



This is the **author's version** of a work that was **accepted** for publication after peer review. This is known as the post-print.

**Citation for author's accepted version**

Black, Paul (post-print). Lexicostatistics with massive borrowing: The case of Jingulu and Mudburra. Retrieved from <http://espace.cdu.edu.au/view/cdu:42483>

**Citation for publisher's version**

Black, Paul (2007). Lexicostatistics with massive borrowing: The case of Jingulu and Mudburra. *Australian Journal of Linguistics*, 27(1):63-71.

**Notice:** *The publisher's version of this work can be found at:*

<http://dx.doi.org/10.1080/07268600601172959>

Cover page for paper submitted to the Australian Journal of Linguistics

Title: Lexicostatistics with massive borrowing: The case of Jingulu and Mudburra

Author: Paul Black

Affiliation: Charles Darwin University

Postal address:

CAESL

Charles Darwin University

Darwin, NT 0909 Australia

Phone: +61-8-8946 7761

Fax: +61-8-8946 7144

Email: <paul.black@cdu.edu.au>

# Lexicostatistics with massive borrowing: The case of Jingulu and Mudburra

*While heavy lexical borrowing can pose a problem to any approach to linguistic prehistory, it has often been regarded as an especially difficult problem for lexicostatistics, especially in such areas as Australia, where some believe that extensive borrowing is the norm. The present paper applies lexicostatistics to what is arguably the most massive case of borrowing known for Australia, namely between the Jingulu and Mudburra languages of the Northern Territory, and finds that it actually leads to what is generally considered the correct genetic classification of these languages. This result is then shown to depend on certain relationships among the lexicostatistical percentages that may not always obtain in other cases of heavy borrowing.*

## **1. Introduction**

It is widely believed that lexical borrowing raises serious questions about lexicostatistical approaches to subgrouping. The present paper considers a particularly striking example of this, namely massive borrowing reported by Pensalfini (2001: 393–4) between the Mudburra and Jingulu languages of the

Northern Territory of Australia. Pensalfini (2001: 395) suggested that his own lexicostatistical results for these languages ‘underline the dangers of taking lexicostatistical information alone as an indicator of genetic relatedness’. However, the present paper will show that a lexicostatistical approach utilizing Dyen’s (1963) technique for detecting and correcting for borrowings actually leads to the correct classification despite the unusually heavy borrowing. After presenting the details, the paper concludes by pointing out the conditions under which borrowing can and can not be detected in a lexicostatistical study.<sup>1</sup>

## **2. The languages**

Even though Mudburra and Jingulu are adjacent, no classification has proposed that they share a common protolanguage more recent than the putative Proto-

---

<sup>1</sup> The research for this paper was originally carried out under a consultancy with the Northern Land Council, which has kindly given me permission to report these results here. For the present paper I have omitted two Yirram (or Jaminjungan) languages included in the original study and added five meanings to my lexicostatistical list, but in any case the results are the same as in the original study. I am grateful to Mary Laughren and to Rob Pensalfini for access to certain sources of data.

Australian ancestor of most or perhaps all Australian languages. In both structure and vocabulary Mudburra appears to be closely related to such nearby languages as Gurindji and Jaru (see e.g. Nash 1999, McConvell & Laughren 2004) within what is generally taken to be a Ngumpin subgroup within the widespread Pama-Nyungan family. Jingulu, on the other hand, is markedly different from other Australian languages. While the early classification of Australian languages by O'Grady, Voegelin and Voegelin (1966) did not group it with any other Australian language, work by Chadwick (1979: 655) suggests that it joins with such varieties as Ngarnga, Gudandji and Wambaya to the east and southeast to form what is now called the West Barkley group. It has also been proposed that West Barkley in turn is a branch of a discontinuous Mindi family along with the Yirram (or Jaminjungan) languages to the northwest (e.g. Chadwick 1997: 95, Pensalfini 2001: 385), but this matter need not concern us here.

While Mudburra and Jingulu thus appear to be genetically remote, Pensalfini (2001: 392) reports that their speakers of Jingulu have lived together for generations. In his thesis Pensalfini (1997: 21) described the situation as follows (note that Jingili is the name for the speakers of Jingulu):

The resulting mixing of Mudburra and Jingili people produced a cultural group who are referred to (by themselves in many cases, and by older

Jingili) as ‘Kuwarrangu’, distinct from either Jingili or Mudburra.

Kuwarrangu households, until recently, would have been places where Mudburra and Jingulu were spoken alongside one another, and there was a great deal of lexical borrowing between languages in these households, with the result that it is possible to identify Kuwarrangu dialects of both Mudburra and Jingulu.

The lexical borrowing between Mudburra and Jingulu was in fact so heavy that Pensalfini (2001: 393) found that they actually shared up to 71% related forms (cognates or borrowings) on a standard 200-item lexicostatistical list. This is an amazingly high figure, as high as one might expect to find between distant dialects of the same language, but it is based on treating meanings as having cognate forms even if the shared cognate is only one of a number of synonyms in one or both languages. Pensalfini (2001: 392) also notes that Jingulu speakers are aware that some of the forms come from Mudburra, which raises questions as to whether such forms should be treated as Jingulu for lexicostatistical purposes, especially since a common lexicostatistical practice is to ignore known borrowings in an attempt to count only cognates. However, even if one uses a very strict requirement that all forms for a meaning be cognate across the two languages, Mudburra and Jingulu still share 40% (Pensalfini 2001:394).

Whether one considers the low of 40% or the high of 71% between Jingulu and Mudburra, the percentage is much higher than any shared with the other languages considered by Pensalfini (2001: 393–394). The following table shows both the lower and the higher percentage for each pair of varieties, of which Yirrim will not be considered further here.

Table 1. Lexicostatistical percentages reported by Pensalfini

Wambaya

21–34 Jingulu

16–24 40–71 Mudburra

13–19 11–26 17–36 Yirram

From the above it may seem that there is no way for a lexicostatistical classification to avoid the incorrect conclusion Jingulu and Mudburra form a subgroup. There is, however, and it depends on extending the comparison to include additional varieties, including ones more closely related to Mudburra.

### **3. A new lexicostatistical comparison**

The present study will compare eight varieties, namely the four West Barkely varieties Wambaya, Gudandji, Ngarnga and Jingulu and the four Ngumpin varieties of Eastern Mudburra, Western Mudburra, Gurindji and Jaru. For most of these varieties Menning and Nash (1981) conveniently provided 168-item word lists, which were thus taken as an initial basis for the comparison. This was the only source of data for Gurindji and Jaru, while for Mudburra I ignored this souce in favour of an electronic Mudburra vocabulary (1997) which allowed the distinction to be made between Eastern and Western Mudburra. For the West Barkley languages I supplemented the Menning and Nash (1981) data with more authoritative sources, including Chadwick (1975) for Jingulu, Chadwick (1971, 1979) for Ngarnga, Aguas (1968) and Chadwick (1979) for Gudandji, and Nordlinger (1998) for Wambaya.

My source for Jingulu was thus quite different from that of Pensalfini (2001), who presumably used his own data (perhaps as found in Pensalfini 1997). While this affects the comparability of my results with Pensalfini's, it seemed appropriate for two reasons. Firstly, as a matter of the principle a lexicostatistical study should not treat known borrowings as cognates, and I believed I could avoid these by using Chadwick (1975), which I expected to include only words that speakers identified as Jingulu and to exclude any that they identified as being from Mudburra. Secondly, since Jingulu is moribund, it

seemed reasonable to use an earlier source that was less likely to be contaminated by Mudburra forms that came to be used by Jingulu speakers only as their own language was going out of use. While my Jingulu percentages will thus not reach the extremes found by Pensalfini (2001), they will nonetheless continue to reflect extensive borrowing.

While the word lists in Menning and Nash (1981) listed 168 meanings, I reduced the number to 148 to avoid more problematic items. I dropped nine meanings for which forms were attested in no more than four of the eight varieties, namely *come here* (as distinct from *to come*), *to know*, *to lay*, *louse*, *rotten*, *sandhill*, *soon*, *to tell*, and *vagina*. The eleven others I ignored included *all* (not entirely distinct from *many*?), *cold* (*in head*; sometimes confused with *cold* (*weather*)?), *dry* (whether the quality or the season), *to kill/strike* (hard to distinguish from *hit*), *to lie down* (not clearly distinguished from *to tell a lie* in some sources), *person* (should it be equated with *people*, as found in other sources?), *that* (*remote*; often not distinct from *that* (*distant*) in these languages), *warm* (*weather*; not always clearly distinguished from the feeling or sensation), *wet* (sometimes confused with the wet season?), *yam* (some sources seem to only list specific varieties), and *yes* (not always distinct from *all right*?).

Only for Gurindji were forms attested for all of the 148 meanings that were used. For Jingulu there were 147, Wambaya 145, Eastern Mudburra 144, Jaru 135, Ngarnga 132, Gudanji 129, and Western Mudburra 129. For each pair of languages the number of meanings that could be compared thus ranged from a high of 147 (between Jingulu and Gurindji) to a low of 114 (between Western Mudburra and either Gudanji or Ngarnga).

For each pair of varieties, the forms for each meaning were counted as ‘cognate’ if any word for that meaning in the one variety seemed a possible cognate of any word for the same meaning in the other. These ‘cognates’ could thus include undetected borrowings, since there is no principled way of distinguishing borrowing from cognates in the data.

In some cases it was difficult to decide whether or not two forms should be treated as cognate. For the meaning ‘bark’, for example, Eastern Mudburra *barnngirri* is identical to Gurindji *panyirri* except for purely orthographic choice between *b* and *p* and, more importantly, the cluster *rnng* in the former as against *ny* in the latter. The latter sort of difference was not otherwise observed between these two varieties (or elsewhere), and it thus makes their cognition problematic, although it is possible that it could result from regular (if rare) phonological change, or even an error in attestation.

To allow for such uncertainties provision was made to treat such forms as ‘uncertainly cognate’, and for each pair of languages two lexicostatistical percentages were calculated, a lower one treating uncertain cognates as non-cognate and a higher one treating them as cognate. Citing the two as a range has the advantage of making the extent of the uncertainty clear while suggesting that the true percentage should be somewhere within the range.

The resulting lexicostatistical percentages are as follows:

Table 1. Lexicostatistical percentages among eight varieties

Wambaya

77–83 Gudanji

54–61 50–57 Ngarnga

28–35 22–32 24–35 Jingulu

15–24 14–21 16–22 40–43 Eastern Mudburra

8–18 10–17 6–14 19–22 89–90 Western Mudburra

10–15 9–15 7–11 9–12 46–50 59–63 Gurindji

6–13 6–13 4–8 6–9 27–36 30–40 35–43 Jaru

Recall that the upper four varieties belong to the West Barkley group while the lower four belong to the very remotely related Ngumpin group. The 40% to 43% shared by the last member of the former group (Jingulu) and the first member of the latter (Eastern Mudburra) is thus anomalously high, if not nearly as high as the 71% found by Pensalfini (2001: 393) when known borrowings were apparently also counted.

While Jingulu's percentage with Eastern Mudburra is 40–43%, note how its percentage with Western Mudburra is only about half this, at 19–22%. This represents an anomaly. Eastern and Western Mudburra are dialects of a single language, as their 89–90% shared cognates suggests. They must have shared a substantial common history of development before they began to diverge, so one would expect their relationships with all outside varieties to be much the same. That Jingulu shares more vocabulary with Eastern Mudburra than with Western suggests that Eastern Mudburra must have borrowed many of these forms from Jingulu, which also makes sense in view of their geographical proximity. (While Jingulu may also have borrowed from Mudburra, as suggested by Pensalfini (2001: 392), this would not explain the differences in its percentages with Eastern and Western Mudburra.) This in turn suggests that the lower percentages Jingulu shares with Western Mudburra are more likely to

represent the true relationship between Jingulu and Mudburra as a whole, at least to the extent they are less likely to reflect such borrowings.

This is the reasoning behind Dyen's (1963) approach to detecting borrowing; see also Black (1997) and Minett and Wang (2004) for recent work along similar lines. To implement this approach in deriving a subgrouping from Table 1, we would first use the highest percentage in the table, the 89–90% between Eastern and Western Mudburra, to group these two varieties into the Mudburra language as a whole. We then need to decide what percentages to use to represent the relationships of Mudburra as a whole with the remaining varieties. Because we have reason to believe that some of the percentages have been inflated due to undetected borrowing, we would select the lower sets of percentages in each case. For example, we would ignore Eastern Mudburra's 40–43% with Jingulu in favour of Western Mudburra's 19–22%.

Proceeding in that way would not group Jingulu with Mudburra, but instead it would yield something very close to the widely accepted classification of these languages. However, perhaps it could be argued that the ability to distinguish between Eastern and Western Mudburra has provided an advantage that we are unlikely to have in similar cases. Accordingly I am going to set the Western Mudburra data aside and proceed with a 'worse case' scenario that assumes that we have only the heavily borrowed Eastern Mudburra data. In this case

deriving a classification from the lexicostatistical percentages would proceed as follows:

1. The first step is to find the highest set of percentages in the table.  
Ignoring the Western Mudburra percentages, this is the 79–85% between Wambaya and Gudanji, which thus group to form the McArthur River language. While their percentages with other languages do not differ greatly, to minimize the problem of borrowing we will use the lower ones to represent the group.
2. Thus ignoring Ngarnga's 54–61% with Wambaya, the next highest percentages are Ngarnga's 50–57% with Gudanji and hence the McArthur River language as a whole. This gives us the Eastern group of the West Barkly languages,. Again favouring their lowest percentages with languages out side the group, our table would now look like this:

Table 3. Lexicostatistical percentages after step 2

Eastern group

22–32 Jingulu

14–21 40–43 Eastern Mudburra

7–11 9–12 46–50 Gurindji

4–8 6–9 27–36 35–43 Jaru

3. Gurindji next groups with Eastern Mudburra at 46–50%. This is only marginally higher than Eastern Mudburra’s 40–43% with Jingulu, and the grouping is problematic because Gurindji’s 9–12% with Jingulu is so much smaller than the latter set of percentages. This is a key step, because if we accept this grouping it suggests that there has been massive borrowing between Jingulu and Eastern Mudburra (as we know there has been), and it gives us reason to ignore the Jingulu-Eastern Mudburra figure of 40–43%.

The alternative would have been to ignore the highest figure and use that 40–43% to first group Eastern Mudburra with Jingulu, with the implication that the borrowing between Eastern Mudburra and Gurindji has been even more massive. However, there is no reason to favour the latter, somewhat more extreme alternative, so we accept the first one, and again we use the lower sets of percentages to represent the relations of the combined ‘Mudburra-Gurindji’ group with the remaining varieties. Our table now looks like this:

Table 4. Lexicostatistical percentages after step 3

Eastern group

22–32 Jingulu

7–11 9–12 Mudburra-Gurindji

4–8 6–9 27–36 Jaru

4. It's now clear that Mudburra-Gurindji groups with Jaru at 27–36% to form the Ngumpin group while the Eastern group joins with Jingulu to form the West Barkley group at 22–32%. The relationship between these two groups is quite remote, at some 4% to 12% shared vocabulary.

This lexicostatistical approach thus arrives at the widely accepted classification of these languages, and in doing so it has also produced evidence supporting the hypothesis of heavy borrowing between Jingulu and (Eastern) Mudburra.

#### **4. Conclusion and discussion**

From the lexicostatistical evidence alone it seems that borrowing between Jingulu and Mudburra has been so extensive as to raise lexicostatistical percentages based on relatively basic vocabulary by about thirty percentage points. Even so, the lexicostatistical data has not only provided a basis for detecting such heavy borrowing, but also for correcting for it, so that it leads to the the widely accepted classification, which is surely correct. It does this even without the data from Western Mudburra, although the addition of that data makes the situation even clearer, suggesting that borrowing from Jingulu into

Mudburra account for about twenty percentage points of the rise, with the remainder due to borrowing from Mudburra into Jingulu.

This is not to say that lexicostatistics can always detect and correct for the effects of borrowing. The reason that Pensalfini's (2001) lexicostatistical percentages suggested grouping Jingulu with Mudburra was in part because he did not consider data from Mudburra's closest relative, Gurindji. If he had done so, and if he had excluded from his Jingulu data forms that Jingulu speakers described as being from Mudburra, then he might have found, as I did, that the lexicostatistical evidence for grouping Mudburra with Gurindji was stronger than that for grouping it with Jingulu. At the same time, however, if it happened to be the case that Mudburra had no close relative like Gurindji, then there would have been no basis for detecting or correcting the borrowing between Jingulu and Mudburra.

## References

- Aguas EF 1968 'Gudandji' in *Papers in Australian linguistics no. 3* (= *Pacific Linguistics* ser. A, no. 14) Research School of Pacific Studies Australian National University Canberra: 1–20.
- Black P 1997 'Lexicostatistics and Australian languages: Problems and prospects' in D Tryon & M Walsh (eds) *Boundary rider: Essays in honour*

- of Geoffrey O'Grady* (= *Pacific Linguistics* ser. C no. 136) Research School of Pacific and Asian Studies Australian National University Canberra: 51–69.
- Chadwick N 1971 'Ngarndji wordlist and phonological key' in *Papers on the languages of Australian Aboriginals* Australian Institute of Aboriginal Studies Canberra: 34–45.
- Chadwick N 1975 *A Descriptive study of the Djingili language* Australian Institute of Aboriginal Studies Canberra.
- Chadwick N 1979 'The West Barkly languages: An outline sketch' in SA Wurm (ed.) *Australian Linguistic Studies* (= *Pacific Linguistics* ser. C no. 54) Research School of Pacific Studies Australian National University Canberra: 653–711.
- Chadwick N 1997 'The Barkly and Jaminjungan languages: A non-contiguous genetic grouping in North Australia' in D Tryon & M Walsh (eds) *Boundary rider: Essays in honour of Geoffrey O'Grady* (= *Pacific Linguistics* ser. C no. 136) Research School of Pacific and Asian Studies Australian National University Canberra: 95–106.
- Dyen I 1963 'Lexicostatistically determined borrowing and taboo' *Language* 39: 60-66. (Reprinted in Dyen I 1975 *Linguistic subgrouping and lexicostatistics* Mouton The Hague: 128–136.)

- McConvell P & M Laughren 2004 'The Ngumpin-Yapa subgroup' in C Bown & H Koch (eds) *Australian languages: Classification and the comparative method* Benjamins Amsterdam: 151–177.
- Menning K (comp.) & D Nash (eds) 1981 *Sourcebook for Central Australian Languages* (pilot edn) Institute for Aboriginal Development Alice Springs.
- Minett JW & WS-Y Wang 2004 *An analysis of the lexical skewing method for detecting language contact* electronic document found at <[http:// www. ee. cityu. edu. hk/ ~lelab/ minett.../ Lexical\\_ skewing. pdf](http://www.ee.cityu.edu.hk/~lelab/minett.../Lexical_skewing.pdf)> on 21 August 2004.
- Mudburra vocabulary* 1997 an electronic text file '9990520masterMudburra. TXT' describing itself as having been prepared by David Nash and others between 1990 and 1997.
- Nash D 1999 *Mudburra language* (created Dec. 1995, last addition 30 Jan. 1999) electronic document found at <[http:// www. anu. edu. au/ linguistics/ nash/ aust/ m/ index. html](http://www.anu.edu.au/linguistics/nash/aust/m/index.html)> on 10 Jan. 2003.
- Nordlinger R 1998 *A grammar of Wambaya, Northern Territory (Australia)* (= *Pacific linguistics*, ser. C no. 140) Pacific Linguistics Canberra.
- O'Grady GN CF Voegelin & FM Voegelin 1966 'Languages of the World: Indo-Pacific Fascicle Six' = *Anthropological Linguistics* 8(2).

Pensalfini R 1997 *Jingulu grammar, dictionary, and texts* PhD thesis

Massachusetts Institute of Technology.

Pensalfini R 2001 'On the typological and genetic affiliation of Jingulu' in J

Simpson D Nash M Laughren P Austin & B Alpher (eds) *Forty years on:*

*Ken Hale and Australian languages* Pacific Linguistics Canberra: 385–399.