Kavalan (Li and Tsuchida 2006), and Puyuma (Cauquelin to appear). These six languages represent five of the probable nine primary branches of the AN language family in Taiwan, leaving Atayal, Seediq, Saisiyat, Bunun, Tsou, Kanakanabu, Saaroa, and Rukai without adequate dictionaries. Less adequate or unpublished dictionaries exist for Atayal, the Truku dialect of Seediq, Bunun, and Tsou, covering three of the remaining four likely primary branches.

1.2 KNOWLEDGE. Given its size and the wide range of typological diversity among its languages, AN has rarely been approached in its entirely by a single scholar. Instead, most comparativists have focused on particular areas: there are Austronesianists who are primarily or exclusively Philippinists, Micronesianists, Polynesianists, specialists in the languages of Sulawesi, the Moluccas, or Vanuatu. Each of these is to some degree faced with the type of problem memorialized in the fable of the seven blind men and the elephant: *Austronesian* to them is often defined largely by the subset of languages with which they are familiar, and to the extent that they do not normally look beyond this artificial limitation, their perception of the family is distorted by areal or subgroup-defined features that they may mistakenly assume to be more general. Those scholars who have specialized in the study of the Formosan languages have been at less of a disadvantage in this respect, since to the extent that their work is truly comparative they must deal with a wide range of genetic and typological variation among a small number of geographically restricted languages.

Among the few scholars whose work has in principle spanned the entire AN language family are the Dutch Sanskritist and comparativist Hendrik Kern (1833–1917), who made a number of contributions to the study of AN languages in the Philippines, Indonesia, and Melanesia; the German comparativist Otto Dempwolff (1871–1938), who laid the foundations for the comparative phonology of the AN languages in his seminal three-volume reconstruction and discourse on method (Dempwolff 1934–1938); the American comparativist Isidore Dyen (1913–2008), whose contributions ranged over much of the language family from Taiwan to Micronesia; and the first author of this research note, whose publications have ranged over much of the language family from Taiwan to Polynesia. This situation is in striking contrast to the division of labor in Indo-European comparative linguistics, where a given scholar may have greater knowledge of say, Germanic, Slavic, or Indo-Iranian, but where everyone is basically trained in the fundamental triad of Sanskrit, Greek, and Latin, with additional attention to Gothic and Old Church Slavonic. Whereas most Indo-Europeanists share a core body of established comparative knowledge, then, and may be relatively ignorant only of the darker corners of the family (Albanian, Armenian, Baltic, Celtic, or Tocharian), most Austronesianists are actively familiar with only a geographically or genetically circumscribed part of the core body of established comparative knowledge.

In Indo-European terms, the typical AN comparativist is, therefore, not the equivalent of an Indo-Europeanist, but is more like a Germanist, a Slavicist, or a Sanskritist. A few, such as the pioneering Indonesianist H. N. van der Tuuk (1824–1894), the methodical Swiss systematizer Renward Brandstetter (1860–1942), or the wide-ranging English comparativist Sidney H. Ray (1858–1939) did comparative work that covered a wider range of languages than most, but fell short of a comprehensive treatment of the language family as a whole. Despite his searching comparative treatment of AN languages from Taiwan to eastern Indonesia (and west to Madagascar), for example, Brandstetter (1916 and elsewhere) deliberately excluded the AN languages of the Pacific, both such non-Oceanic languages as Palauan and Chamorro, for which usable descriptive materials existed in his day, and most dramatically the entire Oceanic group. Similarly, Ray (1926) worked with extensive comparative materials for the languages of Melanesia, but approached the AN languages of insular Southeast Asia only in Ray (1913), a valuable set of vocabularies for the languages of Borneo, but one that he did not collect himself, and which did not lead to any other publication on AN languages outside Melanesia.

The compartmentalization of comparative scholarship in AN is clearly a direct outgrowth of the size and enormous geographical spread of this language family, together with unevenness in documentation (language sizes, which correlate closely with amount of documentation, are largest in the lowland Philippines and western Indonesia-Malaysia, then dip considerably in eastern Indonesia and the western Pacific, before rising again to much more moderate highs in Fiji and western Polynesia). Given the rarity of truly comprehensive scholarship for the entire AN language family, it is fair to ask whether anyone has sufficient knowledge to do broad comparative work on these languages and reach reliable inferences about their history.

1.3 TIME. The last consideration to keep in mind in evaluating the difficulty of doing an AN comparative dictionary is time; even with the huge volume of work that has already been done, given the enormous amount of data that must be processed, one must seriously ask how long it will take to produce a more-or-less complete compendium of etymologies that can be extracted from the descriptive materials now available. In many ways, doing a comparative dictionary for a language family of this size is like counting the stars: there are times when it seems that it will never end. The ACD, like any dictionary, may never be "finished" in the absolute sense that this word conveys, but the prospects of reaching a more-or-less complete body of work that can be passed on to future generations to refine are now more sanguine than might initially have been expected.

Over the past three years, about 1,000 single-spaced pages of new material have been added to the dictionary by Blust, while by no means working full-time on the project. It is impossible to know exactly how much more remains to be found and incorporated, but if we use Dempwolff (1938) as a standard of measurement, we are given some idea of the work that remains to be done. Around 1,159 of the 2,216 comparisons in Dempwolff (1938) have been checked and thoroughly reexamined in the light of the comparative data now available. As a result, a number of Dempwolff reconstructions have been rejected as probable "ghosts," resulting from widespread Malay loanwords in western Indonesia, and in some cases the Philippines. It seems likely that over 600 of the 2,216 etyma in Dempwolff (1938) will have to be treated in this way. As noted below, the dismissal of a Dempwolff comparison does not mean that it is entirely ignored. Instead, it may be assigned to "Loans," a file that contains loanwords with a distribution that is sufficiently wide to potentially mislead comparativists into proposing protoforms that may never have existed.

If the figure of 1,159 out of 2,216 is taken at face value, work on the ACD is about half finished. However, this assumes that, throughout the duration of the project, the only material that will be considered will be the reconstructions in Dempwolff (1938)—which is not how most searching to date has been done. During the funded period from 1990–1995 the strategy was to exhaustively search narrow segments of the lexicon in some 200–250 languages. For example, all potential reflexes of the protosequence *ba- through *bad- were collected and scanned visually for cognate relationships, then all potential reflexes of the protosequence *bag- through *bak-, and so forth, incorporating or rejecting Dempwolff material in the process, but always reaching beyond it. To save duplication of effort, potential reflexes of vowel-initial forms (*a, *e, *i, *u) were searched together with the same sequences preceded by *h-, *q-, or *S-, since the latter phonemes disappeared in most daughter languages, and searching V- and then *hV-, *qV-, or *SV- would have meant repeating nearly all of the same work. In addition, the *b and *w sections (the former very large and the latter very small) were worked in great detail during this period.

As a result of this effort, the most thoroughly researched portions of the ACD so far, and hence the ones that are least likely to be changed by further searching, are reconstructions that begin with a vowel (*a, *e, *i, *u), *b, *h, *q, *S, and *w. A comparison of the number of protoforms beginning with each of these protophonemes in Dempwolff (1938) and the ACD is instructive, as shown in table 1 (though see below regarding *h, *q, and *S).

Due to errors in his reconstruction of the Proto-Austronesian "laryngeals" that were corrected by Dyen (1953), no direct comparison can be made between Dempwolff's *h and *h, *q, or *S in the ACD, although all of the latter correspond to Dempwolff's *h and sometimes to zero where Dempwolff failed to reconstruct the first syllable of a trisyllable, as with *hapejes (his *pəg′ət′) 'to smart, sting', *qasawa (his *[t′]ava[‘]) 'spouse', or *SadiRi (his *ḍiγi') 'housepost'. If *h, *q, and *S are taken *in toto* as equivalent to Dempwolff's *h, then the figure for these is 84 : 650, for a ratio of 1 : 7.7. In every case, the number of forms posited in the ACD greatly exceeds the number in Dempwolff (1938), by anywhere from two to more than seven and one-half times. Taken at face value, then, the *k section of Dempwolff (1938), which contains 214 entries, could grow to over 850 reconstructions if it is done as thoroughly as the *b section has already been done, whereas so far it contains only 343 entries in the ACD. Similarly, the *t section of Dempwolff (1938), which contains 302 entries, could contain over 1,200 reconstructions, whereas so far it contains only 306. These figures leave us in limbo with regard to estimating how much more work needs to be done, and, therefore, how long it will take. Moreover, work during the past three years has tended to focus more on filling obvious gaps and strengthening or rejecting Dempwolff reconstructions rather than exhaustively searching new sections of the lexicon, introducing another element of uncertainty. However, if by an optimistic estimate the work is about half done and has taken about eight years of concentrated effort to achieve, it should be possible to bring the ACD to a state of near-completion in another eight years.

TABLE 1. COMPARATIVE SIZES OF RECONSTRUCTION GROUPS IN DEMPWOLFF (1938) AND THE ACD

Protoform set	Dempwolff	ACD	Ratio
*a	95	241	1 : 2.5
*b	280	1,015	1 : 4.0
*e (schwa)	18	125	1 : 7.0
*i	60	203	1 : 3.4
*u	66	141	1 : 2.1
*w	11	69	1 : 6.3

2. HISTORY.

The first scholar who had sufficient knowledge to seriously approach the reconstruction of Proto-Austronesian was the German medical doctor and linguist Otto Dempwolff who, despite shortcomings that we can see today, demonstrated that it was possible to infer the phonology of a language that he called *Uraustronesisch* (generally equivalent to what is now called Proto–Malayo-Polynesian) based solely on the evidence of three carefully chosen languages—Tagalog, Toba Batak, and Javanese—which served as representatives for perhaps 200 languages that he had already examined closely in a comparative context (Dempwolff 1934).

Dempwolff's reconstructed phonology was then tested on eight other languages in Indonesia, Melanesia, and Polynesia (Dempwolff 1937), and his success in coping with an immense body of data was clearly reflected in volume 3 of his landmark trilogy: a comparative dictionary with 2,216 explicit reconstructions arrived at by strict application of the comparative method, in which all irregularities were carefully noted (Dempwolff 1938). For over thirty years, this collection of reconstructions, drawing on supporting evidence from a total of eleven languages, remained a static repository of comparative linguistic data that other scholars used in discussing problems of comparative phonology, but which was not expanded through further lexical comparison. The only marginal exception to this statement was the work of the German comparativist Wilhelm Milke, which was limited to languages of the Oceanic subgroup, and provided the initial impetus for the reconstruction of Proto-Oceanic as a separate entity within the Austronesian colossus.

The post-Dempwolffian stasis in etymological research ended with the appearance of Blust (1970), a set of 443 new etymologies deliberately chosen to represent 20 percent of Dempwolff (1938). This publication was followed by seven other collections of etymologies from 1972 to 1989 (Blust 1972a, 1972b, 1973, 1980, 1983/84, 1986, 1989) that, together with Mills (1981), more than doubled the inventory of AN reconstructions with supporting evidence. From the beginning, most of these comparisons conformed to stricter distributional requirements than the material in Dempwolff (1938). Whereas Dempwolff proposed many reconstructions based only on comparisons represented by Malay and a few other languages of western Indonesia that have a known history of borrowing from Malay, Blust (1970) required comparisons to be minimally supported by data from a language of the Philippines and a geographically noncontiguous language of western Indonesia. This requirement in turn initiated a renewed interest in the subgrouping of the AN languages, following the highly ambitious but generally disappointing lexicostatistical study of Dyen (1965). Beginning with Dahl (1973), Blust (1974, 1977), and Mills (1975), a new higher-level subgrouping of the AN languages began to emerge, based on evidence of exclusively shared innovations in cases where these could clearly be distinguished from shared retentions, and these basic outlines were further elaborated in subsequent publications such as Blust (1999).

By the late 1980s, with over 2,800 new lexical reconstructions scattered through at least ten publications, requests for a united set of new etymologies began to increase. In response to this interest, Blust applied for and obtained a grant from the National Science Foundation for the period 1990–1993. Work began slowly due to other commitments, but accelerated considerably at the beginning of the second year. When the funding period expired, funds were still available, as they had been used sparingly, and a no-cost extension of the grant was given until 1995. During this period of funded research, the primary collection and analysis of data was done entirely by Blust, with graduate assistants doing the inputting and formatting of data, and general technical assistance provided by David Stampe, who created the original computational design of the dictionary in Emacs. When the project came to a close in 1995, the ACD contained about 2,045 single-spaced pages of analyzed data representing 3,360 cognate sets.

Fieldwork on some of the most endangered aboriginal languages of Taiwan then became a research priority for Blust, and work on the dictionary, which probably was about 25 percent complete at that time, was suspended indefinitely. Despite its inchoate state, the ACD proved of value to other researchers, most notably Ross, Pawley, and Osmond, who began to use it at the outset of their monumental multivolume project, *The lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society*, and have continued to do so through the most recent volume (Ross, Pawley, and Osmond 1998, 2003, 2008, 2011, to appear). Because it was available on the Internet, the ACD also came to the attention of other scholars working on AN languages, and in 2009 Richard Nivens of the Summer Institute of Linguistics, Indonesia Branch, who had received his doctorate in linguistics at the University of Hawai‘i in 1998, contacted Blust by email and offered to improve the computational structure of the online file by eliminating ambiguities, inconsistencies in formatting, variable use of conventions, and so on. In doing this, he produced a Shoebox version of the original Emacs file that had been created by David Stampe. Although the internal logical consistency of the file and the general display were improved, no new material was added during this period.

In February 2010, at a chance dinner meeting arranged by Daniel Koch, Blust mentioned the ACD to Steve Trussel, who had abandoned ship just before completing his doctorate in linguistics at the University of Hawai‘i, and built a career as a custom software developer. Because Trussel, who loves large databases, was clearly interested in seeing the material and having a chance to work on it in his own way, we almost immediately began to correspond about revising and expanding the existing file. After 15 years of near stasis, then, the ACD suddenly revived, and the improvements that Trussel made were so quick and so dramatic that within a very short time the *Austronesian comparative dictionary* became a truly joint project. In the three years since this collaboration began, the ACD has undergone fundamental improvements in online display, in search features, and in various other ways that will be described below, as well as growing in page count by nearly 50 percent through the addition of many new reconstructions and hundreds of pages of supporting evidence, as sketched in table 2.

TABLE 2. ADDITIONS TO THE ACD AFTER THE BEGINNING OF THE BLUST-TRUSSEL COLLABORATION

Legacy data (sets)		3,360
New sets	2010	386
	2011	545
	2012	474
	2013	72
Total additions from 2010		1,477
Current total		4,837 (=3,360 + 1,477)

3. THE LINGUISTIC STRUCTURE OF THE ACD.

This section will provide a description of the linguistic structure of the ACD, which includes everything connected with the form of entries and the kinds of conventions used. The computational structure of the dictionary, which appears as soon as the website is accessed, will be described in section 4.

Features that are common to all entries in the ACD are the following:

- Entries begin with an abbreviation for the protolanguage to which the etymon is assigned. This is followed on the same line by the reconstructed form marked by an asterisk, indicating that it has been inferred by application of the comparative method, and then a gloss.
- The next line contains supporting evidence, beginning with "Formosan," or with a subgroup label followed by language names that are generally cited in a north-to-south and west-to-east order, then the reflex of the reconstructed form and a gloss that is an exact or nearly exact copy of the meaning given in the primary source.

An example of a maximally simple comparison that contains only these features is the following:[2]

(1) PAN *qeCeŋ 'obstruction, barrier'

Formosan: Paiwan *qetseŋ* 'barrier, fence, enclosure; a "no entry" sign to humans or evil spirits (e.g., a stick left in certain position in front of house)'

PMP *qeteŋ 'obstruction, barrier'

WMP: Kayan *teŋ* 'dam in the river'
Karo Batak *henteŋ* 'lie athwart, lie across a path'

Reconstructions in the ACD are assigned to any of nine levels:[3]

1. PAN (Proto-Austronesian)
2. PMP (Proto–Malayo-Polynesian)
3. PWMP (Proto-Western Malayo-Polynesian)
4. PPh (Proto-Philippines)
5. PCEMP (Proto–Central-Eastern Malayo-Polynesian)

[2] Language names and protolanguage abbreviations are given as they appear in the ACD. There are some cases where the name or abbreviation we give differs from the *current* standard or common usage or the standard *Oceanic Linguistics* style, as, for example, with Nggela, now usually written Gela, or PAN, usually written PAN in this journal. For language names, we used the orthography of the major sources (in this particular case, Fox 1955), and to avoid internal confusion we have retained the names we originally used. (Alternate names are cross-referenced in the language index that appears at the end of this volume.) Abbreviations and the spelling of other terms (e.g., protolanguage rather than proto-language) follow *Oceanic Linguistics* style except in what are basically direct quotations from the ACD.

[3] An explanation of some of these subgroup labels is as follows:

MP: Malayo-Polynesian = all AN languages outside Taiwan.

WMP: Western Malayo-Polynesian = the AN languages of the Philippines, Borneo, the Malay peninsula, and islands in peninsular Thailand and Burma, Sumatra, Java and its satellites, Bali, Lombok, western Sumbawa, Sulawesi, Palauan and Chamorro of western Micronesia, the seven or eight Chamic languages of mainland Southeast Asia and Hainan Island, and Malagasy.

CEMP: Central-Eastern Malayo-Polynesian = CMP + EMP.

CMP: Central Malayo-Polynesian = the AN languages of the Lesser Sunda islands and the southern and central Moluccas of eastern Indonesia.

EMP: Eastern Malayo-Polynesian = SHWNG + OC.

SHWNG: South Halmahera-West New Guinea = the 30–40 AN languages of southern Halmahera and the northern Bird's Head peninsula of New Guinea.

OC: Oceanic = the roughly 460 AN languages of Melanesia, Micronesia, and Polynesia except Palauan and Chamorro of western Micronesia.

Unlike these subgroup labels, "Formosan" is used as a cover term for the aboriginal languages of Taiwan which, as noted earlier, appear to belong to at least nine primary branches of the AN language family. In addition, PWMP may not be a valid subgroup, and some forms that are currently assigned to it may have been found in PMP.

6. PCMP (Proto-Central Malayo-Polynesian)
7. PEMP (Proto-Eastern Malayo-Polynesian)
8. PSHWNG (Proto-South Halmahera-West New Guinea)
9. POC (Proto-Oceanic)

Reconstructions at each of these levels have the following requirements:

1. PAN reconstructions require evidence from at least one Formosan language, and one language outside Taiwan.
2. PMP reconstructions require evidence from at least one WMP and one CEMP language.
3. PWMP reconstructions generally require evidence from at least one language of the Philippines and a geographically noncontiguous language of western Indonesia (comparisons limited to languages of western Indonesia are occasionally made where it is felt that borrowing from Malay is unlikely, and the languages are geographically separated).
4. PPh reconstructions require evidence from at least one language of the northern Philippines and one language from the central or southern Philippines, or from either of these and one of the Sangiric or Minahasan languages of northern Sulawesi.
5. PCEMP reconstructions require evidence from at least one CMP language and one EMP language.
6. PCMP reconstructions require evidence from at least one language of the Lesser Sundas or southern Moluccas and a geographically noncontiguous language of the central Moluccas.
7. PEMP reconstructions require evidence from at least one SHWNG language and one OC language
8. PSHWNG reconstructions require evidence from at least one language of southern Halmahera and another from the Bird's Head region of northwest New Guinea.
9. POC reconstructions require evidence from at least one language of the Admiralty islands and any other Oceanic language or, more liberally, from any Oceanic language of New Guinea and its satellite islands, and another from the Southeast Solomons, Micronesia, Vanuatu, southern Melanesia, or Polynesia. The inclusion of POC reconstructions overlaps with the work being undertaken by Ross, Pawley, and Osmond (1998, 2003, 2008, 2011, to appear), and, needless to say, the benefits have been mutual.

Examples of somewhat more complex comparisons that contain additional features are seen in entries (2) and (3):

(2) PWMP *kudis 'scurfy skin disease; scabies' [doublet: *kuris]

WMP: Ilokano *kúdis* 'skinned, flayed, excoriated, peeled'
Malay *kudis* 'scurfy skin disease, esp. true itch or scabies, but also used of mange and shingles'
Old Javanese *kuḍis-en* 'scurvy, scabby'
Javanese *kuḍis* 'scabies'
kuḍis-en 'have or get scabies'
Sasak *kudis* 'a skin disease, *ichthyosis*'

NOTE: also Ilokano *kúrad* 'contagious affection of the skin characterized by the appearance of discolored whitish patches covered with vesicles or powdery scales, and at times itching greatly; a kind of tetter or ringworm’, Karo Batak *kudil* ‘scabies’, *kudil-en* ‘suffer from scabies’, Javanese *kuḍas* ‘ringworm’, Sasak *kurék* ‘scabies, itch’.

(3) PMP *handem ‘think, understand’

WMP: Hiligaynon *hándum* ‘wish, desire, ambition’
Javanese *arem* ‘to sit brooding’
Sasak *arem* ‘think, understand’

POC *adom ‘think, understand’

OC: Nggela *ando* ‘think, understand’
Sa‘a *aro* ‘to brood (cited only in the English index)’

NOTE: with root *-dem ‘think, ponder, brood, remember’

Additional features that entry (2) brings to light are the citing of doublets on the reconstruction line, and the use of notes. Doubletting that cannot be traced in any clear way to borrowing is extremely common in AN languages (Blust 2011), and an effort has been made to cross-reference doublets in the ACD wherever possible. Given the number of comparisons that must be considered, it is likely that not all doublets are currently cross-referenced, but the goal is to see that they ultimately are. A distinction is further drawn between doublets (variants that are independently supported by the comparative evidence), and “disjuncts” (variants that are supported only by allowing the overlap of cognate sets). To illustrate, both Tagalog *gumí* ‘beard’ and Malay *kumis* ‘moustache’ show regular correspondences with Fijian *kumi* ‘the chin or beard’, but they do not correspond regularly with one another. Based on this evidence, it is impossible to posit doublets, since unambiguous support for both variants is lacking. However, since the Tagalog and Malay forms can each be compared with Fijian *kumi*, two comparisons can be proposed that overlap by including the Fijian form in both (like all Oceanic languages, Fijian has merged PMP *k and *g; in addition, it has lost final consonants) . The result is a pair of PMP disjuncts *gumi (based on Tagalog and Fijian) and *kumis (based on Malay and Fijian), either or both of which could be used to justify an independent doublet if additional comparative support is found.

Unlike many comparative dictionaries, the ACD is annotated. Some notes are several lines, while others are a page or more. Notes are used for a variety of purposes. Among the most common are to report other forms that show a likely historical connection with those cited in the main comparison, but which exhibit irregularities other than the usual sporadic assimilation or metathesis, and so raise more serious questions about comparability, as in entry (2) above; to discuss details of the reconstructed gloss; and to note the occurrence of monosyllabic “roots” or submorphemic sound-meaning correlations in reconstructed morphemes. Entry (3) illustrates the last of these purposes, since the sequence *-dem is also found in PAN *demdem ‘think, ponder, consider’, PWMP *qedem ‘think, brood, remember’, and PWMP *tadem ‘remember’, and occurs in many attested languages, as with Tboli *hedem* ‘to think; to ponder; to remember something’, Proto-Sangiric *taRəndum ‘memory; to remember, think of’, Tiruray *tedem* ‘to remember something’, and so on (Blust 1988). Another feature of entry (3) that is also seen in entry (1) is the change of phonemic shape between a morpheme in a higher-order (earlier) protolanguage and its historical continuation in a lower-order (later) protolanguage: PAN *qeCeŋ to PMP *qeteŋ in entry (1), and PMP *handem to POC *adom in entry (3). Conventions regarding the citation of protoforms that differ in shape or meaning (or both) in different protolanguages that are represented within a single comparison will be explained more fully in entry (4), a considerably longer comparison that also contains other features that have not yet been discussed.

(4) PAN *qaCi 'to ebb, of water in streams'

Formosan: Kanakanabu *ʔ-um-á-ʔaci* 'dam up a side stream to catch fish'

PMP *qati 'to ebb, of water in streams; low tide' [doublet: *qeti]

WMP: Ilokano *atí* 'dry, evaporated, dried out, waterless. Exhausted in its supply of liquid'
Wolio *ati* 'land, sandbank, shoal, shallow water; ebb, low tide'

PWMP *ka-qati 'low tide'

WMP: Pangasinan *káti* 'low tide; go out, of the tide'
Tagalog *káti* 'low tide, low ebb; land not covered by sea'
Wolio *ka-ati* 'shallowness, low tide'

PAN *ma-qaCi 'to ebb, of water in streams'

Formosan: Bunun *ma-hciʔ* 'to dam up a side stream to catch fish' (Tsuchida 1976)

PMP *ma-qati 'to ebb, of water in streams; low tide' [doublet: *qeti]

WMP: Wolio *ma-ati* 'shallow'
Chamorro *maʔte* 'low tide'
maʔte i tasi 'the sea is at low tide'

OC: Wuvulu *maʔi* 'low tide'
Seimat *mat* 'tide'
Bipi *mak* 'reef'
Lindrou *mek* 'reef; low tide'
Loniu *mat* 'low tide'
Lou *met* 'reef; dry reef'
Mussau *mati* 'low tide; dry reef'
poŋa-mati 'coral reef'
Tigak *mat* 'reef'
Nakanai *mahati* 'be out, of the tide; low tide; dry season'
Mbula *magat* 'low tide; dry reef'
Manam *mati* 'reef'
mati i bara 'ebb, ebb-tide; low-water (the reef is dry)'
Eddystone/Mandegusu *mati* 'low tide'
Nggae *maɣati* 'reef'
Lau *mai* 'ebb tide; reef; dry reef; to ebb'
Kwaio *mai* 'low tide'
Saʻa *mäi* 'ebb tide, low tide'
Ulawa *mäi* 'low spring tides in August'

ʻĀreʻāre *mai* ʻlow tide, ebb tideʼ
Arosi *mai* ʻlow tide, ebbʼ
Anejom *mas* ʻlow tideʼ
Rotuman *mafi* ʻlow-tide water as containing fish; tide in generalʼ
Fijian *mati* ʻto ebb, of the tide, as opposed to the flow; part of the reef exposed at low tideʼ

NOTE: also Kapampangan *kati(h)* ʻlow water-levelʼ, Mansaka *atiʔ* ʻdry up, (of water); evaporateʼ; Lou *ra-met* ʻreef; dry reefʼ, Mota *meat* ʻebb, low tideʼ

A feature that appears in entries (1), (3), and (4) is a change in phonemic shape between a higher-level (earlier) protolanguage and a lower-order (later) protolanguage. In entries (1) and (4), this is seen in moving from PAN (which distinguished *C from *t) to PMP (which did not), while in entry (3), it is seen in moving from PMP to POC, where PMP *h dropped and PMP *e (schwa) became POC *o. As a general convention, whenever a form or meaning in a lower-order protolanguage differs from its antecedent form or meaning in a higher-order protolanguage, it is written out in full: so in entry (4), PMP *qati ʻto ebb, of water in streams; low tideʼ is written out in full because both the shape of the protoform and its full range of meaning (at least as inferred here) differ from PAN *qaCi ʻto ebb, of water in streamsʼ. Since reconstruction is undertaken for nine distinct protolanguages from PAN to POC, entries with widely distributed cognate sets typically provide evidence for protoforms on multiple levels. In the case of entry (4), protoforms are implied for PAN, PMP, PWMP, PCEMP, PEMP, and POC. As can be seen, not all of these are indicated explicitly. In general, if a reconstruction at any node in the AN family tree has the same shape and meaning as its immediately ancestral form, it is not repeated. PMP *ma-qati ʻto ebb, of water in streams; low tideʼ is, thus, written out in full because it differs from the PAN form, but the POC reconstruction is not written, since it is identical in all inferable respects to its PMP, PCEMP, and PEMP antecedents. In effect, although many comparisons encode information about multiple protolanguages, some lower-order protoforms are explicit and others implicit (needless to say, protoforms in PAN are necessarily explicit).

A second feature of entry (4) that is not seen in the previous three entries is the appearance of subentries for affixed forms of the base. The ACD differs from almost all other comparative dictionaries not only in indicating the level of reconstruction of each etymon based on explicit subgrouping criteria, and in being annotated, but also by including reconstructions of affixed forms, reduplications, and compounds wherever there is comparative evidence to support them. The result is a corpus not only of lexical bases, but also of material on comparative morphology. Entry (4) is a relatively modest example, with just two subentries, but to gain an appreciation for how complex some of the entries in the ACD are, it should be noted that PAN *aNak ʻchildʼ, which is 15 single-spaced pages in print form, is supported by reflexes in 101 languages, and the main entry is followed by 46 subentries, which include prefixed, suffixed, and circumfixed forms of the base, prefixed and suffixed forms of the base, partial and full reduplications, and a number of compounds, such as PMP *anak apij ʻtwinʼ, PMP *anak bahi/ba-bahi/b<in>ahi ʻwife-taking groupʼ, PMP *anak buaq ʻrelativeʼ, PWMP *anak daRa ʻvirgin, girl of marriageable ageʼ, PMP *anak ma-Ruqanay/(la)-laki ʻwife-giving groupʼ, PWMP *anak i haRezan ‘step or rung of a ladder’, PMP *anak i mata ‘pupil of the eye’, and *anak i panaq ‘arrow’. Similarly, PAN *aCay ‘death’ is 16 printed pages, with 26 subentries and a note of 99 lines (one and one-half pages), while PAN *kaen ‘to eat’ is 19 pages in print form, includes 23 subentries, and draws material from over 170 languages.

One other feature of the linguistic structure of the ACD that does not appear in any of the entries cited above is the use of subscripts to distinguish phonemically identical proto-forms, as with PMP *a$_1$ ‘article’, PMP *a$_2$ ‘conjunction: and’, PMP *a$_3$ ‘exclamation, interjection’, or PWMP *bali$_1$ ‘become, happen’ and *bali$_2$ ‘equal, equivalent’.

4. THE COMPUTATIONAL STRUCTURE OF THE ACD.

The *Austronesian comparative dictionary* is an alphabetical arrangement of a large number of words and meanings, but it is for answering neither the question “what does this word mean” nor “how do you spell it.” Rather, it is a presentation of a hypothetical reconstruction of the vocabulary ancestral to the AN language family.

Because of the special nature of a comparative dictionary, a user faces unusual problems in locating an entry. An alphabetical arrangement of reconstructed words makes it possible to locate an entry for a word that is known. But how would one find, for example, whether the ancestor of a modern-day word was included? With a printed copy in hand, one could browse through it, but what about online?

Comparative dictionaries are rare, and online versions rarer still, but generally they provide the user with various sorts of search forms. Such is the case, for example, with the online version of *A comparative dictionary of Indo-Aryan languages* (Turner, 1962–85), the SEAlang *Mon-Khmer etymological dictionary* (2013), and the Sino-Tibetan etymological dictionary and thesaurus (Matisoff, 1987–2013).

The ACD takes the approach of providing a full-text display of the dictionary, similar to what would be found in a print version, but taking advantage of the linking functions inherent in the Internet browser screen. It is browsable like a print dictionary, but includes numerous indexes and supporting sections that link extensively to the main dictionary.

Before examining some aspects of the underlying data structure and the compiler’s interface, we will briefly examine the surface realization of the data, the web page.

4.1 THE ONLINE DICTIONARY.

On all pages, the date of the current ACD version, used for citation, appears in the upper right corner, repeated again at the bottom center of the page. Centered at the top of each page is a menu that provides navigation among these sections:

Index to Sets Cognate Sets Finderlist
Sub-Groups Languages Words Proto-form indexes
References Roots Loans Noise Formosan

Below that is the title of the page, and then the various index lines, all of which are links to pages within this section of the ACD or to sections of the current page. As this appears to be a unique format for comparative dictionaries, the various sections will be briefly described. (This explanation will be facilitated by viewing the online pages at http://www.trussel2.com/ACD, while reading the descriptions below.)

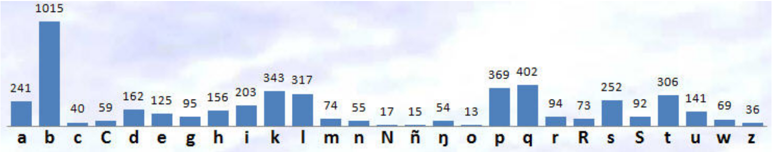
The two main methods of accessing *Cognate Sets*, the main dictionary, are the *Index to Sets* and the *Finderlist*.

4.1.1 Index to Sets. This is the normal opening page of the dictionary, the one you are directed to if you use the above link to the ACD. It is identical in arrangement to the main dictionary, but doesn't show the supporting forms, resulting in a compact display, much shorter, and more easily scanned visually. All reconstructions are links to their entries in the dictionary and, as in the main dictionary, the glosses are links to the English *Finderlist*.

4.1.2 Cognate Sets. This is the main dictionary, including all the supporting evidence for the reconstructions and, in many cases, extensive annotation. References cited in the glosses are to sources in addition to the primary dictionary reference shown in the language lists (below). At the top of each *Cognate Sets* page is a graphic representation of the number of sets by initial phoneme, as seen in figure 1. Because the ACD is continually being expanded and revised, this graph is updated regularly: as noted in 1.3, reconstructions beginning with *b- were most thoroughly examined from 1990–1995, and the *b- sets consequently still far outnumber all others.

The arrangement of the entries is alphabetical, except that a convention used by Dempwolff (1938) has been adopted, namely to ignore a preconsonantal nasal, as this often appears in some reflexes but not others (Blust 1996). Thus ordering sequences such as these appear ... *baban, *bambaŋen, *babaq$_1$ ..., or ... *kad(e)rit, *kandiŋ, *kandis, *kandoRa, *kahiR ..., this same alphabetical order naturally applying to *Index to Sets* as well. Words in the glosses of reconstructions are links to the *Finderlist,* highlighted when the mouse hovers over them.

FIGURE 1. GRAPHIC DISPLAY OF THE DISTRIBUTION OF RECONSTRUCTIONS BY INITIAL PHONEME



4.1.3 Finderlist. This is essentially a concordance of the glosses of the reconstructions, and serves as a semantic index to *Cognate Sets*. (Caution: the *Finderlist* is *not* a representative English vocabulary, and is not intended for use as a guide for Proto-Austronesians learning English.)

In addition to *Cognate Sets,* there are four additional dictionaries: *Roots*, *Loans*, *Noise*, and *Formosan*. Below are the introductions to each (written by Blust), as they appear in the ACD.

4.1.4 Roots. Because many reconstructed morphemes contain smaller submorphemic sound-meaning associations of the type that Brandstetter (1916) called "roots" (*Wurzeln*), I felt that a module collecting these elements into one place would be useful. The "Roots" module of the ACD thus amounts to a continuation of the data set presented in Blust (1988).

4.1.5 Loans. Loanwords are a perennial problem in historical linguistics. When they involve morphemes that are borrowed between related languages, they can provoke questions about the regularity of sound correspondences. When they involve morphemes that are borrowed between either related or unrelated languages, they can give rise to invalid reconstructions. Dempwolff (1938) included a number of known loanwords among his 2,216 "Proto-Austronesian" reconstructions to show that sound correspondences are often regular even with loanwords that are borrowed relatively early, and he marked these with a superscript x, as with *^{x}bad′u' 'shirt', which he knew to be a Persian loanword in many of the languages of western Indonesia, and (via Malay) in some languages of the Philippines. However, he overlooked a number of cases, such as *nanas 'pineapple' (an Amazonian cultigen that was introduced to insular Southeast Asia by the Portuguese). Since widely distributed loanwords can easily be confused with native forms, I have found it useful to include them in a separate module of the dictionary.

4.1.6 Noise. I have included a separate module of the dictionary called "Noise" (in the information-theoretic sense of meaningless data that can be confused with a true signal). The reason for this is that the search process that results in valid cognate sets inevitably turns up other material that is superficially appealing, but is questionable for various reasons. To simply dispose of this "information refuse" would be unwise for two reasons. First, further searching might show that some of these questionable comparisons are more strongly supported than it initially appeared. Second, even if the material is not upgraded through further comparative work, it is always possible that some future researcher with different standards of evaluation will stumble upon some of these comparisons and claim that they are valid, but were overlooked in the ACD. By including a module on "Noise," I can show that I have considered and rejected some possibilities that might be entertained by others.

4.1.7 Formosan. As originally conceived, the ACD excluded cognate sets that are confined to the Formosan languages. Since Taiwan is a relatively small island that has been inhabited by Austronesian-speaking peoples for at least 5,500 years, it was felt that early loanwords that have spread fairly widely among the languages might be difficult to distinguish from forms that were actually present in PAN. I have now reversed that decision on the grounds that it causes too many interesting and potentially valuable comparisons to be ignored. Cognate sets that are fairly widely distributed among Formosan languages belonging to different primary branches of Austronesian are now taken as evidence for PAN etyma. The only cases where I have maintained the original policy is where a cognate set is restricted to geographically contiguous languages and the probability of borrowing cannot easily be excluded. However, given the greater likelihood that undetected early loans might be attributed to PAN if they are attested only in Formosan languages, I have distinguished "Formosan-only" cognate sets as a separate module of the ACD.

The remaining sections of the ACD are *Sub-Groups*, *Languages*, *Words*, *Proto-Form Indexes*, and *References*, various ways to access *Cognate Sets*, and supplementary material.

4.1.8 Sub-Groups. This provides a tree diagram for the AN language family, with links to each protolanguage for which reconstruction is undertaken (all terminal and nonterminal nodes in the tree except "Formosan"). Clicking on these links calls up the subgroup in question, and provides a complete listing of all languages in that subgroup that appear in the dictionary. This is done in parallel columns, the first column listing the languages in alphabetical order, and the second in descending order of citation frequency, with number of citations appearing in parentheses in both lists. Clicking on a language moves to the corresponding language entry in *Languages*.

4.1.9 Languages. This large section indexes all the (approximately 70,000) words from each of the languages used for the reconstructions, including *Loans*, *Roots,* and *Noise*. This permits, for example, checking the forms of all the words used from that language as a group, and facilitates internal orthographic consistency. We are able to make language-wide orthographic changes when such changes are made in the newest dictionaries.

The headline of each language entry includes a variety of different kinds of information; here is an example:

39. Atayal (117) Form. (Egerod 1980) [tay] Taiwan (dialects: Matabalay 1, Cʔuliʔ 2, Mayrinax 9, Squliq 13)

The information in such an entry includes the total number of words in ACD from that language (listed below the entry); the *Language* subgroup (a link to that subgroup in the *Sub-Group* pages); the primary source dictionary (a link to the *References* entry for that source); the three-letter ISO code (a link to the *Ethnologue* page for that language, or one for which the language/dialect is listed as a sublanguage); an approximate geographic location; and the various dialects, if any, including the number of forms in the ACD from each (links to the dialect sublists below the main list, where all the words used from each dialect are shown).

The language indexes at the top of each page include numerous cross-references for the various ways the languages and dialects are referred to. Clicking on a word listed in these language lists brings up the section of the entry in *Cognate Sets* containing the word (just as clicking on a language name in *Cognate Sets* will bring up that language entry in *Languages*).

Achieving internal consistency with language and dialect names has been a challenge. We use a system of (usually) four-letter codes to represent the language and dialect names internally. The internal language table has an entry for the current representation of the code. Changing that entry results in across-the-board modification of the spelling, as, for example, in the case where political name changes are made.

Here, too, is an index to all the protolanguages and forms appearing in the reconstruction sets but reconstructed elsewhere, parallel in structure to the language listings of this section.

4.1.10 Words. This is an index of all the words in *Cognate Sets* and *Formosan*, a sort of 671-language dictionary, stretching from Aklanon *a* 'exclamatory of discovery; "ah" (with high intonation)' < PMP *a$_3$, to Vitu *zuzu* 'breast, milk; suckle' < PAN *susu. Here, too, all starred forms are links to the corresponding entry in *Cognate Sets*.

4.1.11 Proto-form indexes. This is similar to the *Index to Sets* in providing a comprehensive alphabetized list of reconstructions without supporting evidence. However, it is arranged by protolanguage, starting at the top of the tree (PAN) and working down to Proto-Oceanic, the lowest protolanguage for which reconstruction is undertaken in the ACD.

4.1.12 References. This provides a list of all references throughout the dictionary, whether in Notes or glosses, and linked to the reference list. The *References* page provides menu access to two additional sections, *Stats* and *Update Log*. Although intended for internal project management, these provide a picture of the growth and development of the data, and a way to check for updated entries. New material is regularly added to existing entries.

4.2 THE DATA. Although at first glance the organization of a large and diverse data set with reconstructions for nine hierarchically related protolanguages—including forms and languages, loans and roots, doublets and disjuncts, references, and so on—appears daunting, modern programming and data processing techniques have made it possible to produce a coherent, multiconnected body of interrelated elements. However, as described in section 2, this project began with extensive legacy data, so there was a fairly lengthy initial process of conversion involving the two different text-file versions of the data. Extensive formatting codes persisted in the text material, diverse language coding was present, and there was the expected quantity of input errors pervading the legacy data. The original online representation is still viewable at http://www.ohio.edu/people/mcginn/Blust%27s_ACD_dict.txt.

The initial goal was to create a relational database structure that would allow modeling the data to match, as nearly as possible, the hierarchical and feature-based structure of Blust's presentation of the linguistic data, down to the smallest detail, and to import the legacy data into that structure.

The test of the match-up between the actual structure of the reconstructions envisioned by Blust and the computational structure as developed by Trussel is the output, the online pages. To the extent that the data could be modeled programmatically into pages mirroring Blust's vision, the structure was successful. This was also a lengthy process, involving data structure changes to accommodate previously unrecognized distinctions in the output and the like, a process that in some areas of the ACD continues today.

A first goal, then, was to devise a system of programs that would minimally generate pages from data which were more or less identical to the handmade text pages produced by Blust.

4.2.1 Sets. As described above, beyond the main form, reconstructions often include affixed forms, reduplications, and compounds. All of these are grouped together as members of a "Set," which takes its name and gloss from the highest reconstruction in the set. The entire set is displayed within a single box, but the Set name is not assigned a protolanguage; rather, it is an envelope for the group of related protolanguages of the reconstruction set (which may often be only one).

Within the database structure, Sets is a separate table, hierarchically superior to the table containing the reconstructions. As these tables only "look upward," Sets isn't "aware" of its members, only containing collections of forms, sorting material, notes, glosses, semantic codes, and data processing metadata.

4.2.2 Proto-Languages. Within each Set are one or more reconstructions, and these are all hierarchically arranged by protolanguages, of which there are nine (see above).

This is the second level of the data hierarchy, storing all explicitly reconstructed forms and their links to the Set that they are a part of. While these protoforms don't know who their witnesses are, they do know about related doublet and disjunct forms, they may contain their own annotations, and they all know about sort orders, and whether they are to be internally grouped with other reconstructions in the set, or separated into independent subsets.

4.2.3 Forms. At the bottom of the structural hierarchy is observational reality, the actual *forms* and their glosses, which are the evidence for the reconstructions. These forms naturally know what language and/or dialect they are part of, and who their *protolanguages* are, but not directly what *set* they are in. They are happy to go wherever their protolanguages go. And they know whether they were formed by metathesis or assimilation, and what order they should appear in relative to each other, since all the form lists are geographically arranged.

4.2.4 Languages. In addition to the starring hierarchy, there are numerous supporting roles connected to it. The Forms have their *languages* and dialects, which know which *protolanguages* they are descended from, what their various alternate spellings are, and what the primary lexical sources for the data are.

Roots, *Loans*, and *Noise* were prepared for inclusion in the 1995 version, but were not published at that time. They appear online in the current version. Formosan-only forms were added in the current version.

4.3 DISPLAY AND MAINTENANCE.

Just as the ACD is innovative in its many areas of description that are found in hardly any other comparative dictionary, the display and maintenance functions developed for the ACD are similarly innovative, and thus worthy of some detailed description here.

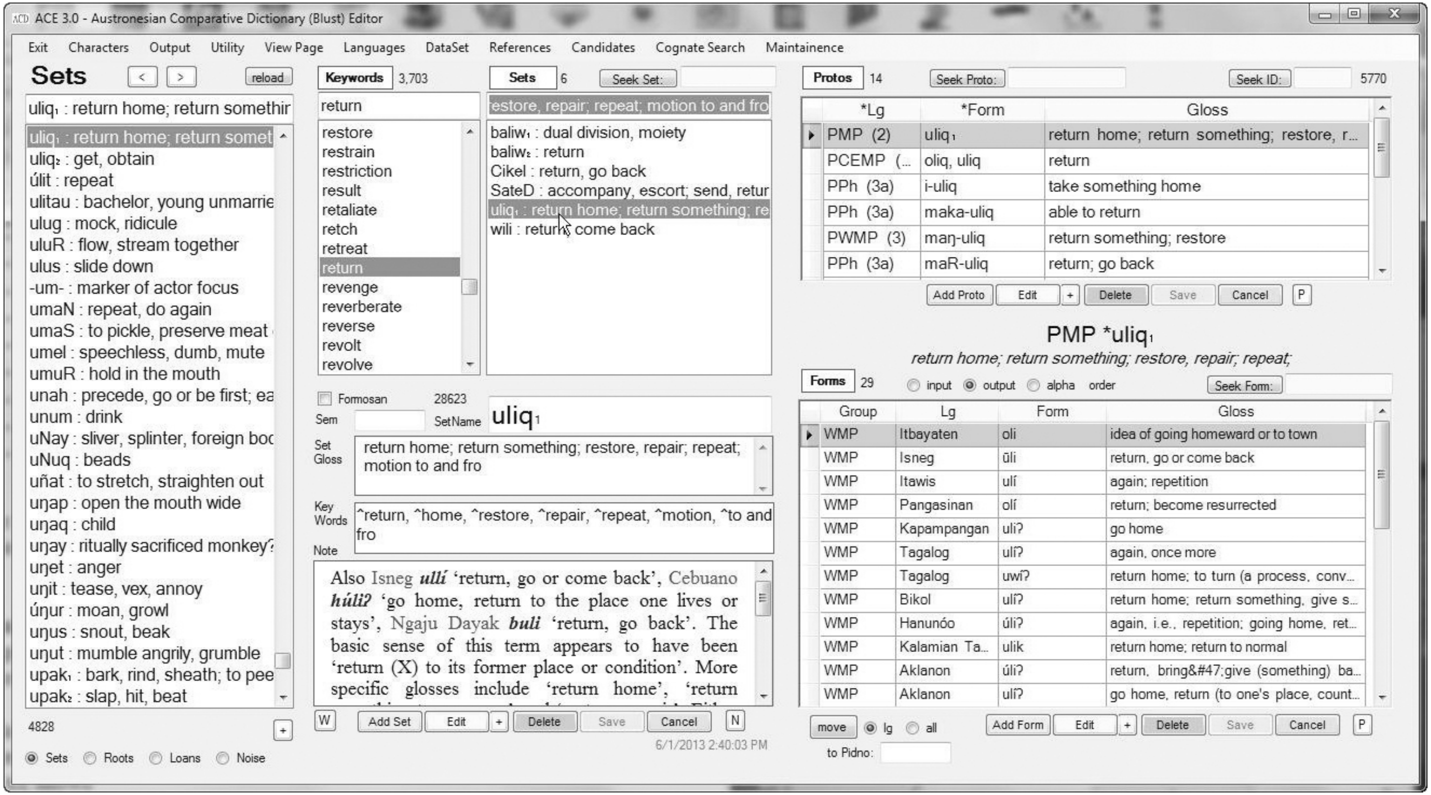
4.3.1 The compiler's interface. Creating a structure to handle the data, features, and associated meta-data was to some extent a trial-and-error process that evolved as new material was added and representations of the data proved inadequate in various areas. Challenges of inputting and editing accompanied the maintenance and development of the ACD.

A critical element was the development of a user interface, in this case, a compiler's interface, that would facilitate extensive addition and revision, including the easy reassignment of hierarchical levels and methods of coding the entries to achieve the sophisticated internal ordering required. Figure 2 shows a main editor screen for a complex entry, *uliq$_1$ 'return home ...', comprising 14 explicit reconstructions, and over 100 forms.

The screen in figure 2 can be considered a combination of five screens, as shown in figure 3. Two of them are dedicated to searching the data, the Index of all the Sets on the left; and the upper left quadrant, Keys, for keyword searching (in this example showing a search for 'return' and the selection of *uliq$_1$). The remaining three quadrants are the main data editing and input screens for Sets (including notes), Protoforms (explicit reconstructions), and Forms (words). Each of the three data quadrants includes a link (+) to an expanded screen for more detailed input for that table.

The expanded Sets form, which appears as figure 2, includes various editing tools for formatting the text. It can display the note text in preview mode, as in figure 4, and in edit mode, as in figure 5.

FIGURE 2. MAIN EDITOR SCREEN FOR A SAMPLE ENTRY


ACE 3.0 - Austronesian Comparative Dictionary (Blust) Editor
Exit Characters Output Utility View Page Languages DataSet References Candidates Cognate Search Maintainence
Sets < > reload
Keywords 3,703
Sets 6 Seek Set:
Protos 14 Seek Proto: Seek ID: 5770
Add Proto Edit + Delete Save Cancel P
Forms 29 ○ input ● output ○ alpha order Seek Form:
W Add Set Edit + Delete Save Cancel N
move ● lg ○ all Add Form Edit + Delete Save Cancel P
to Pidno:
4828
● Sets ○ Roots ○ Loans ○ Noise

uliq₁ : return home; return somethir

uliq₁ : return home; return somet
uliq₂ : get, obtain
úlit : repeat
ulitau : bachelor, young unmarrie
ulug : mock, ridicule
uluR : flow, stream together
ulus : slide down
-um- : marker of actor focus
umaN : repeat, do again
umaS : to pickle, preserve meat
umel : speechless, dumb, mute
umuR : hold in the mouth
unah : precede, go or be first; ea
unum : drink
uNay : sliver, splinter, foreign boc
uNuq : beads
uñat : to stretch, straighten out
uŋap : open the mouth wide
uŋaq : child
uŋay : ritually sacrificed monkey?
uŋet : anger
uŋit : tease, vex, annoy
úŋur : moan, growl
uŋus : snout, beak
uŋut : mumble angrily, grumble
upak₁ : bark, rind, sheath; to pee
upak₂ : slap, hit, beat

return

restore
restrain
restriction
result
retaliate
retch
retreat
return
revenge
reverberate
reverse
revolt
revolve

estore, repair; repeat; motion to and fro

baliw₁ : dual division, moiety
baliw₂ : return
Cikel : return, go back
SateD : accompany, escort; send, retur
uliq₁ : return home; return something; re
wili : return, come back

☐ Formosan 28623

Sem

SetName uliq₁

Set Gloss: return home; return something; restore, repair; repeat; motion to and fro

Key Words: ^return, ^home, ^restore, ^repair, ^repeat, ^motion, ^to and fro

Note

Also Isneg ***ullí*** 'return, go or come back', Cebuano ***húliʔ*** 'go home, return to the place one lives or stays', Ngaju Dayak ***buli*** 'return, go back'. The basic sense of this term appears to have been 'return (X) to its former place or condition'. More specific glosses include 'return home', 'return

6/1/2013 2:40:03 PM

*Lg	*Form	Gloss
PMP (2)	uliq₁	return home; return something; restore, r...
PCEMP (...	oliq, uliq	return
PPh (3a)	i-uliq	take something home
PPh (3a)	maka-uliq	able to return
PWMP (3)	maŋ-uliq	return something; restore
PPh (3a)	maR-uliq	return; go back

PMP *uliq₁

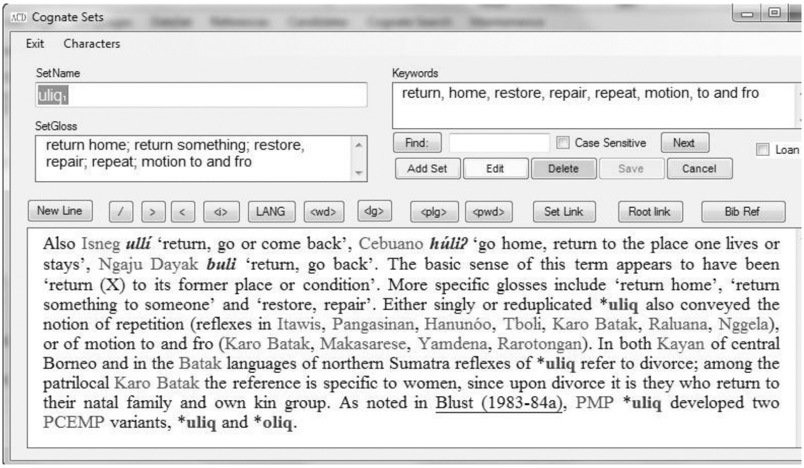
return home; return something; restore, repair; repeat;

Group	Lg	Form	Gloss
WMP	Itbayaten	oli	idea of going homeward or to town
WMP	Isneg	ūli	return, go or come back
WMP	Itawis	ulí	again; repetition
WMP	Pangasinan	olí	return; become resurrected
WMP	Kapampangan	uliʔ	go home
WMP	Tagalog	uliʔ	again, once more
WMP	Tagalog	uwiʔ	return home; to turn (a process, conv...
WMP	Bikol	uliʔ	return home; return something, give s...
WMP	Hanunóo	úliʔ	again, i.e., repetition; going home, ret...
WMP	Kalamian Ta...	ulik	return home; return to normal
WMP	Aklanon	úliʔ	return, bring&#47;give (something) ba...
WMP	Aklanon	uliʔ	go home, return (to one's place, count...

FIGURE 3. UNDERLYING NATURE OF A MAIN EDITOR SCREEN

Index	*Keys*	Protos
	Sets	Forms

FIGURE 4. PREVIEW MODE



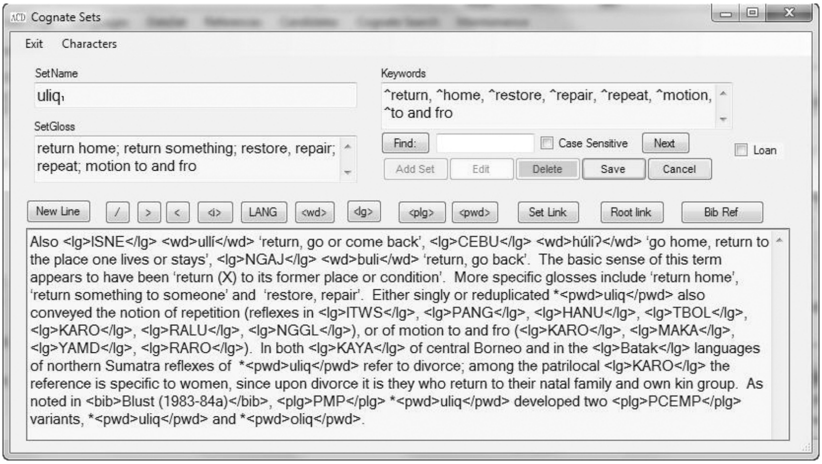
SetName: uliq$_1$

Keywords: return, home, restore, repair, repeat, motion, to and fro

SetGloss: return home; return something; restore, repair; repeat; motion to and fro

Also Isneg ***ullí*** ‘return, go or come back’, Cebuano ***húliʔ*** ‘go home, return to the place one lives or stays’, Ngaju Dayak ***buli*** ‘return, go back’. The basic sense of this term appears to have been ‘return (X) to its former place or condition’. More specific glosses include ‘return home’, ‘return something to someone’ and ‘restore, repair’. Either singly or reduplicated ***uliq** also conveyed the notion of repetition (reflexes in Itawis, Pangasinan, Hanunóo, Tboli, Karo Batak, Raluana, Nggela), or of motion to and fro (Karo Batak, Makasarese, Yamdena, Rarotongan). In both Kayan of central Borneo and in the Batak languages of northern Sumatra reflexes of ***uliq** refer to divorce; among the patrilocal Karo Batak the reference is specific to women, since upon divorce it is they who return to their natal family and own kin group. As noted in Blust (1983-84a), PMP ***uliq** developed two PCEMP variants, ***uliq** and ***oliq**.

FIGURE 5. EDIT MODE



SetName: uliq$_1$

Keywords: ^return, ^home, ^restore, ^repair, ^repeat, ^motion, ^to and fro

SetGloss: return home; return something; restore, repair; repeat; motion to and fro

Also <lg>ISNE</lg> <wd>ullí</wd> ‘return, go or come back’, <lg>CEBU</lg> <wd>húliʔ</wd> ‘go home, return to the place one lives or stays’, <lg>NGAJ</lg> <wd>buli</wd> ‘return, go back’. The basic sense of this term appears to have been ‘return (X) to its former place or condition’. More specific glosses include ‘return home’, ‘return something to someone’ and ‘restore, repair’. Either singly or reduplicated *<pwd>uliq</pwd> also conveyed the notion of repetition (reflexes in <lg>ITWS</lg>, <lg>PANG</lg>, <lg>HANU</lg>, <lg>TBOL</lg>, <lg>KARO</lg>, <lg>RALU</lg>, <lg>NGGL</lg>), or of motion to and fro (<lg>KARO</lg>, <lg>MAKA</lg>, <lg>YAMD</lg>, <lg>RARO</lg>). In both <lg>KAYA</lg> of central Borneo and in the <lg>Batak</lg> languages of northern Sumatra reflexes of *<pwd>uliq</pwd> refer to divorce; among the patrilocal <lg>KARO</lg> the reference is specific to women, since upon divorce it is they who return to their natal family and own kin group. As noted in <bib>Blust (1983-84a)</bib>, <plg>PMP</plg> *<pwd>uliq</pwd> developed two <plg>PCEMP</plg> variants, *<pwd>uliq</pwd> and *<pwd>oliq</pwd>.

Other screens and programs are accessible through the top menu of the main editing screen, which itself includes numerous additional display and search functions for each of the data sections.

5. THE ACD AND OTHER COMPARATIVE DICTIONARIES. It is unclear how many of the world's language families are represented by comparative dictionaries. However, of those that are known to us, it is possible to make some preliminary comparisons of the characteristics and scope of the work that has been done. Those that are known to us include the following:

Algonquian: *A computer-generated dictionary of Proto-Algonquian* (Hewson 1993). Contains 4,066 lexical reconstructions based on four Algonquian languages, but this figure reportedly is inflated by many morphologically related forms (Ives Goddard, pers. comm., March 17, 2004).

Bantu: *CBOLD, Comparative Bantu online dictionary*. Started by Larry Hyman and John Lowe at the University of California at Berkeley in 1994, but defunct since 1999. At that time, this dictionary contained about 445,000 lexical items from 200 languages and 70 sources. More recently, comparative Bantu lexicology has been carried forward by *Bantu lexical reconstructions* (http://www.africamuseum.be/collections/browsecollections/humansciences/blr), run by a team of rotating editors. According to the most current information available on the Internet, "BLR 3 is a database with ca. 10,000 entries that have been proposed as Proto-Bantu reconstructions. BLR 3 is meant to be a working tool for Bantuists and other linguists. BLR 3 is not a finished product, it is continuously being updated by its present editors Yvonne Bastin and Thilo C. Schadeberg."

Dravidian: *A Dravidian etymological dictionary* (Burrow and Emeneau 1984). Contains 5,569 cognate sets, but no reconstructions, reportedly because of uncertainty about the reconstruction of vowels.

Indo-European: Although most Indo-Europeanists consider it badly dated, Pokorny (1959) remains the most substantial single collection of reconstructions with supporting evidence available for Indo-European languages. This comparative dictionary contains around 2,215 base entries, but many of its boldface entries reportedly have "no real existence (they involve old borrowings, or other very shadowy material based on one or two branches only, etc.), and conversely, many of the larger entries would now actually be split up into several different roots" (Brent Vine, pers. comm., January 4, 2005). More recent works propose a larger number of Proto–Indo-European base forms, as Mallory and Adams (1997), which contains about 4,200 entries, but without full supporting evidence.

Mayan: *A preliminary Mayan etymological dictionary* (Kaufman 2003). A freely accessible online resource that is 1,535 pages in length, with around 2,000 reconstructed base forms. Only bases are given, and there is minimal to no annotation.

Mon-Khmer: *A Mon-Khmer comparative dictionary* (Shorto 2006). A posthumous work that contains 2,246 etymologies, and over 17,000 lexical citations drawn from about 130 languages. This study includes affixed forms and is annotated. With this published dictionary as its core, comparative work on the lexicons of Mon-Khmer languages is now continuing in the online Mon-Khmer Languages Project (MKLP), which has incorporated over 250,000 lexical citations from around 170 languages. At present, much of this content consists of raw data, since the MKLP is conceived as "an on-line workspace that includes (among many other lexical resources) a traditional comparative dictionary, but as a whole is still a work in progress" (Doug Cooper, pers. comm., June 5, 2013).

Sino-Tibetan: *The Sino-Tibetan etymological dictionary and thesaurus* (STEDT) is a long-running comparative project, under the direction of James A. Matisoff at the University of California at Berkeley, that has been funded since 1987. The most substantial product of this project to date is Matisoff (2003), a 750-page state of the art summary. Currently, the set of protoforms contains 4,380 records on seven hierarchical levels, of which the two highest are Proto–Sino-Tibetan (26 etyma), and Proto–Tibeto-Burman (2,064 etyma). The dictionary includes some reconstructions that are supported by reflexes in over 200 languages, and it is extensively annotated (Daniel W. Bruhn, pers. comm., June 3, 2013).

Turkic: A Proto-Turkic dictionary that forms part of the Russian Etymological Project appeared in seven volumes published between 1974 and 2003 (Sevortyan 1974, 1978, 1980; Levitskaya 1989, 1997; Blagova 2000; Dybo 2003). It reportedly contains about 2,540 entries (Stefan Georg, pers. comm., June 18, 2013).

Uralic: Collinder (1955) proposed about 1,025 cognate sets for Proto–Finno-Ugric, but without reconstructions. Janhunen (1981) has reconstructed some 140 Proto-Uralic forms through a comparison of Proto-Samoyedic and Proto-Finnic, and has stated the sound correspondences linking the supporting cognate sets in the form of 58 rules for vocalism and 12 for consonantism (John Kupchik, pers. comm., September 4, 2004).

Uto-Aztecan: Stubbs (2011) is an annotated Uto-Aztecan comparative dictionary that contains 2,703 numbered cognate sets with reconstructions, as well as many subsets marked by alphabetic subscript (7a, 7b, etc.). Subscripted items are not morphologically related forms, but problematic words that need special discussion. The author of this work is motivated in part by religious convictions that may cause some scholars to have reservations about his conclusions, although the book has been favorably reviewed in an academically appropriate venue (Hill 2012).

6. CONCLUSION.

In conclusion, the ACD is like the proverbial glass of water that contains water up to the mid-point: from one point of view it is half-full, but from another it is half-empty. Many fundamental words have not yet been entered, and it is a source of continuing frustration to see how long it takes to get to these, given the standards of documentation that have been put in place in comparisons such as *aNak ‘child, offspring’, *aCay ‘death’, or *kaen ‘to eat’. On the other hand, the *b section alone contains more than 1,000 base forms, and is over 500 single-spaced pages in print form—far larger than the corresponding section in the dictionary of any attested AN language. In the end, finding a way to complete the ACD within the lifetimes of the authors of this report may require a compromise between depth and breadth, between thoroughness and an abhorrence of gaps. Most importantly, in the new age of online databases, projects like this enter the public domain, and with archiving and the needed permissions, the project itself need not be limited to a single person or a human lifetime.

APPENDIX. LANGUAGES CITED IN DESCENDING ORDER OF CITATION FREQUENCY

1. Malay	(2,578)
2. Cebuano	(1,811)
3. Tagalog	(1,598)
4. Ilokano	(1,417)
5. Maranao	(1,351)
6. Bikol	(1,348)
7. Old Javanese	(1,292)
8. Javanese	(1,208)
9. Toba Batak	(1,128)
10. Iban	(1,103)
11. Balinese	(1,086)
12. Aklanon	(994)
13. Sundanese	(856)
14. Manggarai	(846)
15. Sasak	(844)
16. Karo Batak	(817)
17. Ifugaw	(812)
18. Paiwan	(806)
19. Makasarese	(775)
20. Bare’e	(772)
21. Tae’	(754)
22. Isneg	(732)
23. Manobo	(727)
24. Bolaang Mongondow	(719)
25. Hanunóo	(719)
26. Kankanaey	(699)
27. Ngaju Dayak	(684)
28. Bontok	(665)
29. Itbayaten	(640)
30. Kayan	(640)
31. Sangir	(630)
32. Malagasy	(618)
33. Arosi	(594)
34. Fijian	(586)
35. Kelabit	(563)
36. Casiguran Dumagat	(535)
37. Samoan	(530)
38. Kapampangan	(522)
39. Tongan	(497)
40. Nggela	(489)
41. Pangasinan	(487)
42. Tiruray	(482)
43. Hiligaynon	(478)
44. Puyuma	(436)
45. Amis	(430)
46. Tontemboan	(421)
47. Kadazan	(417)
48. Sa‘a	(412)
49. Thao	(401)
50. Ngadha	(400)
51. Rotinese	(385)
52. Kambera	(381)
53. Kavalan	(373)
54. Wolio	(370)
55. Rennellese	(368)
56. Tetun	(361)
57. Maori	(355)
58. Lau	(352)
59. Dairi-Pakpak Batak	(351)
60. Mansaka	(351)
61. Buruese	(350)
62. Simalur	(341)
63. Hawaiian	(336)
64. Kenyah	(336)
65. Nias	(328)
66. Niue	(326)
67. Buginese	(312)
68. Motu	(306)
69. ‘Āre‘āre	(297)
70. Acehnese	(291)
71. Yamdena	(291)
72. Itawis	(286)
73. Palauan	(283)
74. Sika	(282)
75. Chamorro	(279)
76. Bimanese	(276)
77. Rembong	(276)
78. Kwaio	(275)
79. Mandar	(274)
80. Pazeh	(267)
81. Binukid	(266)
82. Melanau (Mukah)	(261)
83. Mota	(239)
84. Tboli	(235)
85. Tuvaluan	(229)
86. Tausug	(228)
87. Banggai	(225)
88. Tolai	(224)
89. Bunun	(207)
90. Uma	(207)
91. Bintulu	(206)
92. Buli	(206)
93. Gorontalo	(206)
94. Nukuoro	(205)
95. Hawu	(196)
96. Saisiyat	(196)
97. Gilbertese	(195)
98. Rejang	(194)
99. Masbatenyo	(187)
100. Rarotongan	(181)
101. Muna	(179)
102. Nakanai	(174)
103. Leti	(171)
104. Roviana	(171)
105. Asilulu	(169)
106. Bugotu	(169)
107. Mapun	(164)
108. Bahasa Indonesia	(160)
109. Fordata	(160)
110. Soboyo	(159)
111. Rotuman	(158)
112. Selaru	(158)
113. Anuta	(157)
114. Gedaged	(154)
115. Ibaloy	(151)
116. Berawan	(147)
117. Kapingamarangi	(140)
118. Palawan Batak	(139)
119. Chuukese	(138)
120. Kalamian Tagbanwa	(138)
121. Lun Dayeh*	(137)
122. Manam	(135)
123. Woleaian	(132)
124. Komodo	(131)
125. Pohnpeian	(131)
126. Puluwat	(131)
127. Gitua	(129)
128. Yakan	(129)
129. Numfor-Biak	(125)
130. Wuvulu	(122)
131. Kei	(120)
132. Madurese	(120)
133. Atayal	(119)
134. Wetan	(117)
135. Erai	(116)
136. Wayan	(115)
137. Yami	(114)
138. Tsou	(113)
139. Kanakanabu	(111)
140. Rukai	(109)
141. Lampung	(108)
142. Loniu	(105)
143. Saaroa	(105)
144. Mussau	(103)
145. Seimat	(103)
146. Eddystone/Mandegusu	(101)
147. Paulohi	(99)
148. Mentawai	(98)
149. Minangkabau	(97)
150. Banjarese	(96)

151. Lou	(96)
152. Seediq	(96)
153. Molima	(95)
154. Waray-Waray	(95)
155. Singhi Land Dayak	(89)
156. Marshallese	(88)
157. Maloh	(85)
158. Murik	(83)
159. Kamarian	(82)
160. Lamaholot	(81)
161. Kiput	(80)
162. Futunan	(79)
163. Subanen/Subanun	(79)
164. Mono-Alu	(78)
165. Ida'an Begak	(77)
166. Li'o	(77)
167. Jarai	(74)
168. Nali	(74)
169. Nauna	(74)
170. Dupaningan Agta	(73)
171. Moken	(73)
172. Sambal	(73)
173. Gayo	(72)
174. Dobel	(65)
175. Kosraean	(63)
176. Kédang	(62)
177. Kowiai	(62)
178. Agutaynen	(61)
179. Miri	(61)
180. Mokilese	(61)
181. Rhade	(61)
182. Toqabaqita	(61)
183. Tombonuwo	(60)
184. Alune	(59)
185. Bipi	(59)
186. Bidayuh†	(57)
187. Cheke Holo	(56)
188. Label	(55)
189. Totoli	(55)
190. Leipon	(54)
191. Ma'anyan	(54)
192. Tanga	(54)
193. Lonwolwol	(52)
194. Atoni	(51)
195. Sori	(51)
196. Mbula	(49)
197. Dobuan	(47)
198. Gaddang	(47)
199. Penchal	(47)
200. Samal	(47)
201. Siraya	(47)
202. Likum	(46)
203. Narum	(46)
204. Tigak	(456
205. Waropen	(46)
206. Kisar	(45)
207. Titan	(45)
208. Wogeo	(45)
209. Aua	(43)
210. Dampelas	(43)
211. Limbang Bisaya	(43)
212. Vitu	(41)
213. Kilivila	(40)
214. Lindrou	(40)
215. Watubela	(40)
216. Favorlang	(39)
217. Raga	(38)
218. Nakanamanga	(37)
219. Bilaan	(36)
220. Bwaidoga/Bwaidoka	(36)
221. Timugon Murut	(36)
222. Ulawa	(36)
223. Boano	(35)
224. Agta	(34)
225. Tubetube	(34)
226. Bauro	(32)
227. Simalungun Batak	(31)
228. Ere	(30)
229. Lusi	(30)
230. Tawala	(30)
231. Aborlan Tagbanwa	(29)
232. Numbami	(29)
233. Selau	(29)
234. Yogad	(29)
235. Araki	(28)
236. Tunjung	(28)
237. Ansus	(27)
238. Serui-Laut	(27)
239. Atta	(26)
240. Ibanag	(26)
241. Kairiru	(26)
242. Pak	(26)
243. Rungus Dusun	(26)
244. Seru	(25)
245. Tondano	(25)
246. Bisaya Bukid	(24)
247. Takia	(24)
248. Ivatan	(23)
249. Lauje	(23)
250. Mamanwa	(23)
251. Kadazan Dusun	(22)
252. Makatea	(22)
253. Nehan	(22)
254. Gapapaiwa	(21)
255. Mendak	(21)
256. Moa	(21)
257. Namakir	(21)
258. Romblomanon	(21)
259. Sekar	(21)
260. Duke of York	(20)
261. Isinay	(20)
262. Kayeli	(20)
263. Lenkau	(20)
264. Palu'e	(20)
265. Pendau	(20)
266. Taboyan	(20)
267. Taokas	(20)
268. Yapese	(20)
269. Cape Cumberland	(19)
270. Lawangan	(19)
271. Northeast Ambae	(19)
272. Tialo	(19)
273. Ilongot	(18)
274. Mailu	(18)
275. Southeast Ambrym	(18)
276. Balaesang	(17)
277. Central Maewo	(17)
278. Mori	(17)
279. Paamese	(17)
280. Siang	(17)
281. Windesi	(17)
282. Anejom	(16)
283. Ata	(16)
284. Basai	(16)
285. Elat	(16)
286. Samihim	(15)
287. Bonggi	(14)
288. Hitu	(14)
289. Levei	(14)
290. Melanau (Matu)	(14)
291. Moor	(14)
292. Sonsorol-Tobi	(14)
293. Geser	(13)
294. Kapuas	(13)
295. Sa'ban	(13)
296. Ujir	(13)
297. Bobot	(12)
298. Kalagan	(12)
299. Penan	(12)
300. Balangao	(11)
301. Hoanya	(11)
302. Kuruti	(11)
303. Talise	(11)
304. Ahus	(10)
305. Ba'amang	(10)
306. Bonfia	(10)
307. Kalinga	(10)
308. Kallahan	(10)
309. Kemak	(10)

310. Melanau (Dalat)	(10)
311. Tangoa	(10)
312. Tring	(10)
313. Varisi	(10)
314. Dohoi	(9)
315. Dondo	(9)
316. Dusun Deyah	(9)
317. Kanowit	(9)
318. Lakakai	(9)
319. Mambai	(9)
320. Masiwang	(9)
321. Melanau (Balingian)	(9)
322. Merlav	(9)
323. Roro	(9)
324. Sebop (Long Luyang)	(9)
325. W. Tarangan	(9)
326. Angkola[‡]	(8)
327. Bolinao	(8)
328. Bonkovia	(8)
329. Bukat	(8)
330. Dusun Malang	(8)
331. Ende	(8)
332. Enggano	(8)
333. Itneg	(8)
334. Kokota	(8)
335. Lahanan	(8)
336. Longgu	(8)
337. Lungga	(8)
338. Misima	(8)
339. Sumbawanese	(8)
340. Tarakan	(8)
341. Tolaki	(8)
342. Drehet	(7)
343. Dusner	(7)
344. Ghari	(7)
345. Iaai	(7)
346. Ibatan	(7)
347. Irarutu	(7)
348. Kejaman	(7)
349. Murut	(7)
350. Sichule	(7)
351. Suau	(7)
352. Baluan	(6)
353. Banoni	(6)
354. Buma	(6)
355. Carolinian	(6)
356. Kaidipang	(6)
357. Kakiduge:n Ilongot	(6)
358. Katingan	(6)
359. Marovo	(6)
360. Mekeo	(6)
361. Melanau[**]	(6)
362. Ngwatua	(6)
363. Papora	(6)
364. Sangil	(6)
365. Sarikei	(6)
366. Taje	(6)
367. Tolo	(6)
368. Trobiawan	(6)
369. Ubir	(6)
370. Wandamen	(6)
371. Arguni	(5)
372. Babuza	(5)
373. Belait	(5)
374. Bukidnon	(5)
375. Central Sama	(5)
376. Central Tagbanwa	(5)
377. Dhao/Ndao	(5)
378. Kulawi	(5)
379. Rade	(5)
380. Riung	(5)
381. Santa Ana	(5)
382. Sawai	(5)
383. Sobei	(5)
384. Tabar	(5)
385. Takuu	(5)
386. Tombulu	(5)
387. Vowa	(5)
388. Alas	(4)
389. Alumbis Murut	(4)
390. Ambai	(4)
391. Ampana	(4)
392. Axamb	(4)
393. Bantik	(4)
394. Batu Merah	(4)
395. Bekatan	(4)
396. Bolongan	(4)
397. Cham	(4)
398. Giman	(4)
399. Helong	(4)
400. Hiw	(4)
401. Idate	(4)
402. Kodi	(4)
403. Kurudu	(4)
404. Lenakel	(4)
405. Lundu	(4)
406. Mafea	(4)
407. Mondropolon	(4)
408. Motlav	(4)
409. Murung	(4)
410. Nuaulu	(4)
411. Padoe	(4)
412. Rapanui	(4)
413. Roma	(4)
414. Salako	(4)
415. Tagbanwa	(4)
416. Tambotalo	(4)
417. Teluti	(4)
418. Tonga	(4)
419. Tonsea	(4)
420. Aore	(3)
421. Ayta Maganchi	(3)
422. Baetora	(3)
423. Bagobo	(3)
424. Buludupi	(3)
425. Central Santo	(3)
426. Emira	(3)
427. Fauro	(3)
428. Futuna-Aniwa	(3)
429. Gabadi	(3)
430. Kaniet	(3)
431. Kayupulau	(3)
432. Kele	(3)
433. Kwaraʻae	(3)
434. Lele	(3)
435. Litzlitz	(3)
436. Maleu	(3)
437. Mele-Fila	(3)
438. Mengen	(3)
439. Minansut	(3)
440. Mori Atas	(3)
441. Mori Bawah	(3)
442. Mosina	(3)
443. Nauruan	(3)
444. North Malo	(3)
445. Onin	(3)
446. Paitan	(3)
447. Panayati	(3)
448. Papitalai	(3)
449. Popalia	(3)
450. Punan Kelai	(3)
451. Ratahan	(3)
452. Sentah Land Dayak	(3)
453. Sowa	(3)
454. Tahitian	(3)
455. Tandai	(3)
456. Teop	(3)
457. Tifu	(3)
458. Tolomako	(3)
459. Tonsawang	(3)
460. Tutuba	(3)
461. Uripiv	(3)
462. Vaghua	(3)
463. Vangunu	(3)
464. V’ënen Taut	(3)
465. Vovo	(3)
466. Waiyewa	(3)
467. Amblau	(2)
468. Ampibabo-Lauje	(2)

469. Apma	(2)
470. Babatana	(2)
471. Bali (Uneapa)	(2)
472. Beta	(2)
473. Biga	(2)
474. Bonerate	(2)
475. Delang	(2)
476. Dubea	(2)
477. Dusun Witu	(2)
478. Eastern Kadazan	(2)
479. Fortsenal	(2)
480. Galoli	(2)
481. Halia	(2)
482. Huaulu	(2)
483. Jawe	(2)
484. Kahua	(2)
485. Keo	(2)
486. Kis	(2)
487. Kove	(2)
488. Kulisusu	(2)
489. Laha	(2)
490. Lamogai	(2)
491. Letemboi	(2)
492. Magori	(2)
493. Mamben	(2)
494. Marino	(2)
495. Melanau††	(2)
496. Modang (Long Glat)	(2)
497. Mortlockese	(2)
498. Munggui	(2)
499. Murnaten	(2)
500. Murua	(2)
501. Nalik	(2)
502. Nasarian	(2)
503. Nengone	(2)
504. Nggeri	(2)
505. Paku	(2)
506. Pelipowai	(2)
507. Poro	(2)
508. Ranon	(2)
509. Rerep	(2)
510. Sengseng	(2)
511. South Efate	(2)
512. Southwest Tanna	(2)
513. Sungai Seguliud	(2)
514. Tajio	(2)
515. Tambunan Dusun	(2)
516. Tamuan	(2)
517. Tanema	(2)
518. Tanjong	(2)
519. Tapuh	(2)
520. Tarpia	(2)
521. Tidong	(2)
522. Tikopia	(2)
523. Tinputz	(2)
524. Tokelauan	(2)
525. Tsat	(2)
526. Tuamotuan	(2)
527. Umiray Dumaget	(2)
528. Ureparapara	(2)
529. Vartavo	(2)
530. Vinmavis	(2)
531. Visina	(2)
532. Wedau	(2)
533. Wusi-Mana	(2)
534. Yabem	(2)
535. Yotefa	(2)
536. Zabana	(2)
537. Abaknon	(1)
538. Adonara	(1)
539. Ali	(1)
540. Aroma	(1)
541. Arta	(1)
542. As	(1)
543. Atchin	(1)
544. Babuyan	(1)
545. Bada	(1)
546. Balantak	(1)
547. Bilibil	(1)
548. Bolango	(1)
549. Bonga	(1)
550. Bonggo	(1)
551. Bugis (Soppeng)	(1)
552. Bungku	(1)
553. Canala	(1)
554. Central Palawan	(1)
555. Dali’	(1)
556. Damar	(1)
557. Dehu	(1)
558. Donggo	(1)
559. Dumagat (Polillo)	(1)
560. Fagudu	(1)
561. Filakara	(1)
562. Gah	(1)
563. Ghove	(1)
564. Gondang	(1)
565. Hoava	(1)
566. Hotti/Hoti	(1)
567. Inati	(1)
568. Isiai	(1)
569. Itawit	(1)
570. Jotafa	(1)
571. Kaitetu	(1)
572. Kaiwa	(1)
573. Kantu’	(1)
574. Keherara	(1)
575. Kembayan	(1)
576. Ketagalan	(1)
577. Kilokaka	(1)
578. Kola	(1)
579. Kwamera	(1)
580. Lamboya	(1)
581. Lara’ Land Dayak	(1)
582. Larevat	(1)
583. Larike	(1)
584. Laura	(1)
585. Lelepa	(1)
586. Lengo	(1)
587. Lingarak	(1)
588. Luang-Sermata	(1)
589. Luilang	(1)
590. Malang	(1)
591. Mamboru	(1)
592. Mandaya	(1)
593. Manusela	(1)
594. Mapos	(1)
595. Mapremo	(1)
596. Maragus	(1)
597. Marau	(1)
598. Marquesan	(1)
599. Maskelynes	(1)
600. Mayá	(1)
601. Merig	(1)
602. Minyaifuin	(1)
603. Misool	(1)
604. Morella	(1)
605. Morouas	(1)
606. Mpotovoro	(1)
607. Mukawa	(1)
608. Napu	(1)
609. Nduke	(1)
610. Nemi	(1)
611. Nggae	(1)
612. Nginia	(1)
613. Nobonob	(1)
614. North Ambrym	(1)
615. North Tanna	(1)
616. Orap	(1)
617. Paluan	(1)
618. Panatinani	(1)
619. Patep	(1)
620. Patpatar	(1)
621. Penudjaq	(1)
622. Petapa Taje	(1)
623. Piching	(1)
624. Pom	(1)
625. Port Sandwich	(1)
626. Quop Land Dayak	(1)
627. Ririo	(1)

628. Roglai	(1)	643. Sula	(1)	658. Vaikenu	(1)
629. Ron	(1)	644. Supan	(1)	659. Vano	(1)
630. Sa	(1)	645. Tabun	(1)	660. Vaturanga	(1)
631. Sadong	(1)	646. Tagakaulu Kalagan	(1)	661. Waay	(1)
632. Sakao	(1)	647. Tagol	(1)	662. Wagawaga	(1)
633. Salayar	(1)	648. Talaud	(1)	663. Wahai	(1)
634. Saliba	(1)	649. Tami	(1)	664. Wala	(1)
635. Saparua	(1)	650. Tanimbili	(1)	665. Wampar	(1)
636. Satawal	(1)	651. Tarikukuri	(1)	666. Wangka	(1)
637. Seko	(1)	652. Tau't Batu	(1)	667. Watut	(1)
638. Sengga	(1)	653. Tiang	(1)	668. Weda	(1)
639. Serawai	(1)	654. Toga	(1)	669. Western Subanon	(1)
640. Sesayap	(1)	655. Togian	(1)	670. Woi	(1)
641. Sikaiana	(1)	656. Ulithian	(1)	671. Wusi-Valui	(1)
642. Sudest (West)	(1)	657. Ura	(1)		

* 121 Lun Dayeh is also known as Lun Bawang.

† 186 refers to Bidayuh (Bukar-Sadong).

‡ 326 is actually Angkola-Mandailing Batak.

** 361 refers to Melanau (Dalat – Kampung Teh).

†† 495 refers to Melanau (Dalat – Kampung Kekan).

REFERENCES

Beaujard, Philippe. 1998. *Dictionnaire malgache–français: Dialecte Tañala, Sud-Est de Madagascar*. Paris: l'Harmattan.

Blagova, Galina Fedorova. 2000. *Etimologicheskiy slovar' tjurkskich yazykov*, vol. 6: *Obshchetyurkskie i mezhtyurkskie osnovy na bukvu "Q"*. Moscow: Indrik.

Blust, Robert. 1970. Proto-Austronesian addenda. *Oceanic Linguistics* 9:104–62.

———. 1972a. Proto-Oceanic addenda with cognates in non-Oceanic Austronesian languages: A preliminary list. *Working Papers in Linguistics* 4(1):1–43. Honolulu: Department of Linguistics, University of Hawai'i.

———. 1972b. Additions to "Proto-Austronesian addenda" and "Proto-Oceanic addenda with cognates in non-Oceanic Austronesian languages." *Working Papers in Linguistics* 4(8):1–17. Honolulu: Department of Linguistics, University of Hawai'i.

———. 1973. Additions to "Proto-Austronesian addenda" and "Proto-Oceanic addenda with cognates in non-Oceanic Austronesian languages"—II. *Working Papers in Linguistics* 5(3):33–61. Honolulu: Department of Linguistics, University of Hawai'i.

———. 1974. Eastern Austronesian: A note. *Working Papers in Linguistics* 6(4):101–7. Honolulu: Department of Linguistics, University of Hawai'i.

———. 1977. The Proto-Austronesian pronouns and Austronesian subgrouping: A preliminary report. *Working Papers in Linguistics* 9(2):1–15. Honolulu: Department of Linguistics, University of Hawai'i.

———. 1980. Austronesian etymologies. *Oceanic Linguistics* 19:1–181.

———. 1983/84. Austronesian etymologies – II. *Oceanic Linguistics* 22–23:29–149.

———. 1986. Austronesian etymologies – III. *Oceanic Linguistics* 25:1–123.

———. 1989. Austronesian etymologies – IV. *Oceanic Linguistics* 28:111–80.

———. 1988. *Austronesian root theory: An essay on the limits of morphology.* Amsterdam: John Benjamins.

———. 1996. The Neogrammarian hypothesis and pandemic irregularity. In *The comparative method reviewed: Regularity and irregularity in language change*, ed. by Mark Durie and Malcolm Ross, 135–56. New York: Oxford University Press.

———. 1999. Subgrouping, circularity and extinction: Some issues in Austronesian comparative linguistics. In *Selected papers from the Eighth International Conference on Austronesian Linguistics*, ed. by Elizabeth Zeitoun and Paul Jen-kuei Li, 31–94. Symposium Series of the Institute of Linguistics (Preparatory Office), Academia Sinica, No. 1. Taipei: Academia Sinica.

———. 2003. *Thao dictionary*. Language and Linguistics Monograph Series, A5. Taipei: Institute of Linguistics, Academia Sinica.

———. 2009. *The Austronesian languages*. Canberra: Pacific Linguistics.

———. 2011. The problem of doubleting in Austronesian languages. *Oceanic Linguistics* 50:399–457.

Brandstetter, Renward. 1916. *An introduction to Indonesian linguistics: Being four essays by Renward Brandstetter*. Trans. by C. O. Blagden. Royal Asiatic Society Monographs, Vol. 15. London.

Cauquelin, Josianne. To appear. *Nanwang Puyuma–English dictionary*. Language and Linguistics Monograph Series, W5. Taipei: Institute of Linguistics, Academia Sinica.

Collinder, Björn. 1955. *Fenno-Ugric vocabulary: An etymological dictionary of the Uralic languages*. Stockholm: Almqvist and Wiksell.

Dahl, Otto Chr. 1976 [1973]. *Proto-Austronesian*. 2nd. rev. ed. Scandinavian Institute of Asian Studies Monograph Series, No. 15. London: Curzon Press.

Dempwolff, Otto. 1934–1938. 3 vols. *Vergleichende Lautlehre des austronesischen Wortschatzes*. Zeitschrift für Eingeborenen-Sprachen, Supplement 1. *Induktiver Aufbau einer indonesischen Ursprache* (1934), Supplement 2. *Deduktive Anwendung des Urindonesischen auf austronesische Einzelsprachen* (1937), Supplement 3. *Austronesisches Wörterverzeichnis* (1938). Berlin: Reimer.

Dybo, Anna Vladimirovna. 2003. *Etimologicheskiy slovar' tjurkskich yazykov*, vol. 7: *Obshchetyurkskie i mezhtyurkskie osnovy na bukvy "L", "M", "N", "P", "S"*. Moscow: Vostochnaya Literatura.

Dyen, Isidore. 1953. *The Proto–Malayo-Polynesian laryngeals*. William Dwight Whitney Linguistic Series. Baltimore: Linguistic Society of America.

———. 1965. *A lexicostatistical classification of the Austronesian languages*. Indiana University Publications in Anthropology and Linguistics, and Memoir 19 of the International Journal of American Linguistics. Baltimore: The Waverly Press.

English, Leo James. 1986. *Tagalog–English dictionary*. Quezon City: Kalayaan Press.

Ferrell, Raleigh. 1982. *Paiwan dictionary*. Canberra: Pacific Linguistics.

Fey, Virginia. 1986. *Amis dictionary*. Taipei: The Bible Society.

Fox, Charles E. 1955. *A dictionary of the Nggela language (Florida, British Solomon Islands*). Auckland: Unity Press.

Grace, George W. 1969. A Proto-Oceanic finder list. *Working Papers in Linguistics* 2:39–84. Honolulu: Department of Linguistics, University of Hawai'i.

Greenhill, Simon, and Ross Clark. 2011. POLLEX-Online: The Polynesian Lexicon Project online. *Oceanic Linguistics* 50:551–59.

Hardeland, August. 1859. *Dajacksch–Deutsches Wörterbuch*. Amsterdam: Frederik Muller.

Hewson, John. 1993. *A computer-generated dictionary of Proto-Algonquian*. Seattle: University of Washington Press.

Hill, Kenneth C. 2012. Review of Brian D. Stubbs, "Uto-Aztecan: A comparative vocabulary." *International Journal of American Linguistics* 78(4):591–92.

Janhunen, Juha. 1981. Uralilaisen kantakielen sanastosta. *Journal de la Société Finno-Ougrienne* 77:219–74.

Kaufman, Terrence (with the assistance of John Justeson). 2003. *A preliminary Mayan etymological dictionary* (www.famsi.org/reports/01051).

Levitskaya, Liya Sergeevna. 1989. *Etimologicheskiy slovar' tjurkskich yazykov*, vol. 4: *Obshchetyurkskie i mezhtyurkskie osnovy na bukvy "Dzh", "Zh" i "Y"*. Moscow: Nauka.

———. 1997. *Etimologicheskiy slovar' tjurkskich yazykov*, vol. 5: *Obshchetyurkskie i mezhtyurkskie osnovy na bukvy "K", "Q"*. Moscow: Nauka.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig, eds. 2013. *Ethnologue: Languages of the world*. 17th ed. (web edition). Dallas, TX: Summer Institute of Linguistics, Inc.

Li, Paul Jen-Kuei, and Shigeru Tsuchida. 2001. *Pazih dictionary.* Language and Linguistics Monograph Series, A2. Taipei: Institute of Linguistics, Academia Sinica.

———. 2006. *Kavalan dictionary.* Language and Linguistics Monograph Series, A-19. Taipei: Institute of Linguistics, Academia Sinica.

Mallory, J. P., and Douglas Q. Adams, eds. 1997. *Encyclopedia of Indo-European culture*. London and Chicago: Fitzroy Dearborn.

Matisoff , James A. 1987–2013. *Sino-Tibetan etymological dictionary and thesaurus.* (http://stedt.berkeley.edu/~stedt-cgi/rootcanal.pl).

———. 2003. *Handbook of Proto–Tibeto-Burman: System and philosophy of Sino-Tibetan reconstruction*. Berkeley: University of California Publications in Linguistics.

McFarland, Curtis D. 1977. *Northern Philippine linguistic geography.* Study of Languages and Cultures of Asia and Africa Monograph Series, No. 9. Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa.

Milke, Wilhelm. 1961. Beiträge zur ozeanischen Linguistik. *Zeitschrift für Eingeborenen-Sprachen* 86:162–82.

———. 1968. Proto Oceanic addenda. *Oceanic Linguistics* 7:147–71.

Mills, Roger F. 1975. *Proto South Sulawesi and Proto Austronesian phonology.* PhD. diss., 2 vols., University of Michigan. Ann Arbor, Michigan: University Microfilms International.

__________. 1981. Additional addenda. In *Historical linguistics in Indonesia*, Part 1, ed. by Robert A. Blust, 59–82. NUSA, vol. 10. Jakarta: Universitas Katolik Indonesia Atma Jaya.

Pokorny, J. 1959. *Indogermanisches Etymologisches Wörterbuch*. 2 vols. Bern and Munich: Francke.

Ray, Sidney H. 1913. The languages of Borneo. *Sarawak Museum Journal* 1(4):1–196.

———. 1926. *A comparative study of the Melanesian island languages*. Cambridge: Cambridge University Press (in association with Melbourne University Press).

Reid, Lawrence A., ed. 1971. *Philippine minor languages: Word lists and phonologies*. Oceanic Linguistics Special Publication No. 8. Honolulu: University of Hawai'i Press.

Ross, Malcolm, Andrew Pawley, and Meredith Osmond, eds.1998. *The lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society*, vol. 1: *Material culture*. Canberra: Pacific Linguistics.

———, eds. 2003. *The lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society,* vol. 2: *The physical environment*. Canberra: Pacific Linguistics.

———, eds. 2008. *The lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society,* vol. 3: *Plants* Canberra: Pacific Linguistics.

———, eds. 2011. *The lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society*, vol. 4: *Animals*. Canberra: Pacific Linguistics.

———, eds. To appear. *The lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society*, vol. 5: *Body and mind*. Berlin: de Gruyter Mouton.

SEAlang. 2013. *Mon-Khmer etymological dictionary.* http://sealang.net/monkhmer/dictionary/.

Sevortyan, Ervand Vladimirovich, general editor. 1974. *Etimologicheskiy slovar' tjurkskich yazykov*, vol. 1: *Obshchetyurkskie i mezhtyurkskie osnovy na glasnye*. Moscow: Nauka.

———. 1978. *Etimologicheskiy slovar' tjurkskich yazykov*, vol. 2: *Obshchetyurkskie i mezhtyurkskie osnovy na bukvu "B"*. Moscow: Nauka.

———. 1980. *Etimologicheskiy slovar' tjurkskich yazykov*, vol. 3: *Obshchetyurkskie i mezhtyurkskie osnovy na bukvy "V", "G" i "D"*. Moscow: Nauka.

Shorto, Harry. 2006. *A Mon-Khmer comparative dictionary*, ed. by Paul Sidwell, Doug Cooper, and Christian Bauer. Canberra: Pacific Linguistics.

Stubbs, Brian D. 2011. *Uto-Aztecan: A comparative vocabulary.* Blanding, UT: Rocky Mountain Books and Publications.

Tryon, D. T. 1976. *New Hebrides languages: An internal classification*. Canberra: Pacific Linguistics.

Tryon, D. T., and B. D. Hackman. 1983. *Solomon islands languages: An internal classification*. Canberra: Pacific Linguistics.

Turner, R. L. (Sir Ralph Lilley). 1962–1966. *A comparative dictionary of Indo-Aryan languages*. London: Oxford University Press. (Includes three supplements, published 1969–1985.) http://dsal.uchicago.edu/dictionaries/soas/.

Verheijen, Jilis A. J. 1967. *Kamus Manggarai I: Manggarai–Indonesia*. The Hague: Martinus Nijhoff.

Wilkinson, R. J. 1959. *A Malay–English dictionary (Romanised)*. 2 vols. London: Macmillan.

Zoetmulder, P. J. 1982. *Old Javanese–English dictionary.* 2 vols. Koninklijk Instituut voor Taal-, Land- en Volkenkunde. The Hague: Martinus Nijhoff.

blust@hawaii.edu
steve@trussel.com