Language in Native Title

Edited by
John Henderson
David Nash

Native Title Research Series
Aboriginal Studies Press

Australian Institute of Aboriginal
and Torres Straight Islander Studies

HISTORICAL LINGUISTIC GEOGRAPHY OF SOUTH-EAST WESTERN AUSTRALIA

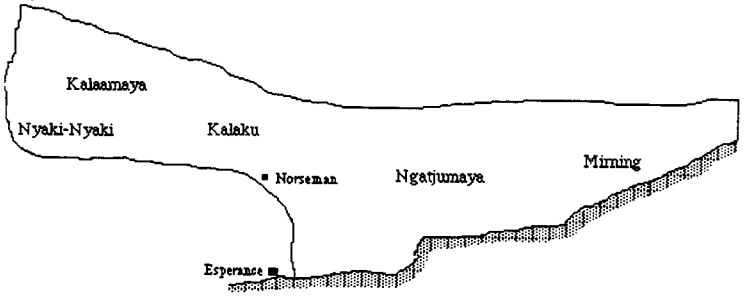
David Nash

Australian National University and AIATSIS

1. Introduction

This paper demonstrates something of the historical picture of relationships between language varieties that can be drawn from the rather meagre records that survive. In particular, we show how the technique of two-dimensional scaling can be applied to graphically represent relationships among vocabulary lists.

Our focus is on the languages of south-eastern Western Australia, the languages which are neither Western Desert nor of the Nyungar group – languages known today as Ngatju, Ngatjumaya, Marlpa, Kapun ('Gubrun'), Karlamayi, and Mirniny ('Mirning').[1] This group does not have an accepted name. O'Grady et al (1966) dubbed them the 'Mirniny Subgroup' which elevates the name of one particular subset. Here we follow Thieberger[2] 'languages of the South-East'; see below for more on language names.



Map 1. '4.1 The South-East' language map

From *Handbook of Western Australian Aboriginal languages south of the Kimberley Region.* Note the compiler's caveat (§1) that 'Maps at the beginning of each section show approximate traditional locations of languages.'

---

[1] This work has been carried out in the context of consultancies since November 1998 for the Goldfields Land Council, whose support is gratefully acknowledged. The work in this paper has been partly in collaboration with Nick Thieberger, who has made helpful comments on an earlier version. I have also benefited from comments made when the paper was presented at the workshop on 2 October 1999 and from editorial comments.

[2] Bates and von Brandenstein used the term 'Dundas District' after the administrative region, in turn from Lake Dundas.

I show that the documentary record of Aboriginal languages of the study region, from 1864 to the 1960s, has pretty much the same vocabulary in the same relative locations over that time, that is, the relative locations of languages (that is, the spatial arrangement of the various places of residence of speakers) show a continuity from the earliest records.

The historical comparative work necessarily concentrates on the distribution of vocabulary items, since few records contain expressions larger than the word. Few collections include numerous full sentences, and connected texts are recorded only in 1969-70 by von Brandenstein (1980). Texts, or at least a good collection of sentences, are needed to adequately study morphology and syntax. Without documentation of these parts of a language, comparisons with other languages can only be partial (even with languages related as dialects). Nevertheless, vocabulary comparisons of language records, even meagre ones, which have identifiable locations can provide useful indications of affiliations with languages more fully recorded in other places. Good examples are CC Hunt's (1864) record of 14 words, and Helms (1896:325) record of Hampton Plains vocabulary (collected in 1891), which are most similar to languages of the south-east rather than any other subgroup, notably the Western Desert subgroup.

A promising avenue of inquiry is to investigate the associations detectable between place and language. A fairly straightforward kind of evidence is a direct relation between a toponym (place name) and a word (both its form and meaning) in a particular Aboriginal language. A couple of such analysable toponyms is indicative; a cumulative pattern of such links is more convincing. Investigation of this kind is postponed until we have a reasonable idea of the varieties of language across the area of study – the kind of understanding which emerges from this paper and from which toponym studies can be a subsequent step.

Another link between language and space is investigated later in this paper. Clines of linguistic properties, or 'linguistic distance', can be correlated with spatial distance. If there is a good fit between the two measures, then we can say that each language variety has a spatial niche, the district where it fits in. Then the variation of vocabulary over south-east Western Australia is studied to bring out its spatial component.

2. Vocabulary comparison

The comparison here reported is of the historical records of vocabulary of the languages of south-east Western Australia. The historical records (to 1970) are considered first. The vocabularies recorded before Aboriginal people's residence moved to the larger towns also provides a detailed linkage of language variety with location, especially with respect to the vocabularies recorded by Daisy Bates (n.d.), which are usually assigned to specific Indigenous locations. While not the subject of this paper, the

study of old records is a good basis also for separate comparison with the language knowledge of people still alive.

As stated earlier, vocabulary is the basis of comparison of the language records since few old collections include expressions larger than the word. Our initial comparison of the historical records of vocabulary has focussed on 'basic vocabulary', from a standard list of meanings used in comparing the 'core' vocabulary of languages. The particular list of meanings used here is that used for 40-odd languages in the *Sourcebook of Central Australian Languages* (Menning and Nash 1981) (hereafter, *SCAL*), and includes the O'Grady (1960) adaptation[3] for Australia of the Swadesh 100-item and 200-item lists used world-wide. Concentration on 'basic vocabulary' goes with a common (though not unchallenged) assumption that such vocabulary is more stably inherited than vocabulary relating to more culturally specific meanings. In addition, there is the practical consideration that historical records tend to record vocabulary for 'basic' meanings, so it is more available for comparison.

3. Previous work

The first quantification of language relationships in south-east Western Australia came in a 1960s project spearheaded by G.N. O'Grady. This lexicostatistically based classification was set out in a special issue of *Anthropological Linguistics* and summarised on a wall map (O'Grady et al 1966a and 1966b);[4] its methodology was described by O'Grady and Klokeid (1969). In the south-east of Western Australia, the map shows three languages grouped in dialectal relationship, that is in pairs with over 70% of shared basic vocabulary. From west to east in south-east Western Australia, they are named 'Kalarko', 'Ngadjunma', 'Mirning'. These three together with Kalamai formed a subgroup, that is at least one of the three dialects shares between 50% and 70% of basic vocabulary with Kalamai. The collection of the four is named 'Mirniny Subgroup'.

The O'Grady et al (1966a,b) classification provided the structure for *A revised linguistic survey of Australia* (Oates and Oates 1970):

XXIX Pama-Nyungan family
A Southwest group
f Mirniny Subgroup
55.1a Mirning A22 [AIAS-A9]
55.1b Ngadjunma A25 [AIAS-A3]
55.1c Kalarko A11 [AIAS-A2]
55.2 Kalamai A12 [AIAS-A4]

---

[3] See http://www.anu.edu.au/linguistics/nash/aust/wl.html

[4] The accompanying article in *Anthropological Linguistics*, and O'Grady and Klokeid (1969:299-301), state the percentage of shared basic vocabulary ranges: dialect over 70%, language 50%-70%, subgroup 25%-50%, group 15%-25%, and family under 15%.

4. Comparison of south-east Western Australia vocabularies

We have taken vocabulary lists representative of the variety of all known sources and compared them to show similarities and differences. The lists range over time and space, from the 1880s to 1970 and from various locations within the south-east Western Australia region or from people who were from that region. As mentioned above, these lists represent language records from the study area, other than of Nyoongar or WD languages.

We undertake this comparison for several reasons:

- to investigate the continuity of the languages spoken in the region,
- to interpolate where a wordlist probably came from if there is no geographic reference on the list, and
- to classify the lists, especially lists for which we only have the name of the speaker and no name for their language.

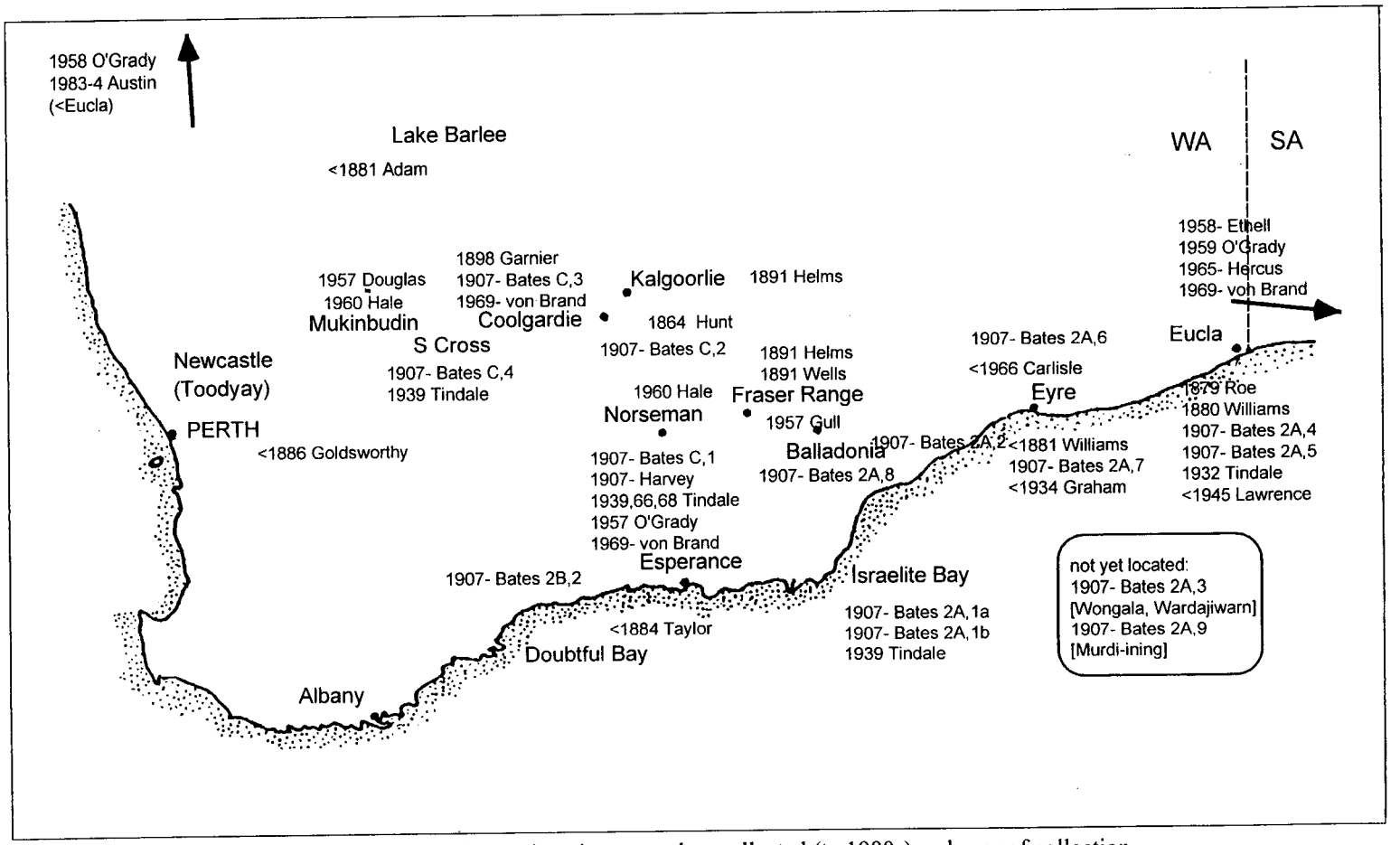
All the sources used in the study reported here are published or on open access in public archives in Australia. Those with more than ten items of the test list of 'basic' vocabulary are listed in Table 1. The sources in this table are located on Map 2 and Map 8. Table 1 also lists in the first column an identification code for those sources, all with more than 30 test-list items, for identification use in Table 2 and the later analysis.

Table 1. Source wordlists in comparative table, in chronological order, with size and location[5]

code	year	source	test list items	Lat S (deg min)	Long E (deg min)	location
A	<1881	Curr 34 Eyre	72	32°08'	126°18'	Eyre's Sand Patch
B	1880	Curr 35 Eucla	81	31°40'	128°52'	Eucla
C	1891	Helms	63	33°?	124°?	Symons Hill Rockhole, Fraser Ra.; Hampton Plains
D	1898	Garnier Coolgardie	50	30°58'	121°10'	Coolgardie

[5] Where the location is vague, I have used the coordinates of the place central to the district (for instance Norseman for 'Norseman district'); similarly where the source mixes several places I have provided coordinates for a point approximately in the middle of the implied area.

code	year	source	test list items	Lat S (deg min)	Long E (deg min)	location
E	1907-14	Bates 2A, 1a 39/2-33 Gauera	105	33°30'	123°30'	Brooks' Place,
F	1907-14	Bates 2A, 1b 39/34-47	58	33°37'	123°53'	70 Mile, Israelite Bay
G	1907-14	Bates 2A, 3 39/55-60	33	32°15'	126°09'	Wongala Wardajiwarn
H	1907-14	Bates 2A, 4 39/62-70	22	31°40'	128°52'	Eucla district
I	1907-14	Bates 2A, 5 part1 9/78-113	106	31°40'	128°52'	Jinyila
J	1907-14	Bates 2A, 6 part1 39/122-132	40	31°55'	126°04'	Nunginija, Yayoudle, Cocklebiddy
K	1907-14	Bates 2A, 8 39/191-212	106	33°03'	123°23'	Deralinya
L	1907-14	Bates 2B, 2 40/30-40	40	33°35'	120°03'	Woolbanup Hill
M	1907-14	Bates 2C, 1 47/2-26	104	32°10'	121°35'	Dundas district
N	1907-14	Bates 2C, 2 47/29-32	25	31°50'	121°40'	Widgiemooltha, Norseman
O	1907-14	Bates 2C, 3 47/35-56	82	30°58'	121°10'	Coolgardie
P	1907-14	Bates 2C, 4 47/59-84	89	30°45'	119°12'	Karratjibbin
Q	1907-14	Harvey in Bates 2C, 5 47/114-119	15	32°10'	121°35'	Norseman
R	1939	Tindale T86 Ngadjunma (Ngadju purara)	40	33°36'	123°53'	Israelite Bay
S	1939	Tindale T89 Kala:mai (includes 4 of genealogy Sheet 116 Kabu:(d)n)	45	31°14'	119°19'	Southern Cross
T	1939	Tindale T83 Mining	37	31°40'	128°52'	Eucla
U	1959	GN O'Grady S source	78	31°51'	132°42'	born nr Kooringibbie
V	1958	GN O'Grady J source	60	31°40'	128°52'	Eucla /Carnarvon
W	1960	GN O'Grady P source	100	32°15'	126°03'	Twilight Cove /NE Norseman
X	1960	Hale & O'Grady	63	30°55'	118°12'	Mukinbudin
Y	1970	CGvB RG	32	32°25'	123°53'	Balladonia / Norseman
Z	1970	CGvB PF	39	32°10'	121°35'	Norseman



Map 2. Locations where language data collected (to 1980s) and year of collection.

Comparison was restricted to equivalents of the 168-term test list used in *SCAL*, ignoring other meanings that may be represented in particular sources. To a limited extent near synonymous glosses are merged, for instance, words for 'man' may be taken from those forms glossed as 'Blackfellow' in an old record. Ideally the whole of each list should enter any comparison, but the restriction to the *SCAL* test list uses forms for meanings which are indeed most widespread in the available data and makes the task more manageable.

Vocabularies were entered in a comparative table with a column for each source and a row for each meaning. The forms were compared row by row. Forms judged to have a common root were assigned to the same correspondence set. (An alternative lexicostatistic method is to require full stem correspondence.) Note that correspondences are counted only if 'homosemantic' (having the same meaning).

Most lists lacked a majority of the 168 *SCAL* items (while some items had several words, that is several synonymous forms for a particular meaning). The number of SCAL items in each list (with at least one word) is given in the third column of Table 1. Since we allowed multiple synonyms, affinities between lists score a little higher than if the classic forced-choice-of-one is used. The resulting percentages are in Table 2, using identification codes given in Table 1.

Table 2. Percentages of corresponding vocabulary between pairs of 26 south-east Western Australia sources.

code	Curr Eyre	Curr Eucla	Helms	Garnier	2A,1 a	2A,1 b	2A,3	2A,4	2A, 5(1)	2A, 6(1)	2A,8	2B,2	C,1	C,2	C,3	C,4	C,5	T86	T89	T83	O'G S	O'G J	O'G P	Hale	vB RG	vB PF
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A																										
B	65																									
C	44	32																								
D	27	27	52																							
E	49	40	49	39																						
F	46	36	54	37	82																					
G	75	76	50	19	65	46																				
H	55	50	75	44	59	75	60																			
I	75	77	43	39	61	50	86	56																		
J	71	68	43	32	62	43	83	33	89																	
K	34	29	55	46	67	57	44	50	44	52																
L	53	41	68	40	64	68	44	50	47	54	69															
M	31	25	61	51	42	46	34	44	31	34	49	55														
N	41	33	89	57	50	61	42	67	43	38	55	82	64													
O	33	29	52	40	29	36	34	36	35	30	30	50	57	50												
P	34	32	56	44	37	43	46	36	39	31	34	55	64	50	79											
Q	58	54	91	50	58	64	67	100	62	64	62	80	69	80	54	58										
R	38	31	62	35	60	67	25	64	37	29	66	60	50	61	42	54	88									
S	48	46	56	32	37	50	36	45	46	33	46	36	55	50	70	86	80	49								

code	Curr Eyre	Curr Eucla	Helms	Garnier	2A,1 a	2A,1 b	2A,3	2A,4	2A, 5(1)	2A, 6(1)	2A,8	2B,2	C,1	C,2	C,3	C,4	C,5	T86	T89	T83	O'G S	O'G J	O'G P	Hale	vB RG	vB PF
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
T	82	86	54	44	65	58	82	50	83	73	57	36	44	38	48	50	86	39	50							
U	73	63	41	32	52	35	75	64	71	85	38	40	29	35	19	27	75	27	36	88						
V	69	63	53	37	58	56	79	67	71	84	46	63	43	47	38	38	80	36	43	95	75					
W	61	51	45	34	56	50	83	56	66	65	43	52	36	33	30	32	60	37	46	92	75	77				
X	45	33	54	35	44	50	38	57	47	29	44	60	74	50	86	88	50	50	92	50	39	56	40			
Y	53	47	74	67	60	63	36	29	54	55	78	75	52	60	47	37	50	71	47	79	44	65	64	53		
Z	44	13	40	31	48	71	33	60	40	25	57	60	42	43	23	29	75	54	31	44	33	47	41	42	75	
#	72	81	64	46	105	58	33	22	106	40	106	40	104	25	82	89	15	40	45	37	78	60	100	63	32	39

= number of items in this source fitting the *SCAL* test list

The percentages in Table 2 range from 13% (BZ) to 100% (HQ), and come from fractions with a large range of denominators, all a lot less than the potential 168. For instance, sources A and G have 72 and 33 test list items respectively, but only 16 of these overlap, that is, are word pairs with the same meaning. Of these overlapping 16 items, 12 are judged to correspond, giving a correspondence percentage of 12/16 = 75%, as can be read in Table 2 at the intersection of column A and row G.

From inspection of Table 2 one can see that most of the vocabularies appear to be related in a way that might be described as a 'dialect web'. Voegelin et al (1963) introduced the term 'family-like language'; 'dialect chain' has also been used. The terms apply to the south-east Western Australia situation of Table 2 in that most varieties (assuming for the moment that each source records a single language variety) can be linked to any other by way of intermediate links and with no link having less than 70% of vocabulary in common. For instance, sources O and U share only 19%, but consider the chain joining them of O-P 79%, P-S 86%, S-Q 80%, Q-T 86%, T-U 88%.

5. Procedural difficulties

Making the comparisons behind Table 2 presents procedural difficulties of well-known kinds.

First, there are the usual difficulties in dealing with vocabulary recorded in old sources. The lists vary considerably in their reliability due to the variable expertise of the person writing the language. Some have internally inconsistent spelling conventions which makes it difficult to determine what the writer was trying to record. Further, spellings for a given word can vary as exemplified in the forms on Map 4. These and similar hazards, and methods for dealing with them, are discussed in Thieberger (1995), a suitable manual.

More particular to this project were these areas where decisions and assumptions were necessary:

(a) Multiple synonyms were recorded for some meanings. For example Source B, the Curr Eucla list, has *jering* 'sandhill', *wanda* 'beach sandhills', and both are entered in the comparative table for the meaning 'sandhill'. Hence Source B is scored as identical to both Source U, which has *jirriny* alone, and with Source V, which has *wantarri* alone.

(b) In some instances we need to decide whether forms are independent, wary of covert double counting. The meanings 'to drink' and 'to eat' are often not expressed independently, for instance, sharing a stem meaning 'ingest', distinguished sometimes by compounding with the word for 'water, liquid', or the word for 'food'.

(c) In some instances a form may be considered to be double-headed so a decision is required as to how to score its correspondences. For instance words for 'to speak' may combine a root 'language ~ mouth' with a verb stem 'to make noise'. We have here scored correspondence by reference to the (inflecting?) verb root but not to the cognate object with which it is compounded.

(d) The 'homosemantic' constraint, usually adopted in lexicostatistics, studiously ignores corresponding forms with different meanings, even where they are probable cognates and the meaning shift is plausible. (A more detailed comparison would take account of presumed semantic changes.)

(e) In a few instances a word is identifiable as a recent borrowing. An example is in source H (Bates 2A,4) "*Mai anmadharra, boorrborr*" 'hungry', where the first of the terms is clearly from the Western Desert language, including a distinctive Western Desert language suffix *-tjarra*.

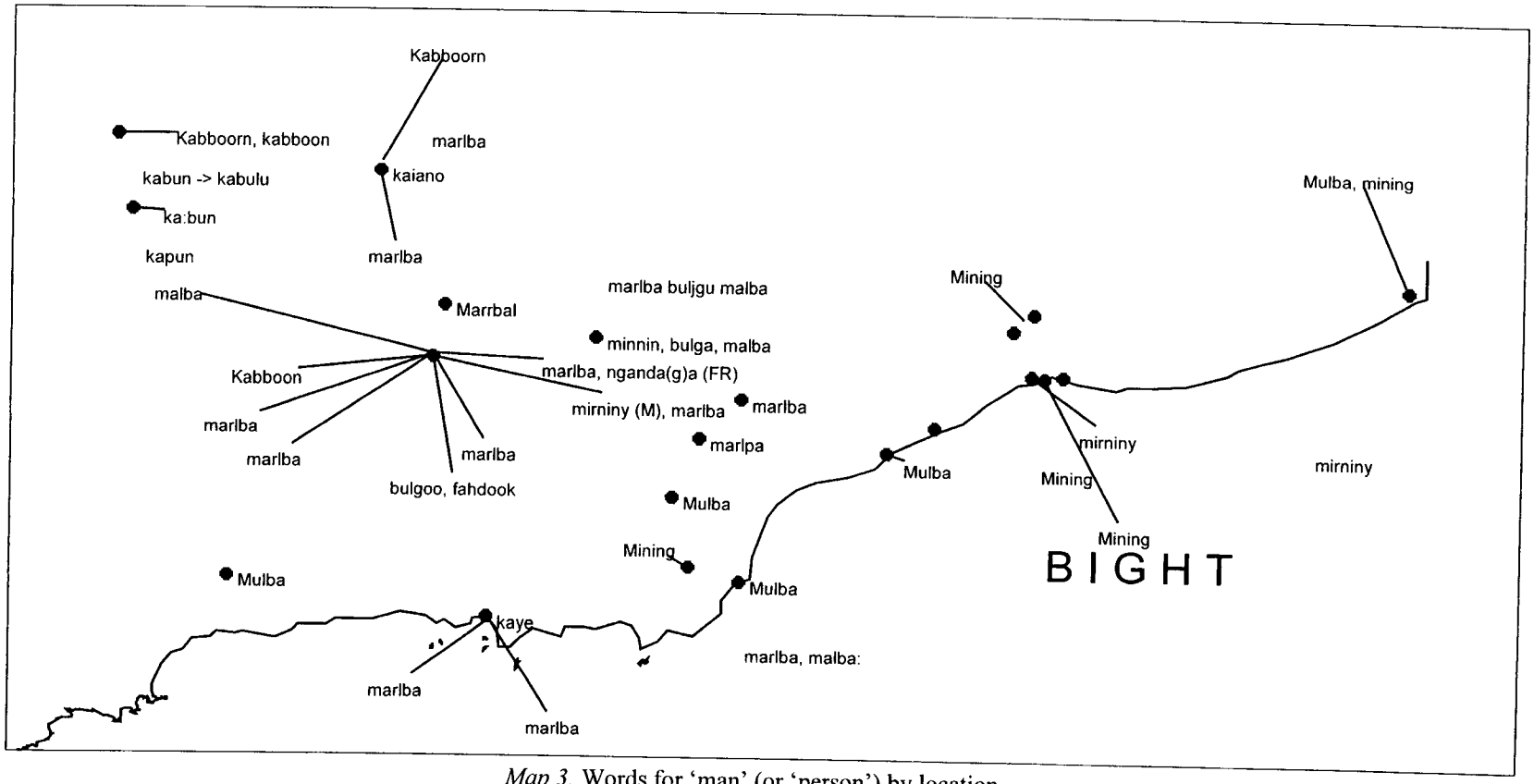
6. Significance

When interpreting the varying percentages in Table 2, we have to keep in mind a sampling error. The test list of 168 items is not large to start with; the sources are missing a proportion of test list data, and any pair of sources overlap in considerably fewer than the potential 168.

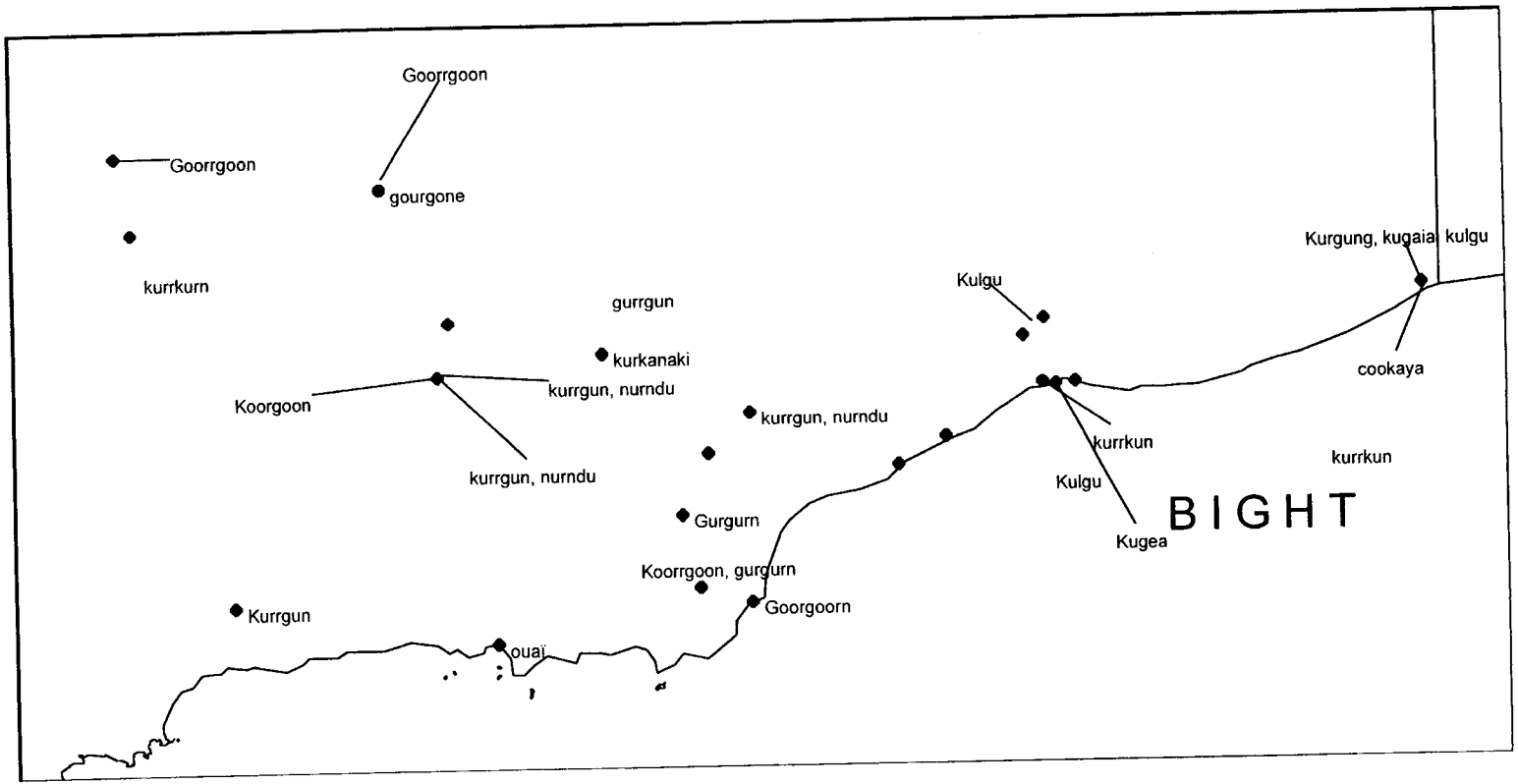
Statistical significance has been studied in the literature on lexicostatistics (for example Embleton 1986 and references there). To draw inferences about subgrouping it is preferable to use a list of something like 200 meanings. The comparative vocabulary data available to us is simply not this extensive. Black 1976:84n3 provides a method for 'a rapid estimation of the percentage point differences required for two percentages to be significantly different'. It may be inferred from Black's table that, for a 100-item list, a 9 percentage point difference is needed for a difference to be significant at the 20% confidence level (and 14 percentage point difference at the 5% confidence level) (for percentages in the range 5%-95%). In the analysis at hand, a 100-item list is about the largest size we have for a pairwise comparison.

Ideally, a more detailed philological comparison should be carried out, with careful attention to all analysable detail. As a practical compromise, it is productive to compare the less numerous sources in the light of findings from having compared the more numerous sources, remembering that the oldest records are commonly among the less numerous ones. Usually the short records show affinity with several of the geographically closest later or longer records.

Another aspect of using small vocabulary lists to compare languages is the possible effect on the calculations of conscious 'marker words' (including the various terms for localised groups and languages). An informant might express group allegiance in their self conscious answer to an inquiry as to the word for 'water', or 'Blackfellow' (a common term in Bates' lists) for instance, and give an emblematic reply, whereas the word provided for, say, 'elbow' (see for example Map 4) is not a conscious marker of group identity (not reported in Aboriginal Australia, at least).



Map 3. Words for 'man' (or 'person') by location



Map 4. Words for 'elbow' by location.

7. Pseudo-maps

Since the data of the languages of the study area are almost entirely vocabulary, thereby limiting us to lexical evidence for language classification, we have little choice but to pursue lexicostatistic similarity measures. Such similarity measures among the recorded language varieties (as in Table 2) are spatially interpretable. Indeed a two-dimensional synoptic layout of the various sources presents them in a way in which clusterings of degrees of similarity may be readily perceived like distances on a map.

This approach falls under the statistical method known as multi-dimensional scaling (MDS). Applied in psychology and many other fields, its use in linguistic applications have been few. The MDS approach was demonstrated in detail for analysing a group of fairly closely related languages by Black 1976, and Dyen et al 1992:70-76 applied MDS to Indoeuropean data. Dobson and Black 1979 is the sole published application of MDS to languages in Australia (specifically those of the Cairns rainforest). Woods et al (1986:262-5) is a textbook presentation of MDS for linguists, and Borg and Groenen (1997) is a recent statistics textbook on the subject.

The MDS graphic output, known as configurations, or pseudo-maps, can in the case of spatially tagged data be compared with the 'true' geographic map of the locational attributes. The comparison is most straightforward when two-dimensional configurations are derived. For languages, or particular language varieties, the known local affiliation of each variety is the obvious two-dimensional spatial attribute. The approach is to apply MDS to some linguistic similarity measure to see what internal relationships are revealed, and whether one of these is the spatial arrangement of the local varieties.

The MDS procedure cannot provide an external orientation for the pseudo-maps – the configurations are equivalent under rotation or mirror image flipping. However when there is some kind of geographic meaning available for each data point, then the configuration can be rotated and flipped so as to make the MDS configuration comparable with the geographic data. The source vocabularies of south-east Western Australia are of this nature, that is all can be geographically located, more or less – some to a particular small named locality, some rather to a district.

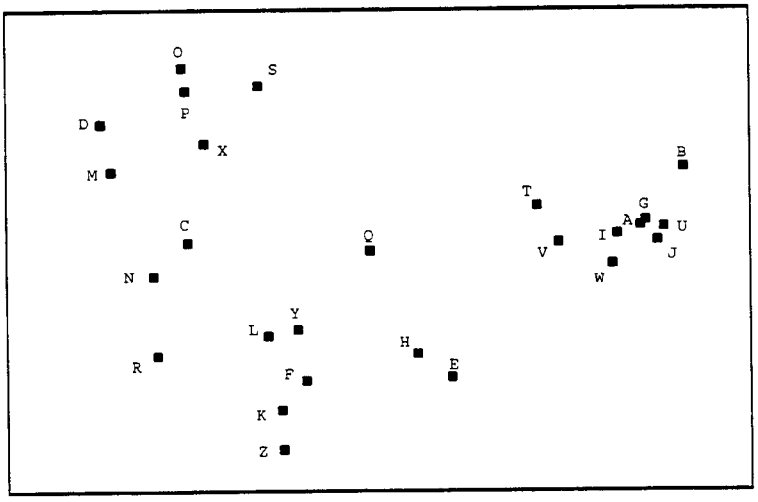
The similarity measure employed in this paper is the correspondence percentages of Table 2. We have allowed multiple synonyms for one meaning, partly because the data comes to us this way (especially from Bates) and there is no reason for choosing one synonym above the others. A consequence is that the pairwise percentages between vocabularies may not satisfy the 'triangle inequality', which theoretically can

Map 4. Words for 'elbow' by location.

undermine the derivation of pseudo-maps. 'In practice, however, such cases rarely result in any great deviations from triangle inequality, and such deviations could in any case be lessened or removed by some suitable nonlinear transformation of percentage into distance' (Black 1976:85n4). In any case, the potential problem is avoided in that we have used ordinal (non-metric) scaling rather than metric scaling.

To produce two-dimensional scaling pseudo-maps, the programs KYST and ALSCAL were chosen because they are the culmination of much refinement of multi-dimensional scaling programs in the 1970s. Both are discussed in recent textbooks. KYST was used by the largest (published?) lexicostatistical project (Dyen et al 1992:72); and the sleeker ALSCAL 84 is in routine use.[6]

Running ALSCAL 84 on the data of Table 2 of all 26 vocabularies produces the diagram or pseudo-map of Map 5.[7]



Map 5. Pseudo-map of 26 vocabularies.

[6] KYST is available from http://www.netlib.org/mds/kyst2a.dos/kyst2a.exe.gz, and ALSCAL 84 Alternating Least Squares Scaling (version 84.1) from ftp://192.41.30.119/pub/visuaftp/ALSCAL/

[7] In the MDS calculations the data points have to jostle each other, as it were, to give an overall compromise configuration which minimises the discrepancy between geometrical distances and the similarity measure of cognate percentages. The degree of mutual squash or squeeze can be summarised in a measure known as 'stress'. A particular method of calculating stress defines a measure known as $stress_1$ (Borg and Groenen 1997:33f). For the configuration of *Map 5*, $stress_1$ is 0.229, which is on the high side, but not unexpected given the assumptions of the model and error in the data. A 'scree plot' (Borg and Groenen 1997:37-38) has $stress_1$=0.129 in 4 dimensions, $stress_1$=0.153 in 3 dimensions, and $stress_1$=0.382 in 1 dimension.

Sources are placed on the pseudo-map so that the distances between any one source and all the others represent the overall relationship of similarity between that source and all others. As Dyen et al (1992:73) put it, 'there is a strong inverse relationship between *lexicostatistical percentage* and *distance*. If the percentage is large the distance is small; and if the percentage is small the distance is large.' The pseudo-map does not aim to show the geographical locations of the sources. The 'dialect web' nature of the relationships deducible from Table 2 are readily seen in Map 5 – every source can be linked to any other by way of short intermediate links, though the extremes are some distance apart.

Note that the reduction of the potential 25 dimensions to two dimensions means that in interpreting the pseudo-maps pairwise distance on the pseudo-map is not uniformly proportional to linguistic similarity, but the 'best fit' aims for a good correlation between pairwise distance on the pseudo-map and linguistic similarity. For example, in Map 5 source D is closer to M than it is to T, while the distance (on the pseudo-map) between D and M is roughly the same as between T and V. Compare the percentages in Table 2; D-M 51% is greater than D-T 44%, but T-V 95% is even closer than D-M. The reason that the distance D-M is much the same on the pseudo-map as the distance T-V is that the relationship with all the other sources is also brought to bear in the formation of the configuration of the pseudo-map; despite being almost the same (having 95% in common with each other), T and V relate somewhat differently to the other sources, for instance T-P 50% versus V-P 38%.

Extremes of mismatch in correspondence between percentages of Table 2 and pseudo-map distances in Map 5 are exemplified by:

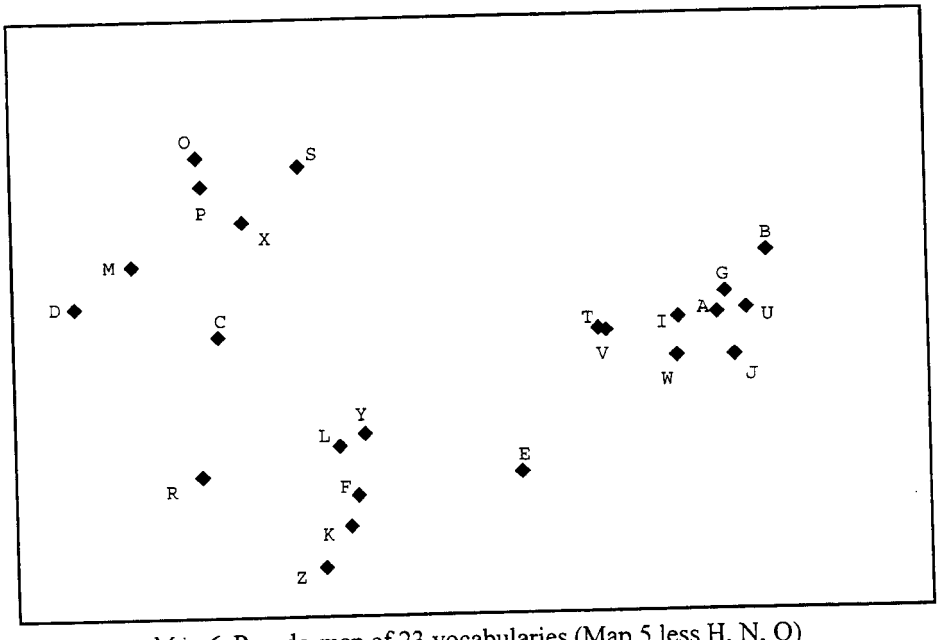
- C-T (54%) which are separated by about seven times the distance separating D-M (51%, much the same percentage);
- K-T (57%) which are separated by about eight times the distance separating K-Z (57%, the same percentage).

The rationale for these mismatches comes from seeing them as extremes in the general distribution of the matches between the 325 (=26*25/2) inter-source percentages and inter-source pseudo-distances. For instance, K-T needs to be more separate than K-Z to accommodate the way that K is generally less similar to other sources with higher percentages in common with T (A, B, G, I, Q, U, V, W, all > 80%) and more similar to other sources with higher percentages in common with Z (F, Q, Y, all > 70%).

The relative configuration of vocabulary similarity matches reasonably well with the known geographic source of each record – compare Map 5 with Map 8 or Map 2 (via the codes in Table 1 or Table 3). The striking mismatches with geographic locations are these:

1. The most striking mismatch concerns source L. L is embedded in the configuration in Map 5 but is actually from the Ravensthorpe district, west of Esperance and well away from all other source locations. L being very much a Marlba language, we cannot explain why Bates associates it with Ravensthorpe (well within the Nyoongar area).
2. Source R was collected by Tindale in Norseman and said to be from Israelite Bay, which is on the coast south-east of Norseman; yet Map 5 places R even further away from sources along the Bight coast (A, B, F, G, H, L, T, V, W) than other Norseman sources (N, Q, Y, Z).

To try to reduce the margin of error somewhat, ALSCAL 84 was then run on the data, extracted from Table 2, of those 23 vocabularies with more than thirty SCAL items, producing the diagram of Map 6.[8] This entails discarding sources H, N and Q, which, coincidentally, include two of the three sources which we know are a mix of language varieties (see the 'ind?' column in Table 3).



Map 6. Pseudo-map of 23 vocabularies (Map 5 less H, N, Q)

Map 6 still resembles a dialect web, with three clusters readily discernible, in the top-left, lower-left, and mid-right of Map 6

[8] $Stress_1$ is acceptable, 0.182. A 'scree plot' has $stress_1$=0.101 in 4 dimensions $stress_1$=0.128 in 3 dimensions, rising to $stress_1$=0.362 in 1 dimension. This pattern points to two being the appropriate dimensionality and to the appropriateness of excluding sources H, N and Q.

(normalised to correspond to the north-west, south-west and east of the study area).

8. Three languages?

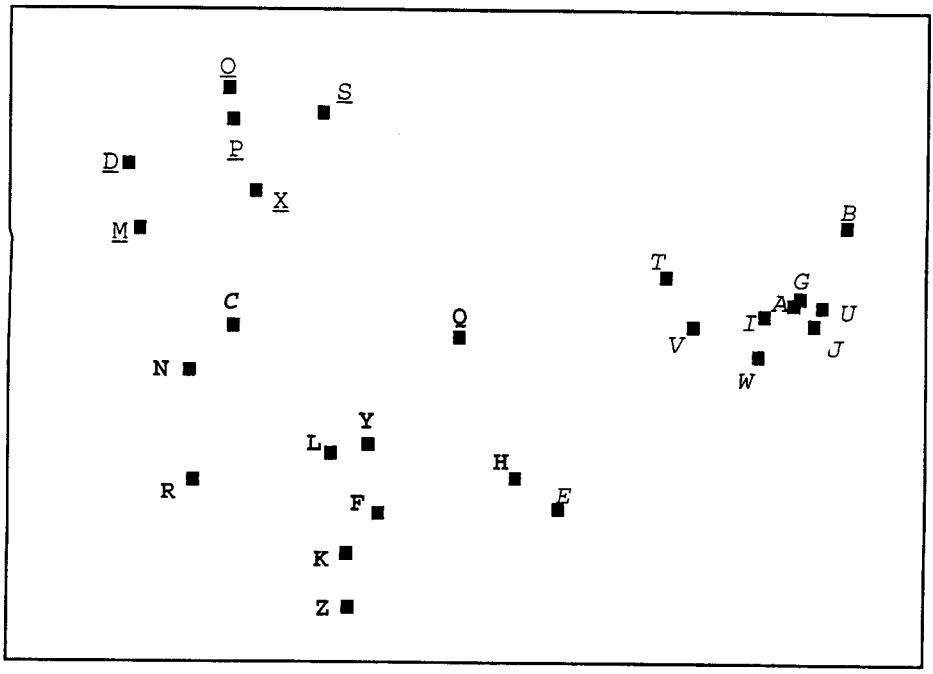
The compelling question asks what is going on in the clustering apparent in Map 6, what are its determining factors? A three way distinction across the study area is made by the label for a language variety according to its word for 'man' or 'person' (usually the same word has both meanings).

The sources are listed again in Table 3 with their word for 'man' or 'person' in the right hand column. If the term is capitalised, then it has been supplied as a name of the language or people.[9] The left hand columns in Table 3 are the identification as in Table 1, including the letter key to the two pseudo-maps Map 5 and Map 6. The middle column indicates (where we know from the provenance) whether the source is a single variety, or probably a compilation from several individuals and a mix of language varieties.

[9] The complexities of language naming in Australia mean that commonly used names should not necessarily be taken to reflect linguistic similarity; see McConvell's paper on ethnonyms, this volume.

Table 3. Source vocabularies and 'person' word

code	source	ind?	'person'
A	Curr 34	?	meening
B	Curr 35	?	majilba, meening
C	Helms	N	minnin, bulga, malba
D	Garnier	Y	kaiano
E	2A,1a	Y	Mining
F	2A,1b	Y	Mulba
G	2A,3	Y	Mining
H	2A,4	N	Mulba, mining
I	2A,5(1)	Y	Mining
J	2A,6(1)	Y	Mining
K	2A,8	Y	Mulba, Baaduk people
L	2B,2	Y	Mulba
M	C,1	Y	Kabboon
N	C,2	Y	Marrbal
O	C,3	Y	Kabboorn
P	C,4	Y	Kabboorn, kabboon
Q	C,5 Harvey	N	Malbert tribe, bulgoo, fahdook
R	T86	Y	marlba, malba:
S	T89	Y	kabun -> kabulu
T	T83	Y	mining
U	GNOG S	Y	mirniny
V	GNOG J	Y	mirniny
W	GNOG P	Y	mirniny
X	Hale	Y	kapun
Y	CGvB RG	Y	marlba
Z	CGvB PF	Y	marlba, nganda(g)a (FR)
	GNOG GAL		marlba [ms. buljgu malba]

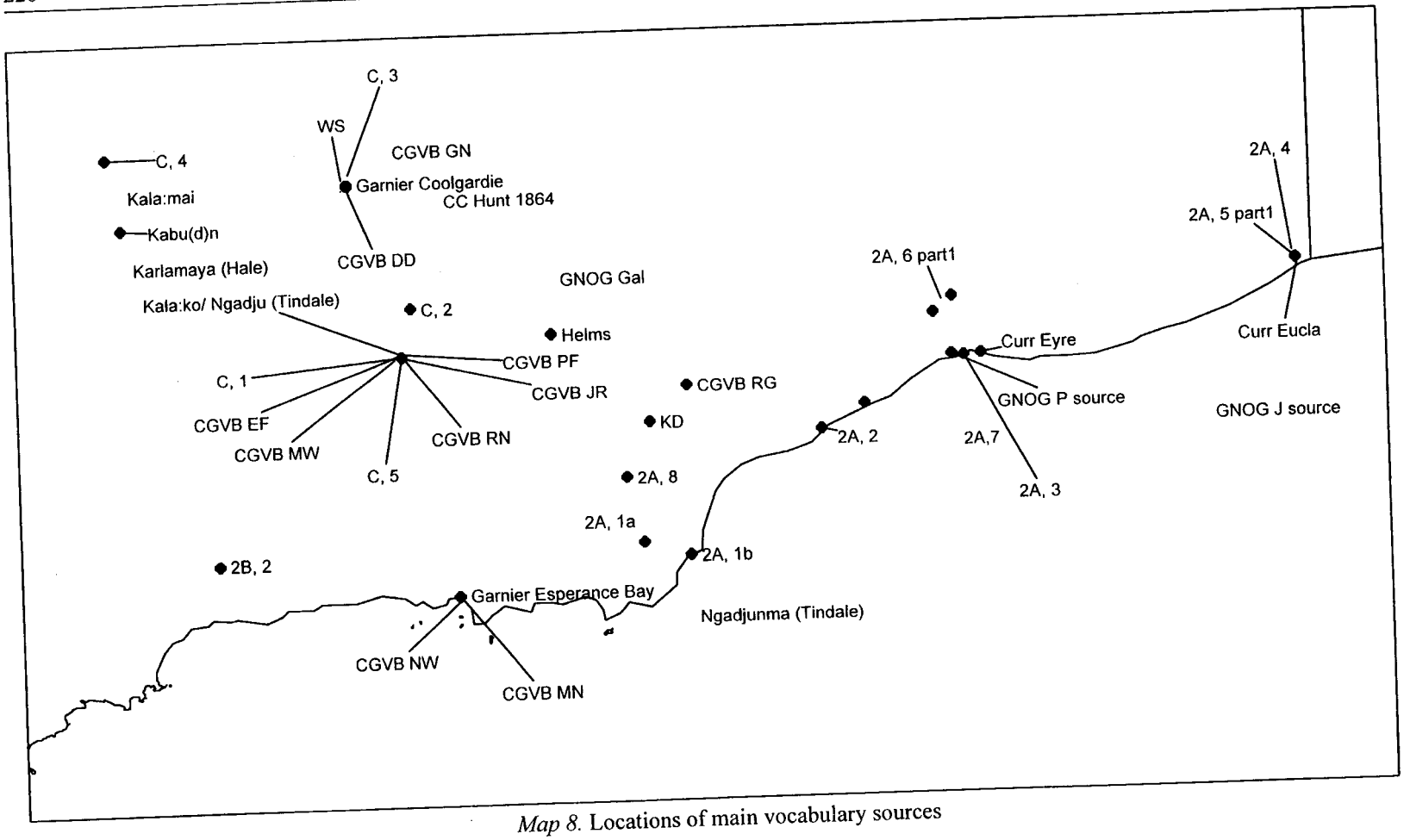


Map 7. Sources of Map 5 with code letter marked according to 'person' word – kabun (underline), **marlba** (bold), or *mirniny* (italic). Map 7 is the configuration of Map 5 annotated with this three-way distinction.

There is an obvious correlation between the 'person' word (and variety name) and relatively high percentage of vocabulary correspondences, as displayed in Map 7 (compare Map 3). The correlation is enhanced in the light of the following comments about sources marginal to the clusters:

- Source Q (compiled by the Norseman doctor) could well be a mixture, and in any case is not numerous.
- Source H has both marlba and mirniny as 'person' words, and is unusual for a Bates vocabulary in being assigned merely to a district (Eucla) and to no named people.
- Source C is said by its compiler (Helms) to be a mixture, and has both *marlba* and *mirniny* as 'person' words.
- Source E has *mirniny* as the 'person' word, but is from a location well to the west of other Mirniny locations, and closer to Marlba locations; as is clear on Map 3.

In short, in the study area the 'person/man' language names, or labels of language distinctions, correspond well to the three observed vocabulary clusters.



Map 8. Locations of main vocabulary sources

9. Interpreting the results of pseudo-mapping

As noted earlier, the relative configuration of vocabulary similarity matches reasonably well with the known geographic source of each record. The extent to which there is mismatch is of further interest. Mismatch could be attributed to a variety of influences. First, there could be bias in the vocabulary attributable to their manner of collection and preservation. Second, there are factors inherent to the language situation. In each individual's interview responses there could well be a conscious 'clustering to the norm' of one of the named languages of the study area.

A third, and perhaps most interesting, possibility is that the vocabulary similarity measure is tapping into older patterns of relationship between the language varieties. Detectable diffusion (borrowing) of vocabulary and other linguistics traits generally proceeds from neighbour to neighbour, and the more recent a borrowing the more readily detectable it generally is. Thus similarities due to borrowing and diffusion are expected to present a spatial cline. Relationships between languages dating from an earlier period, when spatial arrangements may have been different from in historical times, can be expected to leave a trace – the genetic component of the shape of a language, where the vocabulary and other properties of daughter languages are inherited from ancient ancestor languages. Note that the use of correspondence percentages, as in this paper, which include correspondences due to borrowing as well as inheritance, suits the data somewhat more to spatially interpreted MDS (as noted by Dobson and Black 1979:58ff). Thus it is a virtue rather than an drawback that the figures of Table 2 probably include sharings due to borrowing.

Given the general persistence of particular languages short of catastrophes, that is, that a community's language does not (indeed, hardly could) change quickly (within a generation, say), we would expect the historical record of match between location and language variety also to approximate well the situation for some generations prior to the period recorded (that is prior to 1880-1970).

10. Conclusion

The various historical vocabularies are records of a 'dialect web' or 'family-like language' with three discernible clusterings. The three clusters correspond fairly well with the word for 'person' in each vocabulary, words which have been in long use as language names, namely Mirniny, Marlba and Kabun.

Further, the spatial relationships among the vocabularies are comparable across more than a century of records and hence allow a view of locality of linguistic knowledge. The results are evidence for continuity, in that the relative configuration of vocabulary similarity matches reasonably

well with the known geographic source of each record. The span of implied continuity covers the generations of the historical records, and, by reasonable extrapolation, some generations prior, approximating thus the time span relevant to demonstration of continued native title.

I see potential in extending the quantified approach of this paper (test-list percentages, pseudo-mapping) to more recent lexical data, bringing present-day linguistic knowledge into view. It may also prove fruitful to integrate into the comparison various grammatical elements (for instance, some nominal suffixes attested in several sources), which are known to persist as indicators of particular linguistic lineages.

References

A. Vocabulary sources (Table 1)

Bates, Daisy M. n.d. [vocabularies] National Library of Australia, Manuscript Collection, MS365.

Curr, E.M. 1886 *The Australian Race*, Vol 1 Government Printer, Melbourne.

Garnier, Jules 1902-03 Vocabulaire des indigènes de l'Australie occidentale, Société Neuchâteloise de Géographie, *Bulletin* 14:247-251.

Helms, Richard 1896 Anthropology. *Transactions of the Royal Society of South Australia*, 16:237-332.

O'Grady, G.N. 1967-68 [Australian vocabularies computer files] University of Hawaii. Copy deposited at AIATSIS, ASEDA 0169.

O'Grady, Geoff 2001 Two Southern Australian Vocabularies: Parnkalla (Barngarla) and Karlamayi. In Jane Simpson et al (eds) *Forty Years on: Ken Hale and Australian Languages*, Canberra: Pacific Linguistics 512:291–303.

Tindale, N.B. 1939 Australian Vocabularies, ms, South Australian Museum.

von Brandenstein, C.G. 1980 *Ngadjumaja: An Aboriginal Language of South-East Western Australia*, Innsbrucker Beitrage zur Kulturwissenschaft, 48, Institut für Sprachwissenshaft der Universität Innsbruck, Innsbruck.

B. Other references

ALSCAL 84 Alternating Least Squares Scaling (version 84.1) ftp://192.41.30.119/pub/visuaftp/ALSCAL/

Black, Paul 1976 Multidimensional Scaling Applied to Linguistic Relationships, *Cahiers de l'Institut de Linguistique de Louvain* 3.5-6(Dec):43-92.

Borg, Ingwer and Patrick Groenen 1997 *Modern Multidimensional Scaling: Theory and Applications*, Springer (Springer series in statistics), New York.

Dobson, Annette J. and Paul Black 1979 Multidimensional Scaling of Some Lexicostatistical Data, *The Mathematical Scientist* 4:55-61.

Dyen, Isodore, Joseph B. Kruskal and Paul Black 1992 An Indoeuropean Classification: A Lexicostatistical Experiment, *Transactions of the American Philosophical Society* 82:5.

Embleton, Sheila M. 1986 *Statistics in Historical Linguistics*. Studienverlag Brockmeyer. (Quantitative linguistics Vol. 30), Bochum.

Handbook of Western Australian Aboriginal Languages South of the Kimberley Region, Pacific Linguistics, Australian National University, Canberra, 1993; web version © Nicholas Thieberger 1996.

Hunt, Charles Cooke 1864 Entry for Camp 45, 30.9.1864, Journal of Exploration Eastward by York in 1864, Field Book 1. Microfilm copy in State Records, ground floor, LISWA. Unpublished typescript transcript of Hunt's journal in Battye Library, LISWA, catalogue no. Q 994.12 HUN.

KYSThttp://www.netlib.org/mds/kyst2a.dos/kyst2a.exe.gz

Menning, Kathy and David Nash, (eds.) 1981 *Sourcebook for Central Australian Languages*, Institute for Aboriginal Development, Alice Springs.

Oates, W.J. and L.F. Oates 1970 *A Revised Linguistic Survey of Australia*, Australian Institute of Aboriginal Studies, Linguistic series no.12, Australian Aboriginal Studies, no.33, Canberra:.

O'Grady, G.N. 1959 Significance of the Circumcision Boundary in Western Australia, BA thesis, University of Sydney.

_____1960 More on Lexicostatistics, *Current Anthropology* 1(4):338-339.

_____1966 Proto-Ngayarda Phonology, *Oceanic Linguistics* 5:71-130.

O'Grady, G.N. and T.J. Klokeid 1969 Australian Linguistic Classification: A Plea for Co-Ordination of Effort, *Oceania* 39(4):298-311.

O'Grady, Geoffrey N. and C.F. and F.M. Voegelin 1966a Languages of the World: Indo-Pacific Fascicle Six, *Anthropological Linguistics* 8(2):1-197.

O'Grady, G.N., S.A. Wurm, and K.L. Hale 1966b *Aboriginal Languages of Australia (a Preliminary Classification)*, University of Victoria, Victoria, B.C.

Thieberger, Nick, (ed.) 1995 *Paper and Talk: A Manual for Reconstituting Materials in Australian Indigenous Languages from Historical Sources*, Aboriginal Studies Press, Canberra.

Voegelin, C.F., F.M. Voegelin, S. Wurm, G.N. O'Grady and T. Matsuda 1963 Obtaining an Index of Phonological Differentiation from the Construction on Non-Existent Minimax Systems, *IJAL* 29:4-28.

Woods, Anthony, Paul Fletcher and Arthur Hughes 1986 *Statistics in Language Studies*, Cambridge University Press, Cambridge Textbooks in Linguistics, Cambridge.