Swadesh Centenary Conference, 17-18 January 2009, Max-Planck Institute for Evolutionary Anthropology, Leipzig.

On using qualitative lexicostatistics to illuminate language history: some case studies.

Anthony P. Grant.

Abstract: Following certain aspects of the work on lexicostatistics carried out in the 1960s (Hooley 1971, Miller et al 1971, Miller 1984), and thereby working in a tradition which has been practised by Don Ringe, Tandy Warnow and others (Ringe et al, 1997, 2001), I maintain that much of lasting value can be learned about linguistic interrelationships by using techniques developed in work on *qualitative* (rather than merely *quantitative*) lexicostatistics.

1. Introduction.

Lexicostatistics has been invented and reinvented several times since the early 19th century, but the present understanding of the technique dates from Morris Swadesh's work of the 1950s, especially Swadesh (1950) and the papers that followed this (Swadesh 1954 and 1955 are especially of ote here.).[1] Most work on lexicostatistics has concentrated on examining pairs of languages, and sets of pairs of languages, and on putting the results of these pairwise comparisons in matrices.

In this paper I suggest with evidence from case-studies from around the world that a more illuminating approach to lexicostatistics from a diachronic point of view can be gained by examining cognate sets (and also isolated or 'unique') forms for each gloss within a cognacy grid. The use of this technique enables one to identify probable cases of shared lexical innovation between two or more languages.

Each etymological stratum of a language, even those strata comprising words of unidentified origin, has its own different significance within the history of a language, and techniques of analysis of such morphemes that can be applied to one language can often be applied to many or most. (See Swadesh 1951 on archaic residue and cumulative diffusion.) A major consideration to be borne in mind when one is examining the known or hypothesised history of languages. This is something which the use of lexicostatistics can assist if properly applied, is the establishment of the kinds and sources of morphemic etymological strata that are to be found within the language, especially in lexicon, but also among bound and free structural morphemes.

2. Constructing a cognacy grid.

The best approach to linguistic data for lexicostatistics is that of Don Ringe and his associates (e.g. Ringe, Warnow, Taylor, Michalove and Levison 1997, Ringe, Taylor and Warnow 2002), who take a taxonomic approach to this matter. They see entries on lists as characters, and the forms in each language denoting such entries as states of characters. Once the data have been assembled and recurrent sound correspondences have been identified and recognised, they need to be assigned codes for the purposes of quick reference. Coding each language or dialect's (or each isolect's) state of each character in a grid made of individual cells, in which each gloss is displayed horizontally and the character state from each language is displayed vertically, is meant to be maximally meaningful, once the reader understands the

---

[1] Hymes (1971: 327) says the technique seems to have first been presented as a paper by Swadesh at a supper-club session of the Viking Fund for Anthropology at New York on March 12 1948. In this initial and unpublished presentation Swadesh compared its performance in English and German on the one hand, and the Wakashan languages Nuuchahnulth and Kwakwala (Southern Wakashan Nootka and Northern Wakashan Kwakiutl) on the other.

system used. This coding procedure is the same in principle as that used in 'normal' lexicostatistics. The resulting grid, with its rows of a's, b's and so on, is a COGNACY GRID or cognate grid. Ideally no cell is left empty (if so it should be marked with a ?).

Some coding symbols are presented below, as they might be used for processing a dataset, involving several languages, that contains cognate and non-cognate data that have been taken from three languages. The data for each gloss from the first language in the table is always represented by the letter **a**, unless the character is a compound or a loan, or else if it is missing or unavailable in the language. This practice allows for the use of at least 26 columns of linguistic data within a dataset table or grid (though many studies, such as Nakhleh, Ringe and Warnow 2005, use numerals rather than letters to indicate the form for each character). For ease of comprehension, anomalous characters are here almost uniformly represented as occurring in the second of the three languages of the cognacy grid.

Some suggested sample coding rubrics follow:

a a a – this rubric indicates that the forms in all three languages for the item or character in question are cognate: all derive directly from the same ancestral form.

a a b – this rubric shows that the forms for this character in the first two languages are cognate, the last language's form isn't.

a b a – this shows that the forms in languages 1 and 3 (but not in 2) are cognate.

a b b – forms found in languages 2 and 3 are cognate but the form in language 1 isn't cognate with these.

a b c – the use of three different letters to represent the glosses in three languages shows that none of the forms of the character are cognate.

a B a – the use of a capital letter (after the usage of Miller 1984) shows that the character state in the second language is a recognised loan from another language.

a a,b c –the second language has two equally valid means of expressing an idea or encoding a character on the list, of which the first one is also found in the first language which is being surveyed, while the second one is not.

a b,c c – the second form listed as the response to a particular gloss (for we are not restricted to citing a single response to a gloss if this would distort the linguistic facts) is also to be found in the third language being surveyed.

a b,c d – this shows that neither response to the gloss for the second language is found with this sense in either the first or the third language being surveyed.

a a? a – there is some uncertainty about the status of cognacy of the second form with the first and third forms (which themselves are recognised as being cognates).

a a1 a – this is an indication that one of the languages uses a phonologically and/or morphologically irregular but still presumably cognate form of the same stem (what some Austronesianists call a 'sporadic sound-changed form', a linguistic analogue of the biological genetic *allele*) that is found in the other two languages.

a a+ a – this indicates that the second language uses a compound form to represent a certain concept, but that the other two languages use a simplex, and that the compound form includes the stem that is used in all three languages. Further details of the nature of the compounding can be given in footnotes.

a ? a – this indicates that the form for the concept in the second language is unknown or unavailable from the data that we have, although we may safely assume that there is a way of encoding this concept in the language in question. This unfortunate state of affairs should be clearly distinguished from the following rubric:

a 0 a – this indicates that the language lacks a word for the concept in question.

a b$ a – this indicates that the form in the first and third languages was once present in an earlier stage in the second language with the same shape and meaning but it has

undergone semantic shift from its original meaning, a meaning that is preserved in the other languages being surveyed. The replacement form(s) which is now used to express this meaning, and which is not cognate with the form(s) which is or are used in the other languages being surveyed, can be noted in footnotes, and the fact that it is not cognate with the preceding term can be entered in the table. This rubric can also be used in those cases where one or more languages being surveyed use distinct words for two concepts on the list, but where these concepts are bunched together under one phonological form in one or more of the other languages being surveyed.[7]

I used this in a paper on Caddoan interrelationships, drawing on data from the Caddoan languages of the Great Plains (Grant 1995), informed by data from Parks 1979 and unpublished Caddo work by Wallace L. Chafe, taking this approach because it seemed so obvious. Miller et al 1971, Miller 1984 and Hooley 1971 use similar techniques in their studies of other language groups.) Of course this approach should only be used AFTER sets of recurrent sound correspondences have been established.

Table 1 Some comparative Caddoan forms.

Gloss	Arikara	South Band Pawnee	Kitsai	Wichita	Caddo	Cognacy status of the various forms	Cognate sets (after Miller et al.1971)
'three'	*.táwIt*	*.tawit*	*.taawiku*	*.tawha*	*dahaw*	AAAAA	5
'one'	*.axkU*	*.asku*	*.asku*	*.chiass*	*wists'i.*	AAAAB#	4-1
'woman'	*.sápat*	*.capat*	*.cakwákt*	*.kaahiik'a*	*náttih*	AAABC	3-1-1
'fish'	*.ciwáhtš*	*.kaciiki*	*.nitát*	*.kaac'a*	*.batá*	ABCBD	2-1-1-1
'dog'	*.xáatš*	*.asaaki*	*.anosa*	*.wase'ek'a*	*diitsii'*	ABCDE	1-1-1-1-1

The subsequent tabulation of cognacy patterns that I found in the data, using the technique of horizontal set-referenced lexicostatistics advocated in this paper, gave the following findings, presented in Table 2 (# indicates the items is borrowed):

Table 2: Cognacy patterns in Caddoan languages for the modified 100-item list.

AAAAA	15
AABAA	1
AAAAB	18
AAAAB#	2
AABAC	3
ABBBC	2
ABAAC	1
ABCCD	1
AABBC	5
AAABC	18
AABCD	23
ABCBD	1
ABBCD	1
ABCDE	9

[7] In such cases, one can also put a note in the cell of the grid for the relevant language and concept with the number or gloss of the first item on the list which is encoded in the language by the label under discussion. For instance, if Language A has different words for an item numbered #50 and meaning 'day' and an item numbered #170 and meaning 'sun', whereas Language B uses the same term to express both ideas, then in the cell for 'sun' in the Language B column, one can put '(=50)' which indicates that the language uses the same word for 'day' and 'sun'.

3. Vertical and horizontal lexicostatistics.

Blust (2000) has added to lexicostatistical theory, making a terminologically useful distinction between *horizontal lexicostatistics* and *vertical lexicostatistics*. The former technique is the one which is more widely used nowadays (though this was not always so). This technique compares lexical data from languages which are supposed to have been attested in the same time period and to be roughly contemporaneous. Meanwhile the latter technique compares lexical data from an earlier stage of a particular language with data from other languages which are assumed to be descendants from this language. Comparisons between material from Classical Latin on the one hand (Latin being the control case or norm) and French, Spanish, Italian and so on, on the other, would be an example of the use of vertical lexicostatistics. (This is what Rea 1958, 1973, which I used for my 2008 Romance lexicostatistics paper, could have done.) Comparisons between French, Spanish and Italian would be instances of horizontal lexicostatistics. (This is what Rea 1958, 1973 did.) The study offered in this paper uses horizontal lexicostatistics as a point of departure, since most of the languages compared are contemporaries of one another.

For Blust (2000: 320) horizontal lexicostatistics is characterised by a known retention rate (which Swadesh had long since set at 0.81 per millennium, or 81/100 items that are supposed to be retained from the word list after a thousand years), an unknown period of divergence between the two or more contemporary languages that were being surveyed (indeed we may say that the time when these diverged was the question to which we were seeking an answer), and an ability to calculate these figures horizontally. With vertical lexicostatistics the rate of retention was unknown, but the time of divergence between the control case language and the descendant language(s) was supposed to be known, and the figures could be calculated vertically. The unspoken assumption is that in vertical lexicostatistics all the languages concerned diverge from the ancestral language to approximately the same degree and at the same rate. But this is not the case with horizontal lexicostatistics, and this is supposed to enable us to subgroup languages (and then to construct family trees) according to their depth or the recency of split from one another.

Combining the strengths of historical investigation and of the use of a cognate grid in norm-referenced lexicostatistics in which the norm comprises items from a reconstructed language allows one to take advantage of the strengths of the various subfields: the use of a well-selected lexical sample (a choice of material which is especially germane in the case of languages which have minimal inflectional morphology of the sort relied upon for historical linguistic purposes by diachronists), and the ability more clearly to see patterns of lexical distribution within a chosen sample of languages.

There is also the benefit that can be provided by working with due caution from a set of reconstructed forms, which (if we have enough historical background information to make assumptions secure) allows one to recognise whether the equivalent form in a modern language which is being surveyed is an inherited form that the proto-anguage contained, or whether it is one or another kind of innovation. Different kinds of such innovations would include borrowing (including the borrowing of a form which is cognate to one which might have been found in the lexicon of the proto-language under discussion, and thus a ‘false cognate’), internally-developed form, or whatever. Indeed Blust (2000) pointed out that it is the inability of horizontal lexicostatistics to be able to let us distinguish between inherited forms and other forms which are innovations shared between two or more languages (but not between all the languages that are being surveyed), which vitiates this technique.

Using vertical lexicostatistics where this is possible this confusion of the historical status of elements does not happen as readily.

4. Some case studies.

Below I present some case studies of qualitative lexiostatistics from a range of languages around the world. Some are based on data I collated myself; others are from the published literature; the Romance case combines published analysis and additional datasets which I collated from published sources. An Australian example drawing on data from O'Grady and Klokeid (1969) was discussed in Grant (2004).

4.1 Two Uto-Aztecan studies.

Miller et al (1971) compared 100 items in 36 Shoshone (more precisely Numic) varieties including Comanche, using a Swadesh 100-item list with 12 replacements. Only 4 of the 3600 cells were unfilled; 182 different cognates (including 40 uniques) are given. All 36 varieties show the same cognate for 62 of the 100 items, while a further 11 items have a cognate common to all but one lect. The same stem is found across 2/3 or more of the items for 93 items, and in a majority of the 36 lects for all but 4 items. This attests to the close relationship of the lects. The use of 100 cognates would indicate that the languages were essentially identical lects; the use of 3600 different cognates would suggest that they were 36 completely unrelated languages.

Casting his net more widely Miller (1984) used a revised Swadesh list with 12 replacements (not all of them the same as those used in the earlier study) across 32 Uto-Aztecan lects, including 9 lects which had been discussed in Miller et al (1971). 50 out of 3200 cells were left unfilled for lack of data. 3 items are also expressed in some of the languages by loans from Spanish, which Miller marks with capital letters. 952 cognate sets are found across these 100 items, including 441 items which are unique to one Uto-Aztecan language. Only 2 sets, 'path' and 'tooth', appear to be common to all U-A languages. The average cognate set is found among 3.31 Uto-Aztecan languages; if uniques are removed that proportion rises to 6.164 languages.

4.2 Latin and Romance.

Using the principles laid out in Miller et al (1971) and Ringe et al (2002), I turned the data in Rea's articles, which covered Latin (only in Rea 1958), French, Catalan, Spanish, Portuguese, Italian and Daco-Rumanian, and the additional data from Latin, Vallader and Logudorese into a grid, marking each cognate set (even if it had no cognates) with a separate letter. Borrowings were marked with an obelus after the letter; forms which shared a common root but different derivational forms were coded a, a2, a3 etc but were counted as cognates for general purposes. Logudorese was taken as typical of Sardinian; material comes from *Il Sardo in Tasco*. There were two gaps in the Latin list, so the number of boxes filled was two short of 223 x 9, thus 2005 boxes in the grid (but a few dozen boxes had more than one occupant).

The 223 glosses were expressed by 576 cognate sets containing 1 or more items. Yet only 209 items appear as cognate sets on the 100-word list. Were this rate of cognacy found in the 223-item list there would be 466 cognate sets. Only 47 forms out of the Swadesh 100-list items are common to all 8 Romance languages surveyed as the customary expressions for the concepts. This commonality of cognacy is true for 76 out of the 223 items in all.

Subgrouping conclusions for Romance are pretty much as predicted by pre-McMahon techniques. Logudorese Sardinian and Italian cleave to one another, as do

Spanish and Portuguese. Catalan is less clearly marked by Iberoromance innovations and more by its own innovations such as *gos* 'dog'. Lower Engadine (Vallader) Rhaeto-Romance pairs with French and Italian but is generally conservative, as is Sardinian. Daco-Rumanian is more on its own, but lexical evidence naturally suggests that Aromunian is more closely related to it (as are Megleno-Rumanian and Istro-Rumanian) through the use of shared innovations than other Romance languages are.

One thing which is significant from a methodological point of view is that the experiment shows the general robustness of Swadesh's judgments about which concepts tend to be encoded by words which are especially

4.3 Chamic and Malayic.

Chamic is a small group of Austronesian languages spoken in Vietnam and Cambodia, and most closely related to Malay. They have been profoundly influenced by surrounding Mon-Khmer languages over the past two millennia. Tsat of Hainan is also considered a Chamic language, while Acehnese and the Chamic languages are considered a subbranch within Malayic (Thurgood 1999).

In Grant (2005) I found that the number of forms used for expressing the 200 concepts on the sum total of the Chamic Hudson-Blust lists, apart from Acehnese, is 300, that is, a ratio of 1.5 different forms per gloss across a sample of eight languages. (A ratio of 1.00 would indicate to us that all the languages were identical isolects with nothing to distinguish one from another; a ratio of 8.00 would highlight to us that all indications suggested that the eight languages were completely unrelated to one another.) I have full lists for Phan Rang Cham, Western Cham, Jarai, Rade, Chru, Tsat and Northern Roglai in addition to Acehnese; the list for Haroi has 17 omissions

Now these 300 forms cover 1568 filled slots. The number of slots is arrived at as follows: Ideally I would have 200 forms from the 8 sampled non-Acehnese Chamic languages, making 1600 slots on the grid for these languages. But I have 17 gaps on my table for glosses for which I lack a form in one or another language. Additionally there are gaps in the columns for most Chamic languages for the 3pl pronoun form, since it is identical to the 3sg form in nearly all Chamic languages, and the same is true of most 2pl pronouns, which is usually identical to the 2sg pronouns, while some languages also use the same form for 'short' and 'small'. These slots could potentially be filled by 1568 different items, if it were the case that the languages in question bore no lexical resemblances between each other whatsoever. But in fact only 300 separate items are used (excluding a handful of cases in which one language uses two different unique forms to express the meaning of a particular gloss – only one unique is counted for such slots in each case). This makes this a ratio of 5.26667 slots per individual glossed item (for whatever this ratio may actually be worth; please note that this figure is the **reciprocal** of the figure for the average number of cognate sets per word across the eight languages surveyed).

If one adds into this total the forms on the list the Acehnese lst, and if one notes those forms which are only found in Acehnese among the Chamic languages (whether or not they are also retained from PMP or are also found in Malay, or whether they are innovations within Acehnese), the total of different forms rises to 369 and the ratio of cognate sets (plus uniques) per gloss therefore rises to over 1.8 forms per gloss across a sample of nine languages, exhibiting a total of 1756 slots (for we have a full list for Acehnese), making this a ratio of 4.7588 slots (out of nine slots available for each gloss) which are occupied on average by each individual glossed item. On this list, out of 199 items with separate glosses, 101 items are expressed by a single etymon, 54 items by 2 etyma; 25 by 3, 11 by 4, 7 by 5 etyma and 1 by six.

This overall very high degree ofcommonality of the basic lexicon in Chamic is, we should point out, in contrast to the very wide degree of phonological variation (if one views the matter diachronically) which is found across these languages and which is even amply exemplified in the various phonological shapes of those forms which have been inherited from PMP, but which is especially vivid in fully-tonal Tsat.

By comparison, the number of different forms which are used for the equivalents on the eight wordlists which were given for various Malay lects in Blust (1988)[21], a dataset which has fewer overall gaps than the Haroi list has, and one which represents a group of lects whose genetic unity has never been in doubt, is 490, or 2.45 forms per individual glossed item across a sample of eight lects. Of these, 64 forms are represented by the same etymon across all 8 varieties which Blust documented (including some loans such as *pikir* 'to think', < Arabic), 62 glosses by 2 etyma, 37 by 3, 18 by 4, 7 by 5 etyma, 7 by 6, and 3 items by 7 etyma, while only 2 are expressed on the Hudson-Blust list by different forms in all 8 varieties.

5. Conclusions.

Using qualitative techniques in lexicostatistics and drawing upon cognacy grids (which simply tabulate and reify the decisions about cognation upon which quantitative lexicostatistis also relies) enables us not only to compare percentages of similarities between languages or isolects but also to analyse the distribution of sets of forms which are most likely to indicate excusively sared innovations and also possible loans (especially when vertical lexicostatistics is available thanks to material from a soundly reconstructed proto-language which will ideally have been created without reference to the languages being surveyed).

Furthermore first approximations may be made in order to to identify how densely or closely the languages being surveyed are related to one another, and also the extent to which some or all of the languages in the group have undergone periods of separate lexical development away from the other languages. What is more, as the Latin-Romance study showed, there is strong evidence that the items on the 100-word list are more tenacious (and indeed admit of fewer borrowings) than those on the 200-word list, being richer in archaic residue and less full of diffusional cumulation than the 200-word list items (or indeed further set of vocabulary items) are. All this is explicit or implicit in Swadesh's work of the 1950s.

REFERENCES.

Adelaar, K. A. 1985. *Proto-Malayic: the reconstruction of its phonology and parts of its lexicon and morphology*. PhD dissertation, University of Amsterdam. Amsterdam: Kanters.

Blust, Robert A. 1981. 'Variation in the retention rate in Austronesian languages.' Paper presented at the Fifth International Conference on Austronesian Languages, Denpasar, Bali.

---1990. 'Malay historical linguistics: a progress report.' *Rekonstruksi dan cabang cabang bahasa Melayu induk*, edited by Mohammed Thain Ahmad and Zaid Mohamed Zaidi, 1-33. Kuala Lumpur: Dewan Bahasa dan Pustaka.

---

21 The Malay lects which are surveyed in that article are *Bahasa Indonesia, *Banjarese, Medan Malay, Selako, *Iban, *Minangkabau, *Jakartanese (Betawi) and Ambonese Malay (Bahasa Ambon). Adelaar (1991) uses five of these lists (the five which I have asterisked) and also provides a directly comparable wordlist for the Middle Malay language Seraway, providing equivalents for Seraway for 188 out of the 200 items on the Blust list. He additionally reconstructs Proto-Malayic forms from this evidence wherever possible. Neither author provides a list for Kerinci, which is usually classified as a phonologically divergent dialect of Minangkabau, although we do know that it retains 100 out of the 200 PMP forms that are used on the Blust list (Blust 2000b: 329). I have only recently had access to Blust's list for Kerinci and have therefore not used it in the above work.

---1993. 'Central and Central-Eastern Malayo-Polynesian.' *Oceanic Linguistics* 32: 243-292.
---2000. 'Why lexicostatistics doesn't work: the 'universal constant' hypothesis and the Austronesian languages.' *Time depth in historical linguistics*, ed. Colin Renfrew, April McMahon and Larry Trask, 311-331. Cambridge: McDonald Institute for Archaeological Research.
Blust, Robert, Russell D. Gray and Simon Greenhill. 2005-. *Austronesian Basic Vocabulary Database*. Dept. of Psychology, U. of Auckland. Accessible via the World Wide Web.
Grant, Anthony P. 1995. 'Connections in Caddoan: internal subgrouping and divergence.' Unpublished paper, 12 pp.
---2001. 'Possibilities and limitations of lexicostatistics: some indications from Romani and creole languages. *'Was ich noch sagen wollte…': a multilingual festschrift for Norbert Boretzky on occasion of his 65th birthday*, edited by Birgit Igla and Thomas Stolz, 273-286. Berlin: Akademie Verlag. (Sprachtypologie und Universalienforschung, Studia typologica 2.)
---2004. 'Upstreaming Swadesh: making the most of qualitative lexicostatistics.' Paper presented at the Annual Meeting of the Australian Linguistic Society, University on Sydney, July 2004.
---2005. 'Norm-referenced lexicostatistics and the case(s) of Chamic.' Anthony P. Grant and Paul Sidwell (eds.), *Chamic and Beyond: Studies in Mainland Austronesian languages,* 105-146. Canberra: Pacific Linguistics 569.
---2008. 'Swadesh, strata and strange fruit: on some lexicostatistical trees as applied to Romance. Paper presented to the XXXVI Comparative Romance Workshop, 3-4 January 2008, Trinity Hall, University of Cambridge.
Hooley, Bruce A. 1971. 'The Austronesian languages of Morobe District, Papua New Guinea.' *Oceanic Linguistics* 10: 79-151.
Miller, Wick R., James L. Tanner and Lawrence P. Foley. 1971. 'A lexicostatistical study of Shoshoni dialects.' *Anthropological Linguistics* 13: 142-164.
Minett, James W., ands Wang, William S-Y. 2003. 'On detecting borrowing: distance-based and character-based approaches.' *Diachronica* 20: 289-330.
Nakhleh, Luay, Don Ringe, and Tandy Warnow. 2005. 'Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages**.'** *Language* 81: 382-420.
Nakhleh, Luay, Tandy Warnow, Don Ringe, and Steven N. Evans. 2005. 'A comparison of phylogenetic reconstruction methods on an Indo-European dataset.' *Transactions of the Philological Society* 103: 171-192.
O'Grady, G. N. and Terry J. Klokeid. 1969. 'Australian linguistic classification: a plea for coordination of effort.' *Oceania* 39: 298-311.
Parks, Douglas R. 1979. 'The Northern Caddoan languages: their subgrouping and time depths.' *Nebraska History* 60 (2): 197-213.
Rea, John A. 1958. 'Concerning the validity of lexicostatistics', *International Journal of American Linguistics*, 24: 145-150.
Rea, John A. 1973. 'The Romance data of the pilot studies for glottochronology', in Sebeok, T. (ed.) *Diachronic, Areal and Typological Linguistics*, *Current Trends in Linguistics* 11, 355-367. The Hague: Mouton.
Ringe, Donald A., Tandy Warnow, Ann Taylor, Alexander Michailov, and Libby Levison, 1997. 'Computational cladistics and the position of Tocharian.' In Victor H. Mair (Ed.), *The Bronze Age and Early Iron Age Peoples of Eastern Central Asia, volume I:* 391-414. Philadelphia: Institute for the Study of Man in conjunction with the University of Pennsylvania Museum Publications.
Ringe, Donald A., Jr., Tandy Warnow and Ann Taylor. 2001. 'Computational cladistics and Indo-European.' *Transactions of the Philological Society* 98: 57-127.
Swadesh, Morris. 1950. 'Salish internal relationships.' *IJAL* 16: 157-167.
---1951. 'Diffusional cumulation and archaic residue.' *Southwestern Journal of Anthropology* 1-21.
---1952. 'Lexico-statistical dating of prehistoric ethnic contacts.' *Proceedings of the American Philosophical Society* 96: 452-463.
---1955. 'Toward greater accuracy in lexico-statistical dating.' *IJAL* 21: 121-137.
Thurgood, Graham. 1999. *From Ancient Cham to modern dialects.* Honolulu: University Press of Hawai'i.