Claire Bowern*

Data “Big” and “Small” – Examples from the Australian Lexical Database

Abstract: The twenty-first Century has been billed the era of “big data”, and linguists are participating in this trend. We are seeing an increased reliance on statistical and quantitative arguments in most fields of linguistics, including the oldest parts of the field, such as the study of language change. The increased use of statistical methods changes the types of questions we can ask of our data, as well as how we evaluate the answers. But this all has the prerequisite of certain types of data, coded in certain ways. We cannot make powerful statistical arguments from the qualitative data that historical linguists are used to working with. In this paper I survey a few types of work based on a lexical database of Pama-Nyungan languages, the largest family in Aboriginal Australia. I highlight the flexibility with which large-scale databases can be deployed, especially when combined with traditional methods. “Big” data may require new methods, but the combination of statistical approaches and traditional methods is necessary for us to gain new insight into old problems.

Keywords: historical linguistics, Australian languages, phylogenies, databases, comparative method, lexical change, sound change, phonology, Pama-Nyungan

DOI 10.1515/lingvan-2014-1009

1 Introduction

The twenty-first Century has been billed the era of “big data”, and linguists are participating in this trend. We are seeing an increased reliance on statistical and quantitative arguments in most fields of linguistics, including the oldest parts of the field, such as the study of language change. In prehistory and historical linguistics, much earlier work was based on detailed argumentation based on relatively few features. Arguments were qualitative rather than quantitative and judgment of the plausibility of such arguments often rested with other experts in the field rather than being explicitly quantified.

This has changed in the last five years, as statistical methods make increasing inroads into this field. This changes the types of questions we can ask, as well as how we evaluate the answers. But this all has prerequisites of certain types of data, coded in certain ways. We cannot make powerful statistical arguments from the qualitative data that historical linguists are used to working with, and new ways of building relationships have also led to new ways of building data sets. What counts as “big” data may vary from field to field, but common definitions include data sets that are too large to manipulate or analyze using standard techniques (Snijders et al. 2012).

In this paper I survey a few types of work based on a lexical database of Pama-Nyungan languages, the largest family in Aboriginal Australia (Wurm 1972). It highlights the flexibility with which large-scale databases can be deployed, especially when combined with traditional historical, philological methods. “Big” data may require new methods, but I argue that the combination of statistical approaches and traditional methods (from subsampling) are necessary for us to gain new insight into old problems.

***Corresponding author: Claire Bowern,** Yale University, New Haven, CT, USA, E-mail: claire.bowern@yale.edu

2 The Australian lexical database

The Australian lexical database is a database of approximately 775,000 words from languages from across the country. Data were collected with funding from the USA's National Science Foundation from 2007 to 2013.[1] The database is authored in Filemaker Pro v11. Specifically, it is a set of nine linked relational databases containing information about lexical materials, language names and locations, linguistic classifications, references, reconstructions, phylogenetic codes, and the like. Figure 1 gives a brief, schematic overview of the data structure, where each cell in the figure is a relational database.

The database contains information from 1,623 "varieties" (dialects or doculects) from 399 languages. There are 774,607 lexical items and data from over 2,000 references. The lexical materials include the original orthographic representation of the lexical item, a phonemicization in a standard orthography, gloss information, a "standardized"[2] gloss, part of speech information, and other information that was provided in the original source. Each entry also has a series of links to sub-databases (for example, the phylogenetic coding database and reconstructions database). As reconstructions proceed, entries are associated with a cognate-set id, which links to the reconstructions database. All materials are fully sourced. The source database includes bibliography (or archival) details of the source, access restrictions, and status of processing.

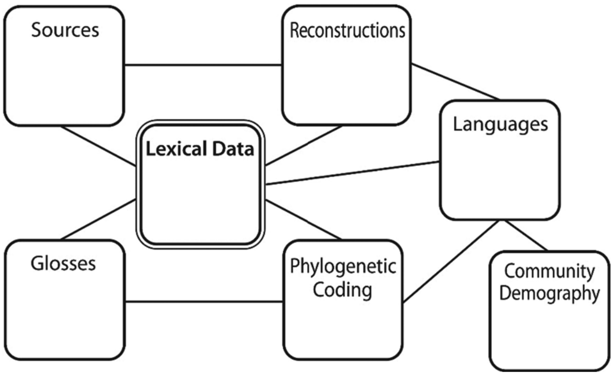


Figure 1 Schema of databases

The language data were compiled from published sources, unpublished fieldnotes and files from archives and personal collections, from the holdings of the (now defunct) Aboriginal Studies Electronic Data Archive, and from websites. Data collection was originally confined to Pama-Nyungan languages but was subsequently extended to the bordering non-Pama-Nyungan languages of the Nyulnyulan, Worrorran, Bunuban, and Maningrida families (Evans 2005). As time and resources permitted, more data were included from other Non-Pama-Nyungan families. Data collection initially focused on converting digital sources from various formats to a standard format for input into the database. Data were also entered by hand by undergraduate students from print materials (both published and unpublished). At this point, holdings are

[1] Data were collected under grant BCS-0844550 "Pama-Nyungan and the prehistory of Australia". Further data collection is currently in progress under NSF grant BCS-1423711.

[2] Original dictionaries differ in whether they gloss words using English infinitives, progressives, or headwords. This field abstracts away from these differences to facilitate searching. It also disambiguates English homophones (e.g. "bank") and zero-derived English word pairs which belong to different parts of speech ("bat"), and amalgamated items which refer to the same concept but with different original glosses. For example, some dictionaries give scientific names for flora and fauna, while others give only the common English name or a local slang name.

comprehensive for some Pama-Nyungan subgroups but are somewhat patchy for other areas, especially those which are heavily reliant on handwritten fieldnotes in non-standard orthographies.

The original aim in collecting materials was to facilitate lexical comparisons for the Comparative Method (Fox 1995; Rankin 2008; Hock and Joseph 2009). However, as data collection proceeded, it became clear that the database had other utility as well, for projects well beyond those for which it was originally intended. Here I describe several of those projects. The reader is referred to Bowern (2010a, 2012) amongst others for other applications, and to Bowern (2010b) for a more detailed description of the database itself.

3 Sound symbolism and the structure of phonological systems

One obvious application of a lexical database is the study of patterns in phonological systems. Australian languages are famous for their near-uniform phonemic inventories (Busby 1980; Dixon 1980; Hamilton 1995; Tabain and Butcher 1999).[3] The apparent uniformity of Australian languages also stands out in worldwide typological surveys when Australian languages are compared to other families and regions (Mielke 2008; Hunley et al. 2012). Otherwise unqualified statements about uniformity in inventory and phonotactics across the continent are easily found in reference grammars of languages in the region (for an example, see Goddard 1985: 21, 43, 66, 323). This assumption should be surprising given that there is no general assumption in phonology that associates inventory size or composition with phonotactic generalizations such as syllable structure constraints or segment frequencies, and indeed, we know that phonotactic constraints vary widely and crosscut inventory composition. Such uniformity in Australian languages, if real, is therefore surprising and unusual, especially given the country’s phylogenetic diversity.

Gasser and Bowern (2014) show the utility of deriving information about the phonologies of Australian languages *directly from lexical data*. Some phonological information is hard to glean from summary statements in reference grammars. For example, unless a frequency study was included in the grammar, there is no information about the relative frequencies of segments for individual languages. Moreover, reference grammars do not contain uniform information; for example, some exhaustively list the clusters found in the language, while others give summary statements by place and/or manner of articulation, while others list only the most common clusters. This makes systematic comparison across languages almost impossible.

However, luckily, Australianists have tended to use practical orthographies for writing the languages which map the phoneme inventories of those languages fairly directly. While data from pioneer and other early sources is not usually transcribed consistently, later work regularly makes use of phonemically-based orthographies to represent words. Wordlists were converted to a single set of standard symbols and 145 languages in the database from across Australia were compared for phonological inventory statistics, mean word length, and positional effects such as phonological contrast collapse versus maintenance in initial and final position. Generalizations were then extracted from the lists with a set of Python scripts which counted the phonemes, natural classes, and clusters, in the relevant positions in the lists and returned statistics for each language and overall throughout the set.

The results in that paper confirmed some generalizations about Australian languages from the previous but found many other exceptions, as well as some new generalizations. Particularly important were “minority” patterns in the data, which appear to be systematically overlooked in Australian phonological typologies. These are features which are not found continent-wide, but nonetheless are not restricted to a small sample of languages. For example, glottal stops or glottalized consonants are found in 32% of the languages in the sample; not a majority pattern by any means, but far more frequent than one might expect given Hendrie’s (1981) claim that it is “rare” (see Figure 2). Another is phonemic voicing of obstruents, found in 34% of the languages of our survey, revealing the inaccuracy of claims by authors such as Miceli

[3] The background information in this section is closely based on Gasser and Bowern (2014).

(2014:716) and many others that phonemic voicing is largely absent, very rare, or confined only to the languages of the northeast in Australia (see Figure 3). Such patterns are, of course, the core of language variation and important for studying the lack of uniformity across the continent. Other features with similar patterns include phonemic voicing and languages with more than three vowels.

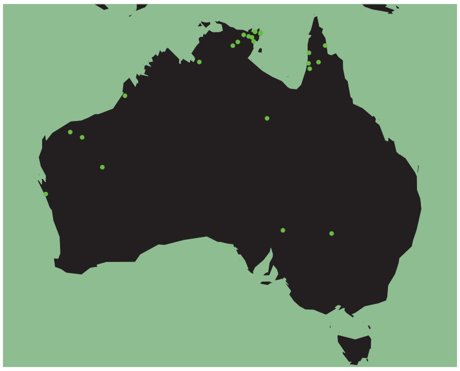


Figure 2 Distribution of glottal stops

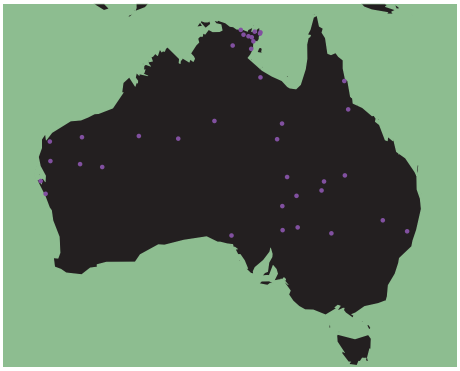


Figure 3 Languages with voicing distinctions

The study also allowed us to provide insight into markedness patterns, as well as questioning several claims in the literature. For example, we have numerous clear counterexamples to claims that typological markedness mirrors intra-language frequency (Greenberg 1966; Paradis and Prunet 1991). For example, for languages which have both /g/ and /k/, /g/ occurs more frequently; and for languages with a lamino-dental vs. apical distinction, the lamino-dental stop is more frequent. This study has convinced us that *all* phonological generalizations about Australian languages require a reevaluation, and that a dataset such as this is the best way to conduct that reevaluation.

4 Large-scale phylogenetics

The recent boost to quantificational work in historical linguistics, particularly computational phylogenetics (Atkinson and Gray 2005; Bouckaert et al. 2012; Holden 2002; Dunn et al. 2011; Gray et al. 2009), has made it possible to use datasets such as the Australian lexical database to infer phylogenies and reconstruct traits to ancestral nodes. Furthermore, beyond their use in estimating historical states these methods allow us to test theories about how language changes. A second area of work in this project is the targeted use of materials for reconstruction and computational analysis.

Work in Australian historical linguistics has long been held back by the difficulties in compiling a family tree for Pama-Nyungan. The relative paucity of distinctive sound changes (compared to some other parts of the world; see Miceli 2014) combined with extensive data in non-phonemic (or more accurately, non-regular) orthographies and a large number of languages with relatively scant descriptive materials has meant that progress has been slow in constructing a Pama-Nyungan tree. While there have been several classifications (for example, O'Grady et al. 1966; Wurm 1972), they list the major subgroups, without providing a detailed picture about how the more than 28 identified subgroups might fit together. This has given Pama-Nyungan a "rake-like" appearance, with large numbers of primary groups and few indications of higher structure in the tree.

Bowern and Atkinson (2012) had the goal of testing the "rake" hypothesis computationally, by testing models of the Pama-Nyungan family tree using basic vocabulary. A wordlist of 189 items was coded for cognate items in 194 languages. Chance resemblances were excluded where possible. This character matrix then forms the input to a Bayesian analysis whereby hypotheses of relationship are evaluated for how well they fit the data.[4] The resulting tree (in Figure 4) is a summary of the relationships hypothesized with the

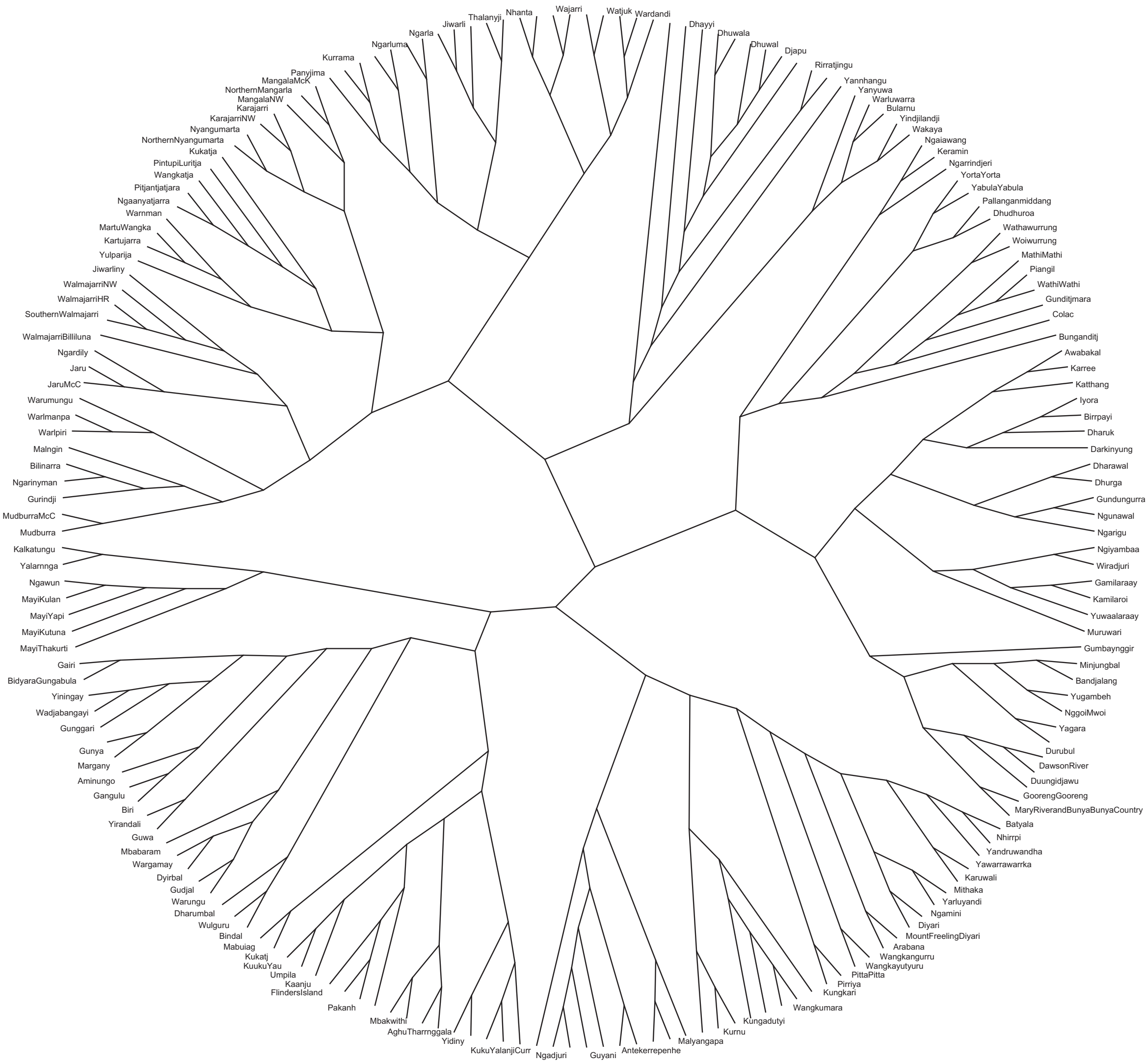


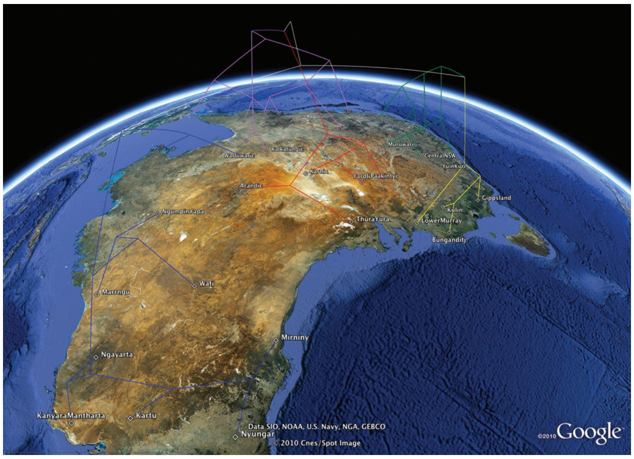
Figure 4 Phylogeny of Pama-Nyungan, redrawn from Bowern and Atkinson (2012)

[4] For further discussion of these methods, see Dunn (2014), Dunn et al. (2005) and Gray et al. (2007a, 2007b).

model, along with their degree of support. Such work is a good example of the way in which phylogenetics can be combined with reconstruction using the comparative method; the comparative method identifies the cognates that the languages share; the Bayesian phylogenetic component allows us to statistically evaluate different classification hypotheses.

Once a hypothesized phylogeny is in place, it can be further refined, as well as used to elucidate relationships between languages in the family not previously covered. In work currently in progress, I extend the phylogeny in Bowern and Atkinson (2012) to identify the most likely relationships of previously unidentified wordlists. A number of the wordlists in Curr (1886), for example, cannot be assigned with confidence to specific languages or subgroups. I use the existing cognate matrix to code items for likely cognacy and use the phylogenetic models to assign the wordlists to subgroups.

Phylogenies can also be mapped into geography to study areal patterns. An example from Pama-Nyungan is given in Figure 5. In this case, subgroups were mapped to their geographic centroids using the programs Mesquite[5] and Google Earth[6] and primary branches were color coded. Particularly obvious here is the way in which the languages of the Western branch (in blue) and northern New South Wales (in green) show splits suggestive of directed migration. In the former case, the migration is westward and then southward, which in the second, migration appears to be northward along the coast.



Data SIO, NOAA, U.S. Navy, NGA, GEBCO
© 2010 Cnes/Spot Image
©2010 Google

Figure 5 Geocoded phylogeny of Pama-Nyungan

Finally, the phylogeny can be studied for other properties, such as the differences in rates of change along various branches and the languages which are relatively innovative or conservative. Figure 6 colors languages by cognate retention; that is, the number of items they show in the basic vocabulary wordlist which have Proto-Pama-Nyungan etyma. Note that the number of languages with extensive etyma of Pama-Nyungan age is small, but the most conservative languages (those in green in Figure 6) are found in several major subgroups. Languages with particularly low retention rates include those from Paman (in the north-east), Victoria (in the southeast), and Yolngu (an enclave of Pama-Nyungan languages surrounded by Non-Pama-Nyungan families).

[5] www.mesquite.org
[6] earth.google.com

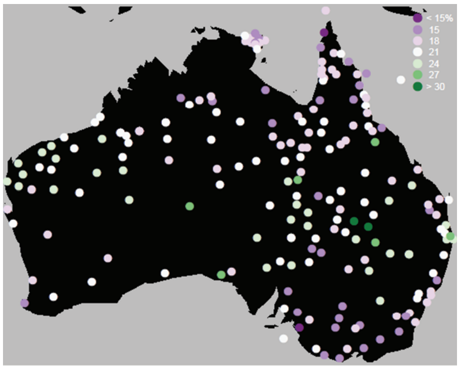


Figure 6 Retention rates across Pama-Nyungan

5 Results based on phylogenetic comparisons

Once there is a lexical phylogenetic tree, this opens up possibilities for further analysis of linguistic prehistory which relies on a prior hypothesis of linguistic relationship. One line of work uses a phylogeny developed from basic vocabulary (or other lexical items) to test hypotheses about processes of language change in other domains. I illustrate this here using data from kinship systems. We assume from some previous work (Dumont 1953; Friedrich 1966, for example) that kinship terms are part of a language’s “basic vocabulary”. However, they are not culturally universal, in that systems differ (though finitely; see Murdock 1968), and anecdotal work shows that not all parts of the kinship system are equally stable. Using tools such as the software program *BayesTraits* (Pagel and Meade 2004; Pagel et al. 2004) allows us to construct models of how a system has likely evolved through a phylogeny, and to compare that with reconstructions of the lexical items for elements in the system using the Comparative Method. That is, we use the language units as proxies for cultural groups and map the cultural evolution to the linguistic change (for further discussion of this practice, see Towner et al. 2012).

We provide preliminary results in this area in Bowern et al. (2014). In this study, we map sibling traits (such as the presence or absence of distinct terms for “older” and “younger” siblings) and reconstruct the sibling system to Proto-Pama-Nyungan. We find that two systems predominate; a three-term system where older brother and sister are distinguished but there is a single “younger sibling” term, and a four-term system where both relative age and the sex of the referent are distinguished. The former system is most common in Western languages, while the four-term system predominates in the East. While we can reconstruct systems to a fair degree of certainty to the major branches of the family, reconstructions to the root only weakly favor a 4-term system.

However, there is considerably more stability and resolution of the systems when considered as systems, in comparison to the terminology used to refer to sibling terms. We find extensive evidence for semantic shift, from sources both within the kinship system (for example, from grandparent terms changing to refer to sibling categories) and from outside.

Through this work we not only uncover information about the likely nature of ancestor languages and the overall patterns of evolution that have shaped Pama-Nyungan, but we also gain valuable feedback about hypothesized mechanisms and constraints involved in various types of language change. By using phylogenetic models to investigate change in these domains we gain a more complete picture of how this language family has evolved, as well as a means by which to answer several important questions about language change. In particular, we are interested in the relative stability and independence of various linguistic systems in this historical context, the apparent constraints on evolution of these systems, and what these facts tell us about the actual processes by which linguistic features change.

6 Conclusions

Once a database has been developed, there are many potential applications. As long as the underlying database is flexibly structured, so that data may be repurposed, it is possible to use comparative lexicographic data for purposes well beyond that for which it was originally collected. Lexical databases can be used to study features both of individual languages and subgroups, as well as regional trends. Data can be extracted for both reconstructions using the Comparative Method and inferences about language families using computational methods. Once a phylogenetic tree is compiled, it can be used to further study relationship between languages and sources, and between linguistic and cultural features. Such methods are likely to be particularly useful for families where there has, thus far, been little work. Important here, however, is the *interplay* between the minutiae of the individual comparisons and the broad-scale comparative work, for without that, it is difficult to evaluate the plausibility of hypothesized evolutionary models.

References

Atkinson, Quentin D. & R. D. Gray. 2005. Curious parallels and curious connections–phylogenetic thinking in biology and historical linguistics. *Systematic Biology* 54(4). 513–26.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, & Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097). 957–960. doi:10.1126/science.1219669 (24 August 2012).

Bowern, Claire. 2010a. Historical linguistics in Australia: Trees, networks and their implications. *Transactions of the Philosophical Society B* 365(1559). 3845–3854.

Bowern, Claire. 2010b. *Database of Pama-Nyungan Languages*. ms: Yale University.

Bowern, Claire. 2012. The riddle of Tasmanian languages. *Proceedings of the Royal Society B: Biological Sciences*. doi:10.1098/rspb.2012.1842. http://rspb.royalsocietypublishing.org/content/early/2012/09/22/rspb.2012.1842 (21 December 2012).

Bowern, Claire & Quentin Atkinson. 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* 88(4). 817–845.

Bowern, Claire, Hannah Haynie & Amalia Skilton. 2014. Lexical Stability and Kinship Patterns in Australian Languages. *LSA Annual Winter Meeting*. Minneapolis, MN. (2–5 January).

Busby, P. A. 1980. The distribution of phonemes in Australian Aboriginal languages. *Papers in Australian Linguistics 4, Pacific Linguistics A-60*, 73–139.

Curr, E. M. 1886. *The Australian race: Its origin, languages, customs, place of landing in Australia and the routes by which it spread itself over the continent*. Vol. 1. Melbourne (Australia): J. Ferres.

Dixon, R. M. W. 1980. *The languages of Australia* (Cambridge Language Surveys). Cambridge: Cambridge University Press.

Dumont, Louis. 1953. The Dravidian Kinship terminology as an expression of marriage. *Man* 53. 34–39. (28 May 2014).

Dunn, Michael. 2014. Language Phylogenies. In Claire Bowern and Bethwyn Evans (eds.), *Routledge Handbook of Historical Linguistics*, 190–211. Abingdon and New York: Routledge.

Dunn, M., S. J. Greenhill, S. C. Levinson, & R. D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473(7345). 79–82.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley & Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309(5743). 2072–2075. doi:10.1126/science.1114615 (27 March 2009).

Evans, Nicholas. 2005. Introduction. In *The non-Pama-Nyungan languages of northern Australia: comparative studies of the continent’s most linguistically complex region*, 1–25. Canberra, ACT: Pacific Linguistics.

Fox, A. 1995. *Linguistic reconstruction: An introduction to theory and method*. Oxford: Oxford University Press.

Friedrich, Paul. 1966. Proto-Indo-European Kinship. *Ethnology* 5(1). 1–36.

Gasser, Emily & Claire Bowern. 2014. Revisiting Phonological Generalizations in Australian Languages. *Proceedings of the Annual Meetings on Phonology* 2013.

Goddard, C. 1985. *A grammar of Yankunytjatjara*. The Gap, NT: Institute for Aboriginal Development.

Gray, R. D., A. J. Drummond & S. J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323(5913). 479–483. (6 August 2012).

Greenhill, S.J. & Gray, R.D. (2009) Austronesian language phylogenies: myths and misconceptions about Bayesian computational methods. In A. Adelaar & A. Pawley (eds.), *Austronesian historical linguistics and culture history: a festschrift for Robert Blust*. Canberra: Pacific Linguistics.

Gray, R. D., S. J Greenhill & R. M. Ross 2007b. The pleasures and perils of Darwinizing culture (with Phylogenies). *Biological Theory* 2(4). 360–375.

Greenberg, Joseph H. 1966. *Language universals: With special reference to feature hierarchies*. The Hague: Walter de Gruyter. http://books.google.com/books?hl=en&lr=&id=_OvPF_bVHmYC&oi=fnd&pg=PR7&dq=language + universals + special + reference + feature&ots=J040e1_BWg&sig=zoXrpoYzCGWIbGoGhhcB_EB3nvk (accessed 28 May 2014).

Hamilton, Philip. 1995. Vowel phonotactic positions in Australian aboriginal languages. *Proceedings of the Twenty-First Annual Meeting of the Berkeley Linguistics Society* 1995, 129–140.

Hendrie, T. R. 1981. Distinctive features matching as a basis for finding cognates. *Working Papers of the Linguistics Circle* 1(1). 32–41. (28 May, 2014).

Hock, H. H. & B. D. Joseph. 2009. *Language history, language change, and language relationship*. *An introduction to historical and comparative linguistics*. Berlin & New York: Mouton de Gruyter.

Holden, Clare Janaki. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 269(1493). 793–799. doi:10.1098/rspb.2002.1955 (19 December, 2012).

Hunley, Keith, Claire Bowern & Meaghan Healy. 2012. Rejection of a serial founder effects model of genetic and linguistic coevolution. *Proceedings of the Royal Society B: Biological Sciences* 279(1736). 2281–2288. doi:10.1098/rspb.2011.2296.

Miceli, Luisa. 2014. Pama-Nyungan. In Claire Bowern and Bethwyn Evans (eds.), *Routledge Handbook of Historical Linguistics*, 704–725. Abingdon and New York: Routledge.

Mielke, J. 2008. *The emergence of distinctive features*. New York: Oxford University Press.

Murdock, George Peter. 1968. Patterns of sibling terminology. *Ethnology* 7(1). 1–24. (29 October, 2009).

O'Grady, Geoffrey N., C. F. Voegelin & F. M. Voegelin. 1966. Languages of the world: Indo-Pacific fascicle six. *Anthropological Linguistics* 8(2). 1–197.

Pagel, Mark & Andrew Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology* 53(4). 571–581. (27 May, 2014).

Pagel, Mark, Andrew Meade & Daniel Barker. 2004. Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology* 53(5). 673–684. (27 May, 2014).

Paradis, Carole & Jean-Francois Prunet. 1991. Introduction: Asymmetry and visibility in consonant articulations. In Paradis and Prunet (eds.), *The special status of coronals: Internal and external evidence*, 1–28. New York: Academic Press.

Rankin, Robert L. 2008. The comparative method. In Brian D. Joseph & Richard D. Janda (eds.), *The handbook of historical linguistics*, 199–212. Blackwell. http://onlinelibrary.wiley.com/doi/10.1002/9780470756393.ch1/summary (accessed 21 March 2013).

Snijders, Chris, Uwe Matzat & Ulf-Dietrich Reips. 2012. Big data: Big gaps of knowledge in the field of internet science. *International Journal of Internet Science* 7(1). 1–5. (23 September, 2014).

Tabain, Marija & Andrew Butcher. 1999. Stop consonants in Yanyuwa and Yindjibarndi: Locus equation data. *Journal of Phonetics* 27(4). 333–357. doi:10.1006/jpho.1999.0099 (7 November, 2009).

Towner, M. C., M. N. Grote, J. Venti & M. Borgerhoff Mulder. 2012. Cultural macroevolution on neighbor graphs. *Human Nature* 23. 1–23. (16 December, 2012).

Wurm, S. A. 1972. *Languages of Australia and Tasmania*. The Hague: Mouton.