

Chirila: Contemporary and Historical Resources for the Indigenous Languages of Australia

Claire Bower
Yale University

Here I present the background to, and a description of, a newly developed database of historical and contemporary lexical data for Australian languages (Chirila), concentrating on the Pama-Nyungan family (the largest family in the country). While the database was initially developed in order to facilitate research on cognate words and reconstructions, it has had many uses beyond its original purpose, in synchronic theoretical linguistics, language documentation, and language reclamation. Creating a multi-audience database of this type has been challenging, however. Some of the challenges stemmed from success: as the size of the database grew, the original data structure became unwieldy. Other challenges grew from the difficulties in anticipating future needs, in keeping track of materials, and in coping with diverse input formats for so many highly endangered languages.

In this paper I document the structure of the database, provide an overview of its uses (both in diachronic and synchronic research), and discuss some of the issues that have arisen during the project and choices that needed to be made as the database was created, compiled, curated, and shared. I address here the major problems that arise with linguistic data, particularly databases created for diverse audiences, from diverse data, with little infrastructure support.

1. Introduction¹ Here I present the background to, and a description of, a newly developed database of historical and contemporary lexical data for Australian languages, concentrating on the Pama-Nyungan family (the largest family in the country). It is named Chirila, an acronym for *C*ontemporary and *H*istorical *R*esources for the *I*ndigenous *L*anguages of *A*ustralia.² The database was initially developed in 2007 in order to facilitate research on cognate words and reconstructions (thereby shedding light on Australian prehistory and Pama-Nyungan historical linguistics); however, it has had many uses beyond its original purpose, in synchronic theoretical linguistics, language documentation, and language reclamation. Moreover, a

¹Many people have been involved in the creation of this database. Portions of the database will be available from pamanyungan.net in fall, 2015 (and released regularly as authors give permission). The database has been funded by the National Science Foundation's Linguistics program within the Behavioral and Cognitive Science Directorate, first through NSF BCS grant 0844550 (2007–2013) and HSD-0902114, and most recently from BCS-1423711 (2014–2017). Contributors to the database are listed at pamanyungan.net/chirila, and their help with all aspects of this work is gratefully acknowledged. Thanks also to the contributors to an online discussion at academia.edu, who provided many helpful suggestions on an earlier version of this paper.

²The homophonous *tyirilya* is also a term for 'echidna' in a number of languages of the Western Desert region.

diachronic database has many uses beyond studying language relationships. Creating a multi-audience database of this type has been challenging, however. Some of the challenges stemmed from success: as the size of the database grew, the original data structure became unwieldy. Other challenges grew from the difficulties in anticipating future needs, in keeping track of materials, and in coping with diverse input formats for so many highly endangered languages.

In this paper I document the structure of the database, provide an overview of its uses (both in diachronic and synchronic study), and discuss some of the issues that have arisen during the project and choices that needed to be made as the database was created, compiled, curated, and shared. In doing so I add to the growing number of descriptions of large datasets (Thieberger 2011, Greenhill et al. 2008). I address here the major problems that arise with linguistic data, particularly databases created for diverse audiences, from diverse data, with little infrastructure support. Extracts of the database itself are available from pamanyungan.net/chirila.

2. Overview and aims of the database Linguists have long stored their data in structured ways. Consider the file card systems (cf. Thieberger & Berez 2012) that are still occasionally in use (Dixon 2007). The extensive reduction in cost of computer technology, particularly over the last ten to fifteen years has allowed for expansion in this area. There is also more work that relies on machine readable, structured data. Almost all work on corpus linguistics, for example, relies on computer files with structured format. Within historical linguistics, structured comparative databases are the norm for both small and large families, from the Austronesian Basic Vocabulary Database (ABVD) (Greenhill et al. 2008) and Algonquian (Hewson 1993) and Siouan (Rankin et al. 2015) work, to other online databases such as CBOLD (Comparative Bantu Online Database),³ RefLex,⁴ the Database of Arabic Dialects,⁵ and STEDT (Sino-Tibetan Etymological Dictionary and Thesaurus).⁶

Structured comparative lexical data is particularly needed for Australian languages and historical work. Prior to this database, there was no comparable source for Australian languages, either in print or online. Lists of cognates can be found in some publications by O'Grady (1990a, 1990b, 1998) and students, and most completely in Alpher (2004). The now defunct Aboriginal Studies Electronic Data Archive (ASEDA) was a repository of digital data for Australian languages, but the files were in many different formats and there was no linking between languages.⁷ Likewise, the Living Archive of Aboriginal Languages (cdu.edu.au/laal) has digital data, but not comparative data, and its scope is the Northern Territory, not the whole of Australia. There are few lists of reconstructed forms, either at the level of Proto-Pama-Nyungan or individual subgroups (though see, amongst others, Alpher et al. 2008; Koch 1997;

³<http://www.cbold.ish-lyon.cnrs.fr/>

⁴<http://reflex.cnrs.fr/>

⁵<http://database-of-arabic-dialects.org/>

⁶<http://stedt.berkeley.edu>

⁷Some, but not all, files formerly hosted by ASEDA are now available through the catalogue of the Australian Institute for Aboriginal and Torres Strait Islander Studies (mura.aiatsis.gov.au), as AILEC. A list of items in the catalogue can be found at aseda.aiatsis.gov.au.

Hale 1976; Black 1980). And finally, until Bown and Atkinson (2012), there was no good reference tree for the languages, particularly for Pama-Nyungan. This made it difficult to distinguish likely true cognates from loans or lookalikes.

To work efficiently and systematically across multiple languages, data needs to be standardized. For example, we want to know which words belong to which language, but different authors spell language names in different ways. For example, Austin (1981, 1990) spells one of the languages as Diyari, but earlier works spell it Dieri. There are circumstances where we want to know exactly how the author referred to the language name, but for the most part, we will want to refer to both ‘doculects’ (to use the term of Good & Cysouw 2013) by the same name. Likewise, for the most part, we will want to analyze data using a single transcription system. But we do not want to lose the information about how the author of the original work transcribed the language, both because that may be a community-preferred way of representing the language, and because the original orthography may itself be the subject of research.⁸ We operate with the principle of being able to recover the information structure of the original source while still being able to standardize the materials so that they can be adequately compared.

The database contains lexical information from languages belonging to both the Pama-Nyungan family and many of the Northern (non-Pama-Nyungan) families. That is, it comprises information about words (and not morphemes, sound systems, or grammatical features). This is because there are many languages recorded in Australia from early sources where morphological information is scarce.⁹

The data were gathered and the database was structured with several distinct research aims in mind. One was COGNATE IDENTIFICATION AND RECONSTRUCTION. That is, we wish to identify recurring correspondences in sounds between languages. From these systematic patterns, we can determine the likely forms of words at earlier stages of the languages; by mapping the sound changes and distribution of words we can do language classification; and from the words which deviate from these patterns we can identify likely loanwords. Thus, the database contains not only synchronic word forms, but also putative reconstructions of proto-language forms.

A second aim was to map the DISTRIBUTION OF LEXICAL ITEMS across the family. This was first done in a preliminary way by O’Grady and colleagues (e.g., O’Grady 1990b), following work by Arthur Capell (e.g., Capell 1956) which mapped the percentage of common Australian items in particular areas of the country.¹⁰ We can use

⁸For example, in work in progress, we are using the orthographic conventions of non-linguists recording languages in Eastern Australia to evaluate the likely meanings of orthographic representations of Tasmanian languages.

⁹Perhaps ironically, there is considerably more comparative work in Australian languages on morphological reconstruction than on lexical reconstruction. This in part reflects the common idea (e.g., Hymes 1956) that morphology is more stable than lexicon. But morphemes cannot be judged as cognate or not without a good understanding of the sound correspondences exhibited between the related languages, and this is gained from lexical work (see also Bown 2012a). Moreover, the low levels of lexical cognacy across the continent make widespread reconstruction difficult.

¹⁰Note, however, that Capell’s identification of common items was flawed in that he chose items which occurred in three different states. It is not surprising, therefore, that the subgroup with the highest degree of retentions (Wati) is one which straddles the borders of Western Australia, South Australia, and the Northern Territory.

a database like this to study patterns of loanwords, cognate retentions, vocabulary stability, and variation in rates of change. For example, there have been claims in the literature that Australian languages are particularly susceptible to borrowing, and that the basic vocabulary is not particularly stable, unlike in other areas of the world (Dixon 1997). A case study of this type was published in Bower et al. (2011).

In addition to the mapping of the distribution of lexical items and the reconstruction of language history, a further aim of the database was to test hypotheses regarding language CLASSIFICATION AND SUBGROUPING. Australian linguistics is a small field and very few people are experts in more than one area of the country. Moreover, there are not many areas where more than one person has worked in detail on languages in an area, especially for historical purposes. Moreover, in some areas where there has been work by more than one person, there are disagreements. Although there is a fair amount of agreement as to the composition of the major subgroups of Pama-Nyungan, there is no consensus as to how those groups might be related to one another, or, in many cases, what the internal structure of those groups is. An overview for one area of Australia is given in Bower (2009; 2010). With an extensive database of this type it is possible to evaluate claims with some degree of objectivity, and to quantify the sources of disagreement.

Finally, a database such as this is instrumental in developing THEORIES OF LANGUAGE CHANGE. Previous work has shown some of the ways in which generalizations about Australian languages have been inadequate: for example, the role of language contact (Bower & Atkinson 2012; Bower et al. 2011; Hunter et al. 2011), or the uniformity of Australian typological profiles (Bower, Brody & Killian 2014; Gasser & Bower 2014). In other cases, the previous work allows us to ask some questions for the first time. Having a well-articulated proposal for Pama-Nyungan language relationships (Bower & Atkinson 2012) allows us to investigate relationships between language change and space and study the evolution of linguistic subsystems (cf. Haynie et al. 2014). Finally, we need to (re)evaluate the use of phylogenetic methods and what they can tell us about language history. We do this by checking the results of our methods against plausible theories of change (cf. Bower & Evans 2015), and by refining the computational models used to generate the analyses.

These aims are all academic. An additional, very important aim of a comparative database is to make a repository of dictionary data for language communities wishing to have digital data for their own needs, such as for language programs. A database like this is useful for communities who wish to use comparative evidence for their language reclamation programs. As Amery (2000) has shown, for example, reclamation programs can make use of comparative information on closely related languages to create words which are not attested from the historical sources on the language. Other groups may simply wish to have more accessible copies of manuscript materials on their language. For example, the Laves manuscript materials compiled in the late 1920s (Bower 2003) are copious and valuable records for more than half a dozen Australian languages, but Laves' handwriting is difficult to read and his glosses can be difficult to follow. Typed versions of these materials are far more useful than copies of the originals.

3. The data The Australian comparative database is a lexical database. That is, the unit of analysis is the phonological word. More precisely, the headword in the database is the citation form of the dictionaries, wordlists, and other manuscript materials that form the basis of the resources.¹¹ The database contains words from all over Australia, linked to language information. The lexical data are also grouped into cognate sets (that is, hypotheses of which words are historically related to words in other languages). The sources used for lexical information for Australian languages are extremely varied. They range from the importing of structured data from digital sources which are already in database formats to unstructured data in text files, to keyboarding of typescripts and manuscripts.

3.1 Orthography The sources represent a wide variety of orthographic conventions, which range from the International Phonetic Alphabet or another standard transcription system, to systems based on English spelling or created ad hoc by someone with no training in linguistics. While we wish to standardize the transcription conventions to allow for easy searching and easier cognate identification, we do not wish to lose the information of the original transcription for several reasons. First, several languages have established community writing systems and displaying the results of searches or quoting the word forms and publications would require the use of that system. After all, when forms are given for languages such as French and German, we write them in the customary orthography, and not in some generalized transcription system.

Secondly, in many cases (especially regarding the 19th century sources recorded by non-linguists) we may not be certain what the accurate and unique transcription is. The English-based writing systems typically under-differentiate crucial distinctions, such as retroflexion. Therefore, while we wish to publicize and analyze a transcription in a standard form, subsequent new data or later analysis may shed new light on the correct interpretation of these forms. Therefore the database contains both a field for the transcription as it appears in the source, and a generalized, phonemicized transcription.

We provide two phonemicizations. The first is based on those symbols in standard use amongst Australianists. For ease of entry and reading, we prefer digraphs to IPA symbols, with the exception of engma /ŋ/, which we use instead of <ng> to avoid problems of representing the cluster of apical nasal plus voiced velar stop /ng/. We use a single <r> for the glide and a double <rr> for the trill, as is standard practice in Australia; for languages which also have a third rhotic, a tap, we use <ř>. We also provide a conversion to IPA, for those not familiar with Australianist practices. A list of characters and their IPA equivalents is given in Tables 1 and 2. Symbols in use in the project are given first, with IPA in // where it differs. In choosing codes, we had to balance well-known conventions within Australia with a need for single symbols across the country. For example, in some languages of the Cape York Peninsula, it is

¹¹In practice, of course, lexicographers make different decisions about what is a dictionary headword. Some use inflected forms; others use roots or stems. This heterogeneity in lexicographic decision making has caused problems for standardization; see further §6, particularly §6.3

customary to represent a central vowel as *v* (see, for example, Alpher 1991). However, this causes problems if we retain this convention and try to compare symbol *v* in Cape York with languages in Arnhem Land, where *v* is a labio-dental fricative. Prestopped nasals and laterals are treated as clusters.

Table 1. Consonantal standard orthography

Labial	Apico-dental	Lamino-dental	Apico-post-alveolar	Lamino-palatal	Velar	Glottal
b	d	dh /d̥/	rd /d̥/	dy /j/	g	
p	t	th /t̥/	rt /t̥/	ty /c/	k	ʔ
m	n	nh /n̥/	rn /n̥/	ny /ɲ/	ŋ	
v	ð					
β						
f	θ	s		sh /s/		
		rr /r/	r /ɹ/			
		ř /ɹ̥/				
w		y /j/	yh	h /ç/		
	l	lh /l̥/	rl /l̥/	ly /ɬ/		

Table 2. Standard vowel orthography

i, i:	u, u:
e, e:	o, o:
ɛ, e:	ə
	ɔ, ɔ:
	a, a:

3.2 Data sources The lexical data come from a wide variety of sources. The major holdings are described here. A significant data source has been the unrestricted holdings of the Aboriginal Studies Electronic Data Archive (ASEDA), curated by the Australian Institute of Aboriginal and Torres Strait Islander Studies. This digital archive was set up by Nicholas Thieberger in 1991 in order to preserve and distribute electronic data for indigenous Australian languages. Between 25 and 30 percent of the entries in the database have come from this source. Files were received in very many different formats and were converted to spreadsheets and then imported. Many of the files were in a basic markup format known as backslash coding, which is commonly used in linguistic dictionaries. It is, for example, the input to dictionaries created in Lexique Pro¹² and Toolbox.¹³ Backslash coded files are easy to convert as long as the files were consistently structured. Unfortunately, most of the files were either inconsistently structured (with fields in different orders in different records) or contained complex internal structure, which made them difficult to import consistently. A student researcher on the database project, Sophia Gilman, wrote a script

¹²lexiquepro.com

¹³www.sil.org/computing/toolbox

in Python to standardize and import data in backslash code format.¹⁴ Further electronic files were given to the author by the researchers themselves. To date we have not made use of Optical Character Recognition (OCR) software in digitizing sources to any extent, because in trials we found that the amount of post-processing that was required meant that there was no saving in time over typing materials directly into the database.

Another sizable source of forms is Curr's (1886) *Australian Race*.¹⁵ This is a series of wordlists from all over Australia. Wordlist length and quality vary substantially but it is an important early resource. Some languages (such as Wiradjuri) have several thousand words in Curr, and show both recorder depth (that is, reports from more than one source) and extensive numbers of unique vocabulary items. Other languages are attested only from Curr's standard item list of a few hundred words. Other early resources which have also been imported into the database include Roth (1897), Teichelmann & Schürmann (1840), Ridley (1875), and Dawson (1881), amongst others. The database thus has historical depth as well as extensive geographic coverage. A histogram of the publications in the database, by year, is given in Figure 1.

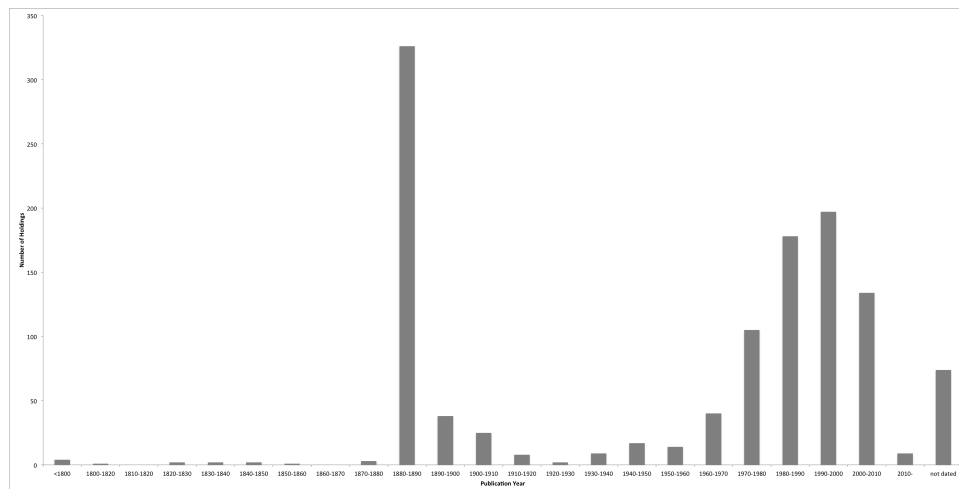


Figure 1. Publications in the Chirila database by year

Most of the other entries came from manual typing of materials where electronic files do not exist or were not available. Materials of this type were found in both published and unpublished sources. A major holding of unpublished (manuscript) sources has been the library of the Australian Institute for Aboriginal and Torres Strait Islander Studies. This library has also provided access to many published but hard-to-find works (so-called 'grey literature,' as Peter Austin¹⁶ has written about), particularly published by regional language centers. In the full database, the clear majority of sources are published (559, compared with 195 unpublished). However,

¹⁴The script is available for download from the author's web page.

¹⁵See further aiatsis.gov.au/collections/collections-online/digitised-collections/collectors-words/edward-micklethwaite-curr for information about the Curr materials and their geographic coverage.

¹⁶elar-archive.org/blog/language-documentations-grey-literature/

when we consider the amount of data holdings, the contribution of unpublished material is much greater: 301,851 records vs 342,386.¹⁷

An important source of reconstructions came from Alpher (2004), which was kindly provided to the author in backslash coded format. Other reconstructions came from the author's own work. Other data donations have come from linguists themselves, including (but not limited to) Peter Austin, Barry Blake, Gavan Breen, Luise Hercus, Alice Gaby, Maïa Ponsonnet, Jeffrey Heath, Peter Sutton, Geoffrey O'Grady, Kenneth Hale, Nicholas Evans, Barry Alpher, and Susan Hanson. Wangka Maya Pilbara Aboriginal Language Centre has been especially helpful in contributing materials from Western Australia. A full list of contributors is available at pamanyungan.net/chirila.

3.3 Language collection policies Others had created databases of Australian languages in the past, but they were of rather different format and scope. Norman Tindale had a set of file cards containing attestations of words from many Aboriginal languages from his fieldwork from the 1930s and later.¹⁸ He was hampered by lack of access to data. Geoffrey O'Grady and Kenneth Hale collected copious material and compiled it into a database. Others also have compiled their own comparative materials, especially for individual subgroups of Pama-Nyungan, or certain geographical areas such as Arnhem Land. The largest of these to my knowledge is the appendix to Alpher (2004), but many others are unpublished.

There are several ways to target data collection for a project such as this. Broadly, one could focus on quality or quantity. That is, a defensible approach would be to use only the most reliable data from the best attested languages, recognizing that this would constrain the scope of inquiry, but one would be more likely to be able to trust results of queries. The alternative approach, and one followed here, is that the primary value of the database lies in the comprehensiveness of its coverage. With a database, searches can always be filtered (e.g., by a tag that identifies the highest-quality sources). But many of the research questions we have investigated were prompted by the process of collecting data itself, rather than the reverse. To take one example among many, if we had been restrictive in the sampling of languages, rather than trying to collect data for all of them, we would not have been led to investigate the question of how many languages were spoken in Australia before European settlement, and why the figure from the database (397) is so much higher than traditional estimates of around 250 (cf. the estimates in Dixon 2002:240 and discussion in Walsh 1997).

Data collection for this project initially focused on the Pama-Nyungan family. This is the largest language family in Australia, covering 90% of the land mass and about two-thirds of the languages. It then expanded to the Non-Pama-Nyungan families adjacent to Pama-Nyungan, particularly the Nyulnyulan, Worroran, Maningrida and Garrwan families. This facilitated loanword comparison in Pama-Nyu-

¹⁷While the total database has 775,000 items, not all sources are currently categorized for publication status.

¹⁸These cards are now held in the South Australian Museum.

ngan ‘border’ areas. Tyler Lau, while an undergraduate researcher on this project, created a digital copy of Plomley’s (1976) database of Tasmanian languages.¹⁹ As part of an expanded data set, we are now aiming to collect data from any Australian language.

As of January 2016, 104 languages have been identified as being underrepresented in the database. There are three geographical areas where sources are substantially incomplete: Cape York Peninsula, where materials exist in handwritten field notes but have not yet been entered into the database, the Nyungar-speaking areas of the far Southwest, and the ‘Top End’ (Central and Coastal Northern Territory, and South-eastern Arnhem Land), where there are substantial numbers of Non-Pama-Nyungan languages. Languages of the Daly Region and the Mirndi, Tangkic, Alawan, and Gunwinyguan families (of Arnhem Land and surrounds) are underrepresented in the database currently. Materials exist in publications, archives, and personal collections. A third obvious data source to include are further early materials which supplement the language sources already entered, such as the wordlists of Richard Brough Smyth²⁰ and Daisy Bates. This would provide further historical depth to the data holdings.

3.4 Database statistics Data entry is still in progress. As of August, 2015, there are 775,814 lexical items in the database from 343 Pama-Nyungan languages, 56 Non-Pama-Nyungan languages, and the entire corpus of extant Tasmanian data (Bowern 2012b; Plomley 1976). I estimate that the database will be substantially complete at about 900,000 items. While I doubt that it will be possible to compile a database of every recorded word for every Pama-Nyungan language, it will be possible to include substantial data for all well attested languages, and most of the data recorded for the less well-known languages.²¹ Figure 2 gives the number of data point for each doculect in the database. Note the heavy skew towards doculects with small numbers of items (below 500).

A map showing sample locations is given below. Tasmanian languages are not represented due to difficulties in identifying sample locations. Locations are shaded by amount of available data. Currently, the database contains at least some data from every region and family in the continent. However, some areas are much better covered than others. The map in Figure 1 is approximately representative of the language density variation in Australia; that is, there are more language varieties in the far north and fewer in the Western Desert regions. Note that while Figure 2 above shows doculects, the map below uses standard language names (hence the higher counts of data per language unit).

¹⁹The Tasmanian database is currently self-standing and has not been merged into the main comparative Australian file, because of the special requirements for dealing with language data where the languages are mostly not identified, the phoneme inventories of the languages are unknown, and the glossing is erratic (see further Crowley & Dixon 1981, Bowern 2012b).

²⁰See <http://www.aiatsis.gov.au/collections/exhibitions/collectors/smyth.html>.

²¹Contact from researchers who are able to contribute data would be much appreciated!

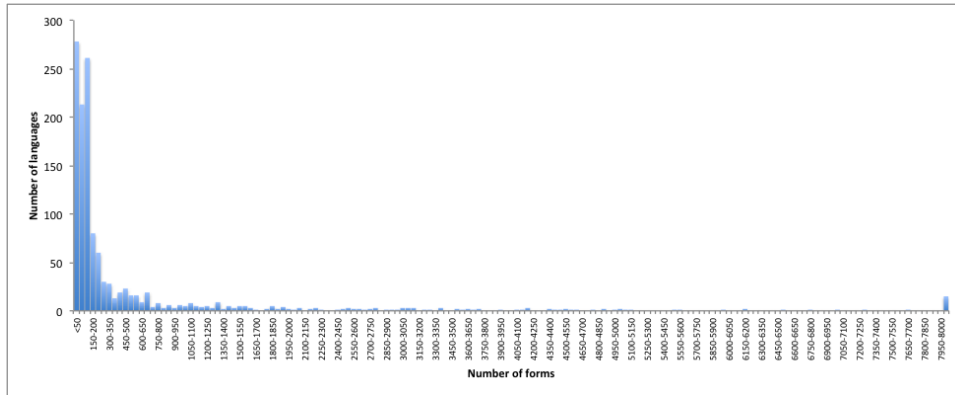


Figure 2. Chirila holdings, by language

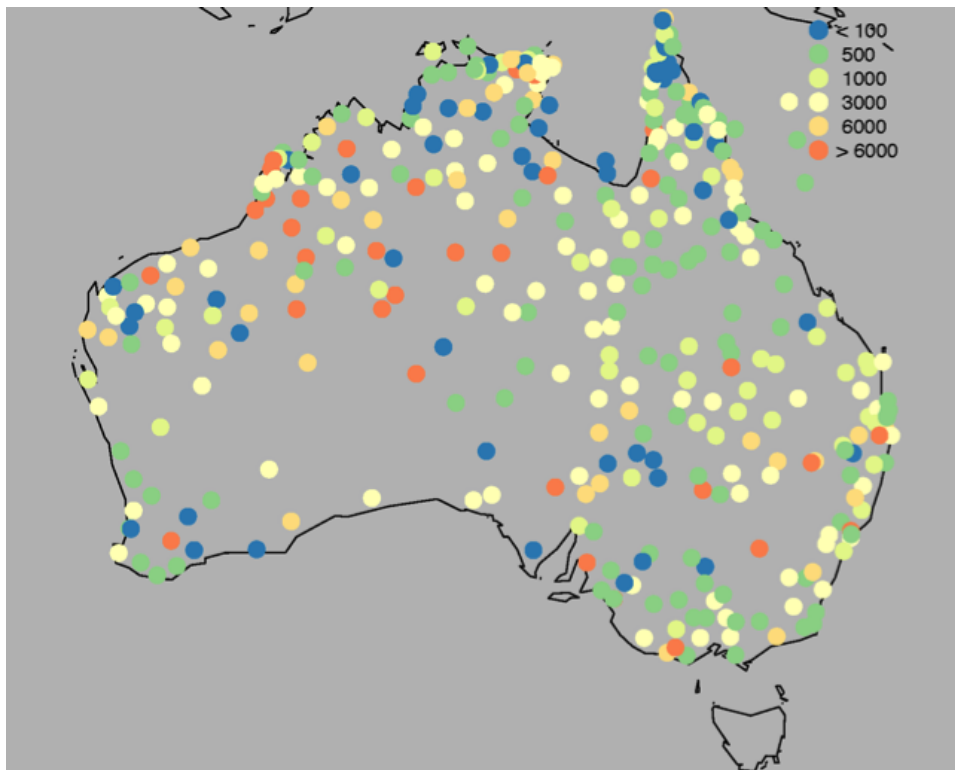


Figure 3. Counts of sources in Lexical Database

4. Structure of the database and software The Chirila database was developed in FileMaker Pro.²² What I have been calling ‘the database’ is in actuality a collection of linked relational sub-databases, with files for lexical data, reconstructions, glosses,

²²The original database was developed in early 2007 using version 8.1. The latest version of Filemaker Pro is 13.09. More on this below.

language data and analytical questions.²³ Some files (such as the sources and data files) are a single table, while others (such as the language file) are a set of relational tables which themselves link to other files in a relational manner. The largest file, the lexical data file, is approximately 400 MB. The total project size is about 750 MB and is hosted on a FileMaker Pro server operated by Yale University. The database is backed up hourly off-site; hourly backups are kept for 24 hours, and daily backups are kept for a week. The weekly backups are kept for 5 weeks, and full back-ups are periodically downloaded to allow for an additional location for secure file storage. The publicly accessible portion of the database is on pamanyungan.net, hosted by bluehost.com.

Data are stored in Unicode (UTF-8) format. We have tried to avoid the use of characters in private-use areas of the UTF-8 character database, although this has not always been possible due to the wide variety of orthographic representations in the different sources, particularly for the older sources. The database is password protected and hosted on a secure server with access restricted to users with accounts with Yale University's IT services. Users log in when opening the database and there is basic user tracking; that is, we can recover which records are modified in a session and who made the changes (but not what the changes are).

Figure 4 shows the core data structures. The following sections give a fuller list of the files and major fields in each table and sample screenshots. As will be obvious from the screenshots of the database in the following sections, graphic design has not been a high priority, and speed, function, and structural design have been prioritized over aesthetic design. We have also, to some extent, made use of Filemaker's capabilities for analyzing data, such as counting records with particular properties (such as the number of records associated with each particular language, the number of words tagged as 'loans' within a semantic field, and so forth).

4.1 Languages The Language information contains tables for core language names (that is, standardized names used in publications); mappings to variety (including dialect) names and alternative spellings; location and subgrouping information; and summary fields which give the resources for each language available in the database, such as the number of lexical items. This allows us to generate language lists with the best-attested languages. Table 1 below gives the language file's tables and major fields. Note that we follow best practice by giving field names without spaces. Throughout the database, each table has a primary key (unique numerical id). These are unmodifiable and auto-generated so that records are uniquely referenced, and tables and fields can be referenced across tables and files, and re-imported without overwriting the wrong record. These unique ID numbers are the basis for all join relationships in the database. Tables also contain 'housekeeping' fields such as the creation and last modification date/time, and the username of the creator/modifier. These fields are not listed in the description below but should be assumed to apply to all tables.

²³The data were originally stored in a single database file, but as more data became available, and as the data structures increased in complexity, a single file became too unwieldy, and a major re-write of the underlying database was completed in 2010.

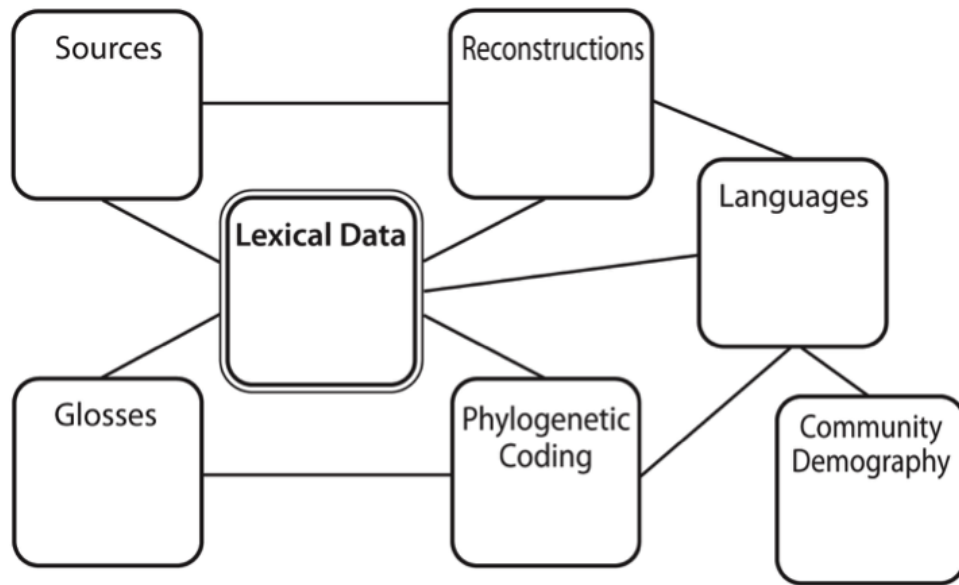


Figure 4. Core data structures for the Australian lexical database

Table 3. Tables in the Language Database

Table	Records ²⁴	Field	Description
Standard Names	671	StandardLanguage-Name	The reference spelling, the standard language name as used in linguistic sources
		AIATSIS_Code	The region-number code used in cataloguing at the Australian Institute of Aboriginal and Torres Strait Islander Studies
		Ascii_Name	Language name without spaces, hyphens, glottal marks or special characters (for use with programs for which those characters create problems)
		EtymologyofName	How the name is formed (whether, for example, it is reduplicated, or based on a phrase such as 'good' language)
		Family	The family to which the language belongs

Continued on next page

²⁴The number of records is given as of January 6, 2016.

Continued from previous page

Table	Records	Field	Description
		ISO-639	The ISO 639-3 code, if present (XXX if there is no code)
		Latitude	Centroid latitude for the language
		Longitude	Centroid longitude for the language
		MasterLanguageList	Name included on the ‘master’ list of 397 languages (considered when answering the question ‘how many <i>languages</i> were spoken in Australia?’)
		NumberofVarieties	Number of variety names associated with the language name
		NumberofEntries	Number of lexical entries associated with the language name
		Subgroup_ID	ID for the subgroup associated with the standard language name
		Include@...	A set of fields which state whether the language is included as a source for various comparative projects
Variety Names	2905	VarietyName	The name of the language variety (doculect)
		ISO_Code	The ISO 639-3 code, if present (XXX if there is no code)
		AIATSIS_Code	The region-number code used in cataloguing at the Australian Institute of Aboriginal and Torres Strait Islander Studies
		Glottolog_Code	Code used by glottolog.org
		DataPoints	Number of data points in the database
		Latitude	Centroid latitude
		Longitude	Centroid longitude
		Std_Lang_ID	ID for the standard language name associated with this variety
		StandardNotes	Notes on the assignment of this variety to the particular standard language

Continued on next page

Continued from previous page

Table	Records	Field	Description
Sub-groups	62	SubgroupName	Standardized name of the subgroup
		Latitude	Centroid latitude
		Longitude	Centroid longitude
		CountofLanguages	Calculation which gives the number of languages in the subgroup
		Family	The language family to which the subgroup belongs (for example, Pama-Nyungan, Nyulnyulan, Mirndi)

The data uses the concept of a ‘doculect’ as the object of study. The term doculect was coined by Jeff Good (cf. Good & Cysouw 2013) to refer to the form of a language as presented in a particular data source. It thus recognizes the fact that attestations of a given language may vary substantially depending on the time at which the source was recorded, the accuracy of the recorder(s), sociolinguistic considerations, the geographic location of the speaker, and so on. Such a model is particularly appropriate for the Australian data contained in this database because of the very wide variety of data preservation traditions, the degree of familiarity of the researchers with the languages, and the fluency of speakers at the time they were recorded. We also, however, need a mapping from the ‘doculect’ to the concept of a language, which is also useful in such a database. Typological sampling, for example, usually works at the level of a ‘language’ (see wals.info, for example) because samples taken from languages with many (minimally differing) dialects will skew generalizations. To take a simple example, if we were studying whether languages tend to have a dual pronoun, and we included in the sample 15 varieties of English (American English, British English, Australian English, etc.) and 3 Pama-Nyungan languages, we would likely conclude that dual is a rare category, even though the English samples are not independent.

Variety names are as the author gives them. They thus range from language names (e.g., ‘Diyari,’ ‘Bardi,’ ‘Yuwaaliyaay’) to ‘group’ names (‘Muliarra Tribe’) to locational information (Mary River and Bunya Bunya Country). Languages are geocoded. That is, all language (and variety) names are given latitude/longitude references, which allow data to be plotted on a map. The reference is the approximate ‘centroid’ of the language range: that is, a point at approximately the middle of the language range, as assumed at European settlement. All geocoding was done manually, and compiled from existing sources (mostly printed maps and text descriptions of traditional language locations). An example of this is given in §7.3 below. The georeferencing for both languages and words has in itself been a very useful outcome of the project. It is being published separately and the information has been used in

both works such as Hunley et al. (2012) and work in progress using the phylogeographic modeling tools of Bouckaert et al. (2012). A field tracks the etymology of the language name. This is not yet complete but is part of a research project to study naming practices across Pama-Nyungan languages.

There is some redundancy in the database. For example, both varieties and standard languages are geocoded. The field ‘Family’ is included both in the standard language table and the subgroup table. In a strictly relational database, this is redundant, since the information could be looked up through a join relation between the variety, its standard language name, and its subgroup. That is, in theory, the language family does not need to be included in the standard language table because it is inherited from the subgroup table. However, when creating views of the data for export, having many inherited join relationships of this type creates unwieldy data structures and slows down database performance. Therefore, the decision was made to have some redundancy, and instead to allow these fields to be looked up from their parent tables periodically.

Other items may appear to be redundant, but in fact are not so. For example, both varieties and standard languages have ISO 639-3 codes, because the ISO itself is inconsistent here. According to ISO 639-3’s principles, *languages* get codes, not varieties, but in practice there are cases both where multiple varieties have codes (and so a standard language reference could point to more than one ISO code) and where a code references a specific variety but not the standard language. Creation of the language dataset involved the submission of more than 100 requests to change, create, or merge ISO codes for Australia in 2012. All but two of those requests were accepted, and Chirila now uses the updated codes. The database also contains reference to the codes used by the Glottolog database (glottolog.org). These codes were matched to variety names and then generalized to the standard language names used here, since Glottolog’s primary reference name often uses slightly different spelling from that used in Chirila.

Most tables also have summary fields which count the records associated with that value in another database (such as the number of entries for each language, the number of languages in a subgroup, and so forth). Finally, there are fields which track whether a language is included in a particular sub-dataset. An example of this is the series of fields called ‘included@’ in the table above. Not all languages in the list have enough data to be included in all the analytical projects that have been undertaken with the dataset. An example of both of these types of fields is given in Figure 5.

Not all data are complete: in fact, there is still considerable work to be done. While most varieties are associated with a standard language, a subgroup, and a family, some vocabulary lists cannot be linked yet with a standard language. Out of 2910 varieties, 228 (or 7.8 percent) are currently unlinked. Some of these are unidentified lists, such as some vocabularies in the periodical *Science of Man*. Others come from compilations of names (such as the serialized “Aboriginal Names of Places in Western Australia,” also in *Science of Man*). Moreover, the assignment of languages to subgroups is itself the outcome of research based on the database, as is the association

ID	Name	ISO639-1	ISO639-2	Family	Subgroup	Status	Priority	Number	Count@lex	Count@lex	Include@P	Basic	Y	Priority	Grammar	Code	Count@lex	CheckPhone	Count@Basic
1114	Abanik																		
1115	Abanik																		
1116	Abanik																		
1117	Abanik																		
1118	Abanik																		
1119	Abanik																		
1120	Abanik																		
1121	Abanik																		
1122	Abanik																		
1123	Abanik																		
1124	Abanik																		
1125	Abanik																		
1126	Abanik																		
1127	Abanik																		
1128	Abanik																		
1129	Abanik																		
1130	Abanik																		
1131	Abanik																		
1132	Abanik																		
1133	Abanik																		
1134	Abanik																		
1135	Abanik																		
1136	Abanik																		
1137	Abanik																		
1138	Abanik																		
1139	Abanik																		
1140	Abanik																		
1141	Abanik																		
1142	Abanik																		
1143	Abanik																		
1144	Abanik																		
1145	Abanik																		
1146	Abanik																		
1147	Abanik																		
1148	Abanik																		
1149	Abanik																		
1150	Abanik																		
1151	Abanik																		
1152	Abanik																		
1153	Abanik																		
1154	Abanik																		
1155	Abanik																		
1156	Abanik																		
1157	Abanik																		
1158	Abanik																		
1159	Abanik																		
1160	Abanik																		
1161	Abanik																		
1162	Abanik																		
1163	Abanik																		
1164	Abanik																		
1165	Abanik																		
1166	Abanik																		
1167	Abanik																		
1168	Abanik																		
1169	Abanik																		
1170	Abanik																		

Figure 5. Summary counts

of varieties with standard language names (and the form of the standard language names themselves).

Finally, a note is warranted about the number of records in the standard language list. I operate with a working number of ‘standard languages’ at 397. However, there are more language names in the ‘standard’ language list for several reasons. Some work used a denser sampling than a single ‘variety’ per language. For example, the sampling for Bowers & Atkinson (2012) included some dialectal materials, in part because at that stage it wasn’t always clear which varieties should be excluded on grounds of mutual intelligibility. The easiest way to ensure database integrity was to create dummy standard language entries for the relevant varieties.

4.2 Sources The sources database is a straightforward bibliographic database which has a single table. This table includes information about the author of the source, date of publication (if published; otherwise date of creation if known), archival manuscript information, notes on the source’s legibility and reliability, and the number of entries. It is linked to the lexical data and reconstruction files. This file currently has 2607 sources. Many of the sources are also listed in the Zotero²⁵ bibliography management system. The purpose of this file is to allow users to recover the citation and source information for a particular wordlist; it is not meant to be a fully functioning reference management system.

The abbreviated source name is displayed in the lexical database. Most names are about six characters long, often the first few letters of the author’s name and an abbreviated year.

²⁵ www.zotero.org. The files can be viewed at <https://www.zotero.org/groups/pamanyungan>

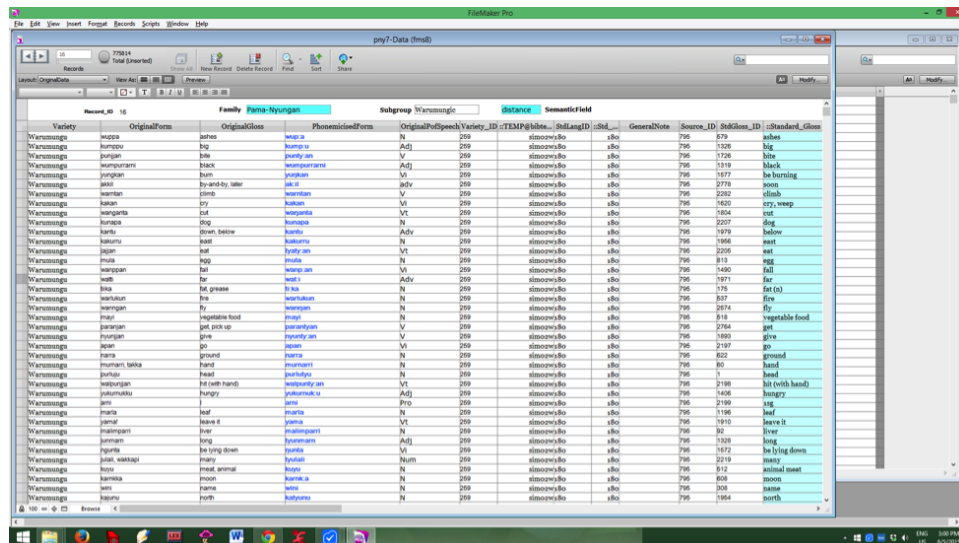
Table 4. Structure of the Sources database

Field	Description
Author	The author of the work, in Last Name, First Name format
Year	The year of creation or publication
Title	Title of the work
Publisher	Publisher of the work
Journal	Journal in which the work appeared
Pages	Page numbers of the source
CurrVarietyID	For data from Curr (1886), the vocabulary number
Processed	A field indicating whether the source has been entered into the lexical database (see workflow information in §5 below)
CurrLocation	For data from Curr (1886), the vocabulary location of collection
ItemCount	The number of lexical items in the source
AccessRights	Access restrictions and rights on the data (see §6.3 below)
Own	Yes/no field indicating whether the project has a physical or digital copy of the source
PrioritySource	1–3 scale indicating the transcription reliability of the source
AbbreviatedName	An abbreviated source name, equivalent to a BibTex key, which allows a short mnemonic reference to the source.

Sources are tagged for their ‘priority’ on a scale of 1–3. Priority 1 sources are the most likely to contain reliable data. They are transcribed phonemically and, as far as we can tell, accurately glossed. Priority 2 sources contain valuable information but they have some problem which makes them unsuitable for use as priority 1 sources. For example, they may have idiosyncratic transcription, or they may have other structural issues. For example, one source was a multi-dialect dictionary where the imported data contained words from several dialects in each lexical entry. Until those items can be split, the data cannot be used for phonological analysis or reconstruction (because most fields contain more than one word). Priority 3 sources are those which require expert knowledge of the language to interpret. Most priority 3 sources are 19th century wordlists in idiosyncratic spelling systems. Currently, 60 percent of the data in the lexical database comes from priority 1 sources, 24 percent from priority 2 sources, and the remaining 16 percent from priority 3 (of which the largest holdings are from Curr 1886). In future, we may implement a more fine-grained approach, such as that described in Cooper (2014).

4.3 Lexical data The core of the database is a set of lexical tables holding information about words in the different varieties. Each record in the database is a single lexical item in a single dialect. This data structure allows for linkage to other databases (for example, to the reconstructions data or to phylogenetic coding). It

also allows easy export to other formats for other projects (see §7). As of September 2015, there are 775,814 lexical items in the database. Because of the wide variety of information that the original sources chose to record about the lexical items, there is a great deal more incomplete data than for the source or language table. For example, only some of the wordlists give examples. Some include synonym or antonym information, while the majority do not. We did not want to discard information from the original sources (especially given that we did not know in advance which fields would be most useful). Other fields are incomplete because the work has yet to be fully processed. For example, many of the sources were not originally transcribed in standard orthographies, and so phonemicizing the data is a substantial research project in itself. More information on the types of data included in the database was given in §3 above.



Variety	OriginalForm	OriginalGloss	PhonemicizedForm	OriginalPurSpeech	Variety_ID	(TKM)ID	(Lang)ID	(S)ID	GeneralNote	Source_ID	(S)ID	(Standard_Gloss)
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	878	ama
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1326	big
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1728	big
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1319	black
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1877	be burning
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	2778	boat
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	2282	climb
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1823	dog
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1824	dog
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	2207	dog
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1490	fall
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1971	fat
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	175	fat (s)
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	807	file
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	2874	fly
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	218	vegetable food
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	2765	fat
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1893	give
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	2187	go
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	822	ground
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	80	hand
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1	hand
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	2198	hit (with hand)
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1428	hungry
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	2199	kg
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1198	leaf
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1810	leave it
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	82	live
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1933	long
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1872	be lying down
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	2179	man
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	812	animal parent
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	808	room
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	308	same
Warramungu	ama	ama	ama	N	209	st000000	st00	209		795	1954	north

Figure 6. Screenshot of main lexical database

The fields of this part of the database cluster into three types. First is the information from the original sources, such as the original form of the word, its part of speech, and gloss. This information is entered directly from the original source, with minimal editing. The second set of fields are processed forms of that data, such as the conversion of the original form into a standard transcription system (see §3.1 above) and the conversion of the original gloss to a standardized form (see further below). The third type of field comprises links to other parts of the database, such as wordlist sets for phylogenetic coding and links to the source information. The lexical database is the feeds data to reconstructions and coding tables. Information about how the reconstructions are set up is given in §4.5 below. The coding tables are sublists of lexical items for specific purposes. For example, the data for Haynie et al. (2014) come from a list of English translations of words denoting size, such as *big*, *little*, *small*, and so forth. Data with these glosses are tagged in a field in the lexical database, allowing for export (and re-export, if needed).

Table 5. Structure of the Lexical data database

Field	Description
OriginalForm	Form of the word as given in the original source
OriginalGloss	Gloss of the word as given in the original source
OriginalPofS	Part of speech of the word as given in the original source
PhonemicisedForm	Form of the word as in a standard phonemicized alphabet
PhonemicisedIPAForm	Form of the word as in IPA
Variety	Name of the doculect from which the word was recorded
StdGloss_ID	ID number linking the lexical item's gloss to the standardized gloss
StdLang_ID	ID number linking the lexical item's variety to the standardized language name
Source_ID	ID number linking the lexical item to its source (bibliographic) record
Rec_ID	ID number linking the lexical item to a reconstruction list/cognate set.
Rec_Code	Information on the type of relationship to the reconstruction list/cognate set (loan, inheritance, etc)
Coding@...	A series of fields providing ID numbers to standard sublists for data coding.
GeneralNote	Note on the form
DialectNote	Source's note on dialect (where the doculect contains information about more than one dialect)
EtymologyNote	Source's note on etymology
LiteralGloss	Literal gloss of the word, as given in the source
OtherNote	Other notes on the word, from the original source
ParadigmNote	Note on grammatical structure of the word, from the original source
SourcePages	Page or pages from the original source
Speaker	Speaker's name or initials, if given in the original source
Synonym	A synonym, if given in the original source
Examples	Example usage sentences, if in the original source
MacroSpelling	An automatically calculated field which removes vowel length, voicing, and retroflexion from transcription of the phonemicized form, facilitating sorting records by similar form.

We include lexical data headwords as they are presented in the original sources. This means that there is a variety of types of items as headwords, including inflected forms, stems, affixes, and phrases. Some authors include verb conjugation class information as part of the form (e.g., by suffixing the verb with a capital letter marking the conjugation class). Standardizing such data was out of the question without specialist knowledge of all the languages, and it is not clear that standardizing would produce a more useful result, since, after all, the languages vary in morphological complexity. Instead, we include as much information as possible in the part of speech tag. While

affixal information is not complete at present, we are currently gathering such information from reference grammars as a future release of Chirila will include affixes as well as headwords.

4.4 Glosses As the number of entries in the lexical database expanded, it became necessary to create a set of relational tables to deal with the glosses for lexical items, rather than keeping that information within the main lexical database. Most crucial was the creation of a set of standard glosses. Authors of dictionaries gloss words in many different ways, and choose different citation forms for the English gloss. For example, some authors gloss all verbs with the present progressive (*running, stepping, seeing*), while others use an infinitive (*to run, to step, to see*) and others use bare stems (*run, step, see*). All these words gloss the same concept. The input data vary extensively in how words are glossed. Some (most) have single word glosses or brief phrases, while a minority have full sentence definitions. Some sources are explicit about senses, with numbered senses, while others have implicit marking of senses, with different glosses set off by semi-colons or commas. Some authors include information in the gloss field with is not part of the meaning of the word, *stricto sensu*. Dialect information and grammatical information are the most common.

The database has a set of tables which look up ‘original glosses’ and map them to a set of standard forms. There is, however, a great deal of manual processing which needs to take place in order to map multiword glosses to standard categories, and this part of the database is still to be implemented fully. At present, only about 30 percent of the glosses are standardized.

Table 6. Tables for glosses

Table	Records ²⁶	Field	Description
Original-Glosses	239,141	CountofForms	Count of the number of times the original gloss appears in the lexical database
		GlossForm	Gloss as it appears in the OriginalGloss field in the lexical database
		StandardGloss_ID	ID associated with the standard form in the Standard_Glosses table
Standard-Glosses	7,430	Entry_Count	Count of the number of times the standard gloss appears in the lexical database
		ChangeNote	Annotations of decisions made about gloss forms

Continued on next page

²⁶The number of records is given as of September 30, 2015.

Continued from previous page

Table	Records	Field	Description
		English_PS	English part of speech (for disambiguation purposes)
		IDChapter	Chapter in the Intercontinental Dictionary Series List
		IDEntry	Entry number in the Intercontinental Dictionary Series List
		LinnaeanName	For flora and fauna terms, the genus and species information in binomial notation (e.g., <i>Chelonia mydas</i>)
		ID@...	(e.g., BasicVocabulary@ID) A series of fields which cross-reference the standard gloss with a subset wordlist, such as basic vocabulary, flora and fauna, or material culture.
		Standard_Gloss	The form of the standard gloss
		SemanticField_ID	ID which links the standard gloss to a list of standard semantic fields
		SWCategory	Category/chapter information for the Sutton and Walsh (1979) wordlist
		SWEntry	Entry number for the Sutton and Walsh (1979) wordlist
Semantic-Fields	55	Count	Number of linked entries in the standard gloss database
		Note	Note on choice of form for semantic field
		Semantic_Field	Semantic field entry
Kin-ship_Terms	679	Abbreviation	Abbreviation in standard anthropological form (e.g., Z for sister, M for mother)
		EnglishTerm	Spelled out in words (e.g., Mother's Sister's Child)
		Kin_ID	ID number which links with standardized glosses

Continued on next page

Continued from previous page

Table	Records	Field	Description
		Type	Superset terms, so that it is possible to extract all the in-law terms, grandparent terms, sibling terms, etc
Basic_Vocabulary	204	Term	Word
		Note	Note on word's inclusion, disambiguation information, etc
		PartofSpeech	English part of speech (for disambiguation)

Glosses are coded by semantic field. We use the categories developed by Sutton & Walsh (1979), with some additions, most notably categories for abstract terms such as emotions (which are missing from the original list). We also include ID numbers for the Sutton & Walsh standard list, as well as those for the *Intercontinental Dictionary Series* standard wordlist.

As discussed in §7 below, a focus of work arising from the lexical database has been exhaustive etymological coding, analyzed quantitatively. This requires subsets of standard wordlists. Thus several other standard wordlists have been set up with the same structure as the basic vocabulary list. For example, the set of sound symbolism terms used in Haynie et al. (2014) is coded as parallel to the Basic Vocabulary list used for Bower et al. (2011), as is the list of flora/fauna terms used in Bower, Haynie, et al. (2014). These lists can be automatically matched to data in the lexical database. Other lists, such as the list of standard kinship terms, require additional fields, such as kinship abbreviations.

4.5 Reconstructions The reconstructions database has a similar underlying structure to the lexical database, in that each record contains a single data point for a single language. In this case, however, the single data point is a proto-language reconstruction rather than an attested language. The reconstructions database is a flat table and each entry is labeled for the level to which it is reconstructed. This obviously presents problems, since reconstructions would ideally be represented in a set of hierarchical relationships.

Table 7 presents the major fields for the reconstruction portion of the database. The data include basic information about the reconstruction form, hypothesized gloss and part of speech, the subgroup or family to which the reconstruction pertains, and an indication of the researcher's confidence in the reconstruction, using the same 'priority' code system used for synchronic sources (where 1 is solid, 2 is probably

correct but with problems, and 3 is uncertain). A major problem in the development of the reconstructions database was determining the way in which to represent the hierarchical and evolving nature of both subgroup-level and proto-language level reconstructions. To see the problem, consider the type of information that cognate sets encode. Cognate sets comprise a list of words in a set of languages. Those languages are typically a subset of the languages in the family. The forms of the words in the daughter language provide the information for the form and meaning of the reconstructed word in the proto-language. The subgroup membership of the languages which attest the term provides the information about which level the word should be reconstructed to. This is illustrated below by selected entries for the word *parrkulu* ‘two’ in Pama-Nyungan. The form is fairly easily reconstructed as *parrkulu*. Pitta-Pitta and Pirriya have *parrkula*, with final a, but to reconstruct **parrkula* would be to assume several independent changes of a > u. Assuming the opposite is more parsimonious. The meaning is also fairly transparently reconstructed as ‘two.’ Diyari has the word in the meaning ‘three,’ but this is probably the result of syncope of a former compound form for three based on ‘two-one,’ as is found in many Australian languages (see further Bower & Zentz 2012). The languages which show this term, with one exception (Baagandji), belong to the Karnic subgroup of Pama-Nyungan, a group which is well established, though its internal structure is a matter of debate (Bower 2009, Breen 2007). Baagandji’s classification is in doubt; it is unclear whether it is an immediate sister to Karnic or more remotely related. This classification has implications for the status of *parrkulu*, however. If we assume that Baagandji is a sister to Karnic and *parrkulu* is an inheritance in the language, that means that we should assign the reconstruction **parrkulu* to the common ancestor of Baagandji and Karnic. If, however, we think that **parrkulu* originated in Proto-Karnic only, the form in Baagandji is most likely a loan. The trouble is that these sorts of decisions emerge from the process of cognate coding itself, and so the reconstructions database is thus both a record of assigning reconstructions to subgroups and a way of defining evidence for those subgroups themselves. This means that the subgroup structure cannot be hard-coded into the database, and also that there needs to be some way to mark items whose inheritance status depends on the subgrouping hypothesis, and not on other evidence such as loan phonology. The current solution has been to code all relationship types: loans, inheritances, doublets, backformations, and the like, in a single ‘etyma set,’ and to be flexible about the labels for the levels to which reconstructions are assigned. Reconstructions are grouped with the highest level where the information is constant. That is, if a form is reconstructed to an intermediate subgroup in a different form or meaning, it is given a different reconstruction code and cross-referenced with the related reconstructions.

While the initial focus of database development was to facilitate reconstructions across the Pama-Nyungan family and within individual subgroups, this reconstruction work has been overtaken by phylogenetic coding, as used in work such as Bower & Atkinson (2012). Reconstructions are not currently included in the public version of Chirila, but will be so in a future database release.

Table 7. Fields in the reconstruction database

Table	Records	Field	Description
Reconstructions	7023	Reconstruction_Form	Reconstructed proto-form
		PartofSpeech	Part of speech of the reconstructed proto-form
		Notes	Notes on the reconstruction
		RecSourceID	ID to link to source of reconstruction (in sources database)
		RecLevel	Subgroup or family to which reconstruction applies (e.g., Proto-Pama-Nyungan, Proto-Nyulnyulan)
		Gloss	Gloss of reconstructed proto-form
		SemanticFieldID	ID to link to semantic field
		Status	Status of the reconstruction
		Reliability	Score relating to the confidence in the reconstruction
		Author	Researcher who made the original proposal for the reconstruction
Reconstruction-CrossRefs	1632	ReferringRecord	Record ID of the reconstruction that is the source of the cross-reference
		ReferringTo	Record ID of the destination cross-reference
		XRefType	Type of cross-reference (e.g., merge records, see also)
		Note	Note on the cross-reference
		ReferringtoForm	Form of the destination cross-reference

Variety	Form	Part of speech/source	Note	Dig	Loan	Loan Source
Arbana	partuku	two	Num	reand	1	
Oyari	partuku	three	Num	awd1	1	
Garfall	partuku	two	Num	sp57	1	
Kunglari	barocda	two	Num	bre90	1	
Kunglari	boolara	two	Num	bre90	1	
Mitlala	partuku	two	Num	bre121	1	
Nbrngu	partuku	two	Num	low-rhu	1	
Pirinya	partuku	two	Num	bre90	1	
Punthamara	partuku	two	Num	ho08	1	
Punthamara	partuku	two	Num	ho08	1	
Wanglangu	partuku	two	Num	reand	1	
Wanglanyuru	partuku	two	Num	standw	1	
Wanglunara	partukubaru	two, an extra	Num	rotnd	1	

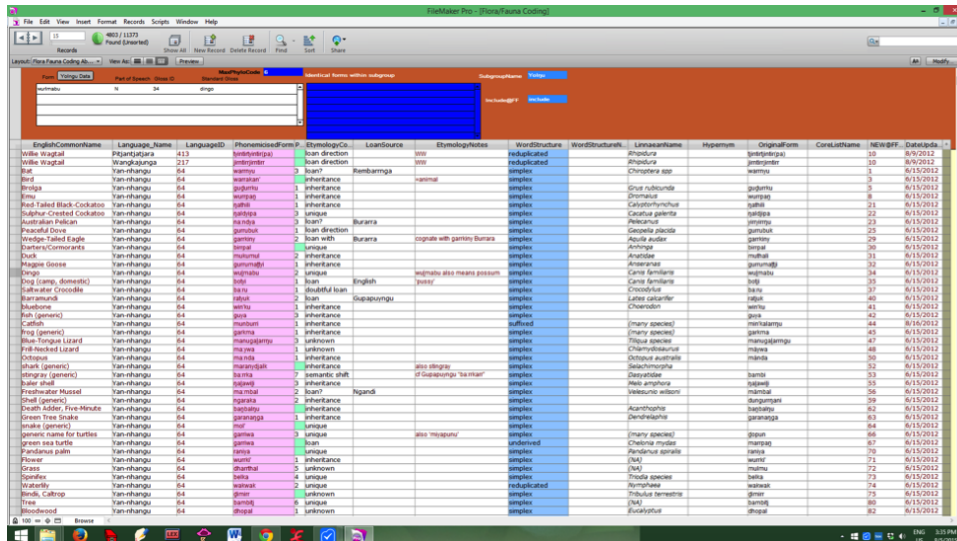
Figure 7. Screenshot of the reconstructions database

4.6 Community demography A file contains links to information about population size, population density, mobility, exogamy, and specific population estimates. This information was used in the modeling of the Australian case study in Bower et al. (2011) and Bower, Haynie, et al. (2014) and was compiled from published sources (such as population estimates in reference grammars) and the authors' knowledge of the region. It will eventually be expanded but at this point contains data only for the languages in those case studies.

4.7 Phylogenetic coding Several ancillary databases pull records directly from the master lexical database for coding for various features. The biggest of these by far is the lexical cognate coding database, which contains 204 words of basic vocabulary coded for phylogenetic analysis (see especially Bower et al. 2010; Bower & Atkinson 2012). Other coding includes analysis of several subdomains: body parts, color, and flora/fauna. The screenshot in Figure 6 gives an example from the coding that was done for Bower, Haynie, et al. (2014).

The phylogenetic databases are linked to the main database, rather than including the coding directly in the main database, for several reasons. First, because of the setup of the main database, there are often multiple attestations of a single English word in a given language. For example, the entry for 'father' in Warlpiri has 20 different records from different sources at different times with slightly different meanings.²⁷ These would either all need to be coded, or a way to work out which entry was the 'reference' entry would need to be developed. This way, the coding database pulls up all relevant records while keeping the coding consistent.

²⁷Because the language has trirelational kinship terms, the English translation 'father' ignores the relevance of the speaker's relationship to the propositus and/or referent which is also encoded in the Warlpiri term.



English/Common Name	Language Name	Language ID	Phonemised Form P	Etymology C	Loan Source	Etymology Notes	Word Structure	Word Structure	Linnæan Name	Hapemym	Original Form	Core Linnæan	NEWSFF	Date Added
Willie Wagtail	Pitjantjatjara	413	lamjetyep(ə)	loan direction	Wily	reduplicated	Wily	Wily	Ptilinopus	lamjetyep	lamjetyep	lamjetyep	10	8/9/2012
Willie Wagtail	Wangkajunga	237	wanyerpe	loan direction	Wily	reduplicated	Wily	Ptilinopus	wanyerpe	wanyerpe	wanyerpe	wanyerpe	30	8/9/2012
Red	Van-rhango	84	wanyer	loan?	Pambarrnga	simplex	simplex	Chrysura	wanyer	wanyer	wanyer	wanyer	1	6/15/2012
Red	Van-rhango	84	wanyer	inference		simplex	simplex	Chrysura	wanyer	wanyer	wanyer	wanyer	2	6/15/2012

Figure 8. Screenshot of sample database used for coding etymology in flora/fauna coding terms (used in Bown, Haynie, et al. 2014)

4.8 ‘Housekeeping’ and metadata This database contains metadata of different types, not only about the language information in the database, but also information about the sources, and the workflow. Adequate management of the workflow and processing has proven very important for the database but also rather difficult.

The database tracks who created records, when they are created, and who last edited them (and when). There is no automatic record restoration, though backups are made through Yale University’s ITS backup system every hour. Manual backups can also be made and are done so before any major work on the database. We keep a local clone of the database structure to test out major data manipulation before it is done on the main database.

5. Workflow One of the biggest challenges of this database stems from its dual status as research tool and data presentation device. That is, so many of the decisions about how best to present data can only be solved by research using the data itself. For example, assigning language varieties to standard names involves knowing how similar two varieties are, which (in the absence of statements on this topic in the literature) can only be determined by inspection of the data. Thus it was crucial to have a maximally modular and flexible workflow which minimized the analytical dependencies between steps.

5.1 Import raw data Data are digitized or prepared from digital sources for importing into the database. The exact tasks depend on the item. Fieldnotes and other print sources are typed into spreadsheets for later importing. We experimented with directly typing records into the database, but student data processors found it less cum-

bersome to work on their own laptops with a spreadsheet program (such as Google Sheets, Microsoft Excel, or Open Office). Students worked from a data template which included places to enter the language name, word form, gloss, part of speech, page number of source, and other information as needed. Original orthographies and glosses were preserved as far as possible. In some cases (such as in old materials with extensive diacritics), we had to substitute some symbols with others so that we could render the text within Unicode.

Digital data was prepared for import by converting it to tab delimited text. An undergraduate in the project, Sophia Gilman, wrote a Python script to convert backslash coded data to a tab-delimited table.²⁸ Many of the dictionary files that we received were in backslash code-format, since it is a common format used by field-workers with SIL's Toolbox (formerly Linguist's Shoebox) and Lexique Pro software programs, and earlier with Nisus Writer. The earliest digital files were from 1987, in backslash coded text format. We also had files from early versions of Microsoft Word and Wordstar. Text from these files was recovered and reformatted. We also made extensive use of manual text manipulation using Regular Expressions to convert text to a format that could be systematically imported.

5.2 Preliminary and periodic processing Once files were processed in a tabular format, they were imported to the Filemaker database. As the records were imported, several pieces of information were entered automatically using scripts in Filemaker Pro. The name of the doculect is filled down to all records, and then the associated standard language ID reference is auto-populated to the relevant field. The source is entered into the sources database and the source ID is copied to all records in the lexical data table. Further scripts construct a preliminary phonemicized wordlist from that of the original source, using Filemaker's 'substitute' script function. This phonemicized list is then examined for errors and any cleaning up required is done. Certain other information is also looked up automatically. For example, standard gloss IDs are looked up and populated from the conversion table of original to standard glosses. This in turn triggers associations with standardized wordlists (for example, the basic vocabulary list used in Bower et al. 2011). Finally, the records are 'eye-balled' (that is, cursorily examined by a human) for any data import problems and reimported or manually edited if necessary.

Certain other tasks need to be completed periodically. For example, the table that maps original glosses to their standardized IDs needs to be updated when a new wordlist is entered, since each new wordlist brings with it new glosses that need to be mapped to a standard ID. We also do periodic consistency checks, so as to ensure that fields are linked properly and that data is being auto-populated correctly. Occasionally, associations between original and standard gloss IDs change and must be corrected.

²⁸This script is available on the project site at pamanyungan.net.

5.3 Analytical and long-term processing Once vocabulary records are entered into the database and the records are correctly linked with the appropriate source and language information, they can feature in analytical work. For example, words can be tagged as belonging to reconstruction sets, coded for phylogenetic status, and used in other research projects (as discussed in §7 below). Associating original glosses with standard glosses is extremely time-consuming (see further § 6.2 below) and as yet, only 30 percent of the entries are associated with a standard gloss. This aspect of the database work is considered long-term research and curatorship.

Other aspects of work with the database are done at all stages of the workflow. For example, we are continually identifying new sources for importation into the database and adding them to the workflow.

6. Problems and Decisions Creating a database of this type is not without problems. This is especially the case when the database is to be used both for the compilation of cognates and primary reconstruction on the one hand, and for coding for computational work on the other. The first requires very flexible data structures with qualitative annotation. Coding for computational work, however, requires a rigorously enforced structure and precise decisions about what should be included. Further problems arise with missing or underspecified data. Moreover, importing data from many different sources makes enforcing consistency difficult. A final set of issues are not so much problems as decisions which have consequences. I outline some of these in this section.

6.1 Software program Given the current push in linguistics towards open-source software solutions (Bird 2009, Thieberger 2004), it might be wondered why such a complex and valuable linguistic resource would be programmed in a proprietary, commercial program such as FileMaker Pro, a software program which is neither open-source nor open-format. The answer is simple. None of the open-access, open-source solutions were adequate for the task. They would have required specialist programming, which would have been substantially more expensive than the cost of a few FileMaker licenses. Moreover, specialist programming would have considerably delayed the start of data collection. Other solutions were not flexible enough to allow us to easily create new queries and layouts as the database evolved. As a database that is simultaneously a data presentation and curation tool and a place where content is being actively created as research progresses, it was vital that the database could be easily set up and modified as need arose.

Because of the special needs of the database, off-the-shelf databases were inappropriate. The structure was complicated enough and the data copious enough that a spreadsheet or flat database (such as SIL's Toolbox) was not appropriate. FileMaker Pro was chosen because it was possible for the principal investigator to develop the resources necessary for the project in a reasonable time frame without relying on someone outside the project with programming experience. FileMaker is flexible enough to be able to handle the data structures required (with some necessary compromises, such as how to handle subgrouping; see above). It was also possible to create a user

interface which undergraduate students could work with without intensive training. This had a considerable advantage that we were able to make progress in adding data almost immediately. FileMaker's import and export options are extensive enough for our needs, and handling of Unicode characters is adequate. For example, some software programs treat accented characters as variants of unaccented characters, and a search for a would also return *á* and *à*. This behavior can be modified in FileMaker.

There have been disadvantages to the program, however. Because of the size and complexity of the database, it is difficult to share data within FileMaker. A web-hosted database using a variety of SQL would have solved this problem, but would have created others. For example, having a programmer create a user interface to a SQL database with all the functionality required would have been prohibitively expensive. The hosting computer for the database is locked-down to on-campus use, so only users who can authenticate against Yale's network (either by being on campus or through the Virtual Private Network) can use it at present. While this is a security advantage and makes the database less vulnerable to malicious use, it is a disadvantage for sharing. We have worked around this problem by exporting the database and hosting it on an external server (see §6.4) below. Another major disadvantage is the lack of a record-level 'undo' function – once a change is committed to a record, it cannot be undone. And there are several functions which can lead to massive loss of information. For example, a single keystroke allows the user to replace the contents of a field in all records in the view. This could lead to all the Original Forms (for example) being overwritten with null data. Our solution for this has been a combination of frequent (hourly) backups, limiting editing privileges to grant personnel who are experienced with the software, backing up before doing major database editing, checking record consistency after any major edits, and doing some data entry and manipulation externally and subsequently reimporting those entries into the database.

An aspect of this project that was important in the choice of software was the *emergent* nature of the data coding. That is, while the basic structure of the database was known in advance (and had been tested in a relational database model on a smaller scale), a number of other aspects of the project were mutually dependent and could not be resolved in advance. In short, the structure of the database has evolved as the project has evolved. The original estimate for database size was approximately 350,000 items. The database is now more than double that and still growing. We did not anticipate how many spin-off projects would be possible, but each of those projects required either extension to the database structure (through the creation of new views and queries) or additional data import.

While FileMaker does have fields for keeping track of record creation dates and times (and the users who created or modified the record), it is not straightforward to keep track of which items have been modified and what the changes were. In theory it would be possible to keep a 'history' log (like Greenhill et al. 2008 describe for the Austronesian Basic Vocabulary Database), but FileMaker is somewhat inefficient computationally and implementing such a script across all the files and tables would probably be prohibitively inefficient.

Data may be exported from individual tables and table views which contain information from multiple tables or files. This has been very helpful in creating custom views for research projects. Files are exported in XML and csv (comma-separated-values) text format. Exporting the data in these formats mimics the use of an open format program and ensures long-term data accessibility. An example of the xml output is provided below of the Warumungu word *kumppu* ‘big’.

```
<?xml version="1.0" encoding="UTF-8" ?>
<FMPDSORESULT xmlns="http://www.filemaker.com/fmpdsoreresult">
  <ERRORCODE>0</ERRORCODE>
  <DATABASE>PNY7-Data.fmp12</DATABASE>
  <LAYOUT></LAYOUT>
  <ROW MODID="76" RECORDID="496029">
    <OriginalForm>kumppu</OriginalForm>
    <OriginalGloss>big</OriginalGloss>
    <OriginalPofSpeech>Adj</OriginalPofSpeech>
    <PhonemicisedForm>kumppu</PhonemicisedForm>
    <Record_ID>2</Record_ID>
    <Source_ID>795</Source_ID>
    <StdGloss_ID>1326</StdGloss_ID>
    <StdLangID>180</StdLangID>
    <Variety>Warumungu</Variety>
    <Variety_ID>259</Variety_ID>
    <StandardLanguageName>
      <DATA>Warumungu</DATA>
    </StandardLanguageName>
    <SubgroupName>
      <DATA>Warumungic</DATA>
    </SubgroupName>
    <Standard_Gloss>
      <DATA>big</DATA>
    </Standard_Gloss>
  </ROW>
</FMPDSORESULT>
```

6.2 Workflow and import issues Curating a database requires decisions about content. Initially, we had to decide on which languages to focus on, whether to collect data systematically or opportunistically, whether to try to digitize complete sources or just concentrate on lexical items, and whether to retype data or use optical character recognition. There were also decisions regarding how to structure the import and how to standardize the many different conventions that linguists have used when creating records of the lexicons of Australia’s languages.

We decided to focus on Pama-Nyungan languages initially, rather than the whole of Australia and to collect and import data opportunistically, focusing on the sources

which were simplest to import, which were most readily available, and which were most likely to be relevant to immediate research projects. This meant the sources which were already digital (although not in database format), were easily available from libraries in the USA (or already in the author's private collection), typed rather than handwritten sources, and wordlists rather than full dictionaries. Data entry began in 2007 with languages of the Karnic subgroup of Pama-Nyungan and the comparative files developed by Barry Alpher for Alpher (2004). It was then expanded to the wordlists in the five-volume *Handbook of Australian Languages* series (edited by R.M.W. Dixon and Barry Blake), while materials from the former *Aboriginal Studies Electronic Data Archive* (ASEDA) were being requested and received. We then focused on ASEDA materials and continued data entry from print sources, gradually expanding to more complex lists as time permitted and as import and processing procedures were refined.

A major issue has been finding student researchers with enough experience in reading manuscripts to be able to complete data entry from field notes with speed and accuracy. The student workers on this project were, almost without exception, born in the 1990s and have used computers their whole lives. While many learn cursive writing in elementary school, they have never used it, and they very seldom read any handwriting but their own lecture notes (and many use a computer or tablet to take notes as well). They thus lack experience in reading handwritten sources and have great difficulty deciphering even relatively clear notes, although this improved to some extent with training in epigraphic techniques.

For simple sources such as wordlists, we imported all information. Where the source had more information, such as semantic field, Linnaean name of flora and fauna species, or ancillary lists of place names, we also imported this information. A decision was made early on in the import and collection process not to type example sentences, but to import them if they were already in a digital source. We include encyclopedic information such as ethnographic information about the use of artifacts if it appears. This decision aims to balance the utility of having more information rather than less with the reality that for a comparative database, material is of limited use if it only exists for a few languages. Since the clear majority of our sources do not contain example sentences, we would never be able to study comparative usage of lexical items using this database.

Linguists format the information in dictionaries in different ways. For example, some structure senses hierarchically, using sense numbers. Others list all glosses in the same entry, with commas or semi-colons as delimiters. Some use subentries, while others use all items as headwords, whether or not they are related to another word in the lexicon. We decided to remove sub-entry and sense number information, and treat all items as independent headwords. The disadvantage of this approach is that obviously identical lexical items are listed multiple times, and each entry has to be associated with the reconstruction set. It also means that we lose the linguist's intuition about differences between homophones and polysemous terms; essentially, we treat all items as homophones. However, there are also advantages, including the simplification in searching and association when there is a single standardized gloss,

and it means we can use the database to study colexification patterns without the filter of deciding between polysemy and homophony.

While importing the original data entries was fairly straightforward once a workflow was in place, creating a standardized output for comparison across sources has been very challenging, primarily because of the way in which lack of consistency introduces ambiguity. Since most modern descriptions of Australian languages use phonemicized orthographies, converting one set of characters to another is not difficult; however, there are many orthographic variants in use, even among consistent orthographies (e.g., <ty, tj, j, dj, dy, ɽ, ʈ, č, t, ch, c> for IPA /c/) which need to be examined by a human. Standardizing glossing also presented many challenges. Some decisions were clear: for example, associating original glosses such as *run*, *running*, *to run*, and *ran* with a single item *run* was straightforward. Associating singular and plural nouns with the same standard entry (such as *digging stick*, *digging sticks*) was also usually straightforward, though some languages have suppletive singular and plural terms (compare Yan-nhaŋu *ratha* ‘child’ versus *yitjiwala* ‘children’). Flora and fauna terms produced numerous problems. Sometimes words were glossed with slang terms which couldn’t always be identified precisely. For example, ‘five-minute snake’ (a snake so poisonous that it only takes five minutes for someone to die if bitten) is a gloss in a number of older wordlists from north-western Australia, but it is unclear if the term refers to a tiger snake or death adder. Sometimes comparative work within the lexicon itself can help resolve these issues, but we do not want to discount the possibility that the word might have undergone semantic shift in one or more languages, and thus not mean the same thing in all varieties.

There were many cases where items were glossed with differing degrees of generality, or with items that are near (but not total) synonyms in English. For example, some wordlists have an item ‘bloodwood,’ while others are more specific about genus and species.²⁹ Other lists have several items glossed as ‘bloodwood’ but no further information. In that case, we cannot tell if they are synonyms, different species, words for trees at different life stages, or different parts of the tree. In these cases we are limited by the accuracy and specificity of the source wordlists, and the accuracy is unknowable. Moreover, speakers of the languages may themselves disagree as to the precise meaning of particular words. We thus adopt an approach that aims to facilitate searching within the database, but not ‘over-analyze.’ Thus in addition to specific glosses such as ‘father’s sister’ and ‘mother’s sister,’ we have a general gloss ‘aunt.’ For imprecise glossing of this kind, we associate entries with the most specific ‘standard’ term that we have evidence for. For example, for sources which provide Linnaean names, we associate the standard gloss with that name; for sources that provide only common English names, we use those; and for sources which only use glosses such as ‘kind of tree’ or ‘gum tree,’ we retain that level of generality.

Other issues in creating standardized glosses come from difficulties in interpreting the English glosses. English homophones create a particular problem in decontextu-

²⁹Note that ‘bloodwood’ brings up an additional issue in using Linnaean classification to identify species, since they were reclassified from the genus *Eucalyptus* to the genus *Corymbia* in the 1990s; most of the lexical sources which use Linnaean names were compiled before the reclassification, and give the earlier names.

alized wordlists. For an item glossed as *spear*, for example, are we dealing with the noun or the verb? For a word glossed as ‘aunt, father’s sister,’ how should we interpret the comma? Does the word mean ‘aunt (generally), but particularly father’s sister?’ or ‘father’s sister, one of the meanings of English ‘aunt’”? For ‘sky, cloudy,’ is this a ‘cloudy sky’ (as opposed to a clear one,) or the noun ‘sky’ that may also be used as a modifier meaning ‘cloudy’? And so on. In some cases, we can recover the information from the order of the original wordlist. For example, if nouns and verbs are listed in different parts of the wordlist, we can tell whether *spear* should be a noun or a verb. In other cases, however, we cannot tell. Because of these issues, associating glosses with standard codes is extremely time-consuming.

One of the first decisions is the degree to which the database should enforce standards and completeness. From early on in the project, the decision was taken to make the data as useful as possible, as quickly as possible. If every entry had been proofed before using any of the information in the database and before the next source was imported, we would never have been able to use the database for analysis. Therefore we adopted a ‘proof as you go’ approach—that is, we spot-check data on import, look for obvious problems, and do automatic data processing of items such as source look-up, phonemicization, and the like. But we do not, for example, exhaustively associate all records with a standard gloss marker at import. This limits the utility of some of the database functions. For example, searching by standard gloss misses forms. But this limitation is outweighed by being able to use the database in its entirety and to fix issues as they arise.

A final issue for import concerns how to deal with superseding sources. Like the Chirila database itself, some of the input dictionaries are works in progress, while others have complex dependencies. For example, Aklif’s (1999) Bardi dictionary is included in my 2003 (unpublished) Bardi dictionary and supplement; the latter source was built on the former but contains more than double the number of entries, along with numerous additional examples and senses. In that case, the latter source contains all the same information (and more) as the earlier one, and so it would be justified to include only the latter. However, many other cases are not so clear. For example, some modern sources are built on 19th century materials, but re-transcribed and annotated in ways that are not always transparent. The latter source clearly ‘adds value’ to the earlier source, but is not strictly superseding, because the earlier source contains different information (different glosses, a different type of transcription, etc). Therefore we have erred on the side of including multiple sources, even at the risk of duplication. We are currently testing a filter which combines all entries for a standard language which have the same phonemicization and standard gloss, however, this is hampered by the difficulties in associating entries with standardized glosses, as described above.

6.3 Access issues A different type of problem concerns access to the content of the database. In the original compilation of the database, we had assumed that both researchers and communities would not want to make materials more generally available. Currently the only people with access to the data are the research team and a

few other people who requested subsets of the database. It seems to me to be a shame that such a useful tool is not more widely available. There are also inevitable glitches in importing data from so many types of sources, and broadening the number of people with access will hopefully allow us to identify and fix many of those problems.

A considerable amount of material was compiled from field notes or other unpublished sources. Some of the material is work in progress and files were provided to the project on the condition that they were treated as confidential. This project has relied extensively on the goodwill of Indigenous communities and other researchers in Australian languages, and with very few exceptions researchers have responded generously. Thus far, only one researcher has declined to allow materials to be used in the database, and only a few of the researchers contacted about publicly releasing material have asked that we not do so.³⁰

The second concern involves Indigenous communities. There has been steady politicization of language—even everyday words without ceremonial or controversial content—as indigenous language use has receded. This is particularly the case in urban communities. There have been vocal protests at conferences such as meetings of the Australian Linguistic Society and the Australian Institute of Aboriginal and Torres Strait Islander Studies, where researchers have been accused of breaching copyright and theft of materials (see, for example, some of the comments in Amery & Nash 2008:15–24). It is difficult to gauge how widespread these views are, because while the protests over the dissemination of language data are very vocal, they are balanced by others who wish to see their languages recognized and treated with the same respect that is given to national and more widely known languages. My discussions with Aboriginal community groups on this database have been uniformly positive, once the purpose of the database has been clarified, and the communities I have worked with have all been happy for materials to be included in the database and shared for educational purposes.

The database has been an excellent informal resource for Indigenous communities working on their languages. The project team has been in contact with language centers in Queensland and Western Australia, as well as Aboriginal individuals from all states and territories in Australia. Materials from the database have been shared informally, where allowed under the terms of the original acquisition. They have been used in language workshops, in the development of school and community language resources, and by individuals (particularly members of the Stolen Generations) reconnecting with their heritage.

The most restrictive and protective group by far have been the archives (both in Australia and overseas) with whom I have negotiated for access to materials. This is, perhaps, to be expected, if the archive views their role as both protecting the rights of individuals and housing and preserving the physical copies of language records. However, the policies of some of these institutions restrict materials extensively. For example, one archive's terms of release of materials includes a clause that prevents me from disclosing that I have a copy of the materials, let alone quoting from them.

³⁰In every case, the request was because they were unable to contact the relevant community members for permission. We have, of course, complied with these requests.

My experience in talking with researchers and communities here is that archives with restrictive policies like this are very much out of step with the general wishes of both Aboriginal groups and researchers.

6.4 Chirila online The original database development (2007–2012) did not include plans to make the sources freely available, in large part because I had assumed that most people would not want to make them publicly available. However, continuing informal discussions with researchers and community members revealed that this assumption was unfounded. Therefore, in 2014 I began plans to release the database. Transitioning the database from an internal research tool under active development to something that could be used by others and quoted has been time-consuming and complex. For example, we have contacted as many of the original depositors as we can in order to clarify their wishes regarding the materials they gave us (we have not released any data without explicit permission to do so). The format of data release required careful thought. Though many online databases have a web search interface, we decided at this point to release text format and excel files of different types. This is because discussions with potential Chirila users suggests that the two main uses of the database will be people who want wordlists of particular languages and researchers who want to download the entire database for use in their own research with their own data-manipulation programs. Phase one of the database release includes approximately 180,000 records from more than 80 sources. They include both historical sources, such as Curr (1886), and modern sources. Further information is available from pamanyungan.net/Chirila, including a list of sources and languages, plans for future data releases, and ways for researchers and community members to contribute data and correct errors.

7. Projects arising from the database In addition to the reconstruction and prehistory research described above, there have been other projects that have come out of the database. This section provides some examples of the uses to which data in the database has been put thus far.

7.1 Language reclamation As mentioned in §6.3 above, data from Chirila has been used in informal work with communities and individuals working to preserve, teach, and revitalize their languages. Because the data are in a standard format, we are able to export data quickly to generate wordlists for language projects. Turnaround time for sources is typically within 24 hours; we can export all holdings, which for some languages represents all the records of the language ever made. We can also generate source lists and reference lists for resources for people who want to see the original materials. Furthermore, because our data entry pipeline is very flexible, we have been able to partner with some language centers and Aboriginal communities to prioritize sources which would be most useful for them. We hope very much that this aspect of the database work will continue and expand as more data are released and Chirila's existence becomes better known.

7.2 Synchronic language work The database can be used for studying patterns of diversity within Australian languages. We have so far looked at acculturation terms (in unpublished work), tendencies for sound symbolism (Haynie et al. 2014), and generalizations about phonotactic constraints in Australian languages (Gasser & Bower 2014).

In Gasser & Bower (2014), we show the utility of deriving information about the phonologies of Australian languages directly from lexical data.³¹ Australian languages are famous for their near-uniform phonemic inventories (Busby 1980, Dixon 1980, Hamilton 1995, Butcher 1994, Voegelin et al. 1963:24). The apparent uniformity of Australian languages also stands out in worldwide typological surveys (Mielke 2008, Maddieson 1986). Otherwise unqualified statements about uniformity in inventory and phonotactics are easily found in reference grammars of languages in the region (e.g., Goddard 1985:21, 43, 66, 323). This assumption is, in itself, 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. Such uniformity, if real, is surprising and unusual given the country's phylogenetic diversity.

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. Moreover, reference grammars do not contain uniform information; for example, some exhaustively list the clusters found in the language, while others give only summary statements by place and/or manner of articulation, while yet others list only the most common clusters. This makes systematic comparison across languages impossible. Gasser & Bower (2014) is a proof-of-concept study, where 145 languages 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. Wordlists were converted to a single set of standard symbols. The relevant 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 but found many exceptions. Particularly important were 'minority' patterns, which appear to be systematically overlooked in Australian phonological typologies. 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.'

Other work that uses the database as a data source includes Hunter et al. (2011); Bower, Skilton, et al. (2014); Bower & Zentz (2012); and Zhou & Bower (2015). Thus far, sixteen refereed articles have appeared that primarily use data from the lexical database.

³¹The background information in this section is closely based on Gasser & Bower (2014).

7.3 Diachronic language work The major historical publication from the data at this point is Bower & Atkinson (2012), which describes the first fully articulated phylogeny for Pama-Nyungan. Bower and Atkinson (2012) published the first fully resolved phylogeny of Pama-Nyungan using data coded from 194 languages. Since then, the language sample has been expanded to 315 languages and to Pama-Nyungan's closest putative relatives (the Tangkic and Garrwan families). In the original paper, 189 words of basic vocabulary were coded for etymology status from languages and dialects from across the Pama-Nyungan family. These etymological patterns were then analyzed computationally to produce a tree (more information on the methods used can be found in Bower & Atkinson (2012) and Bower (2015)). The resulting family tree is shown in Figure 10.

We have also been able to use the dataset to study loanword rates. Language contact has featured prominently in historical analyses of Australian languages. Australian languages are also seen in the wider literature (e.g., Haspelmath 2004) to represent a case of high borrowing; this is confirmed with Gurindji, the sole example of an Australian language in Haspelmath & Tadmor's (2009) typology of loanwords. The real picture for Australia, however, reveals that Gurindji is quite atypical, and despite contact, the number of loan items in basic vocabulary for most languages is small. The data reveal considerable variation in loans, even among languages which had extensive contact with their neighbors. The results range from under 10 percent loans to almost 50 percent. Moreover, the distribution of loan percentages across languages almost directly tracks that from Haspelmath & Tadmor's (2009) worldwide survey (see Figure 11 below). In this work (published as Bower et al. 2011), loans were identified using standard procedures in comparative linguistics. For example, a word in language A is likely to be a loan from language B to language A if it shows morphology which is interpretable in B but not in A, if it violates the evidence of sound change or synchronic phonotactic patterns in language A, and so on.

Information from the database has featured in several philological publications tracking the borrowing of lexical items from Aboriginal languages into Australian English (Nash 2013, Nash 2014). For this type of project, the historical depth of sources has been crucial. The database was also used by researchers at the Australian National Dictionary Centre in updating entries for English words of Aboriginal origin.

7.4 Organizing and presenting information about Australian languages Quite apart from elucidating the history of Australian languages through their lexicon, the database has produced a number of research results which challenge received wisdom about Australian languages. Perhaps the most crucial result here is the number of languages attested in the country. Previous works suggest that 250 languages is an accurate estimate, with approximately two-thirds of those languages belonging to the Pama-Nyungan family. Dixon (1980, 2002) gives this figure, for example, and it has been widely repeated. However, compiling data for 'standard' languages from the database suggests that a more accurate figure is 397, 299 of which (or 75%) are Pama-Nyungan. It is possible that there is some over-counting, and that given that both linguists and language speakers draw the boundaries between 'languages' and

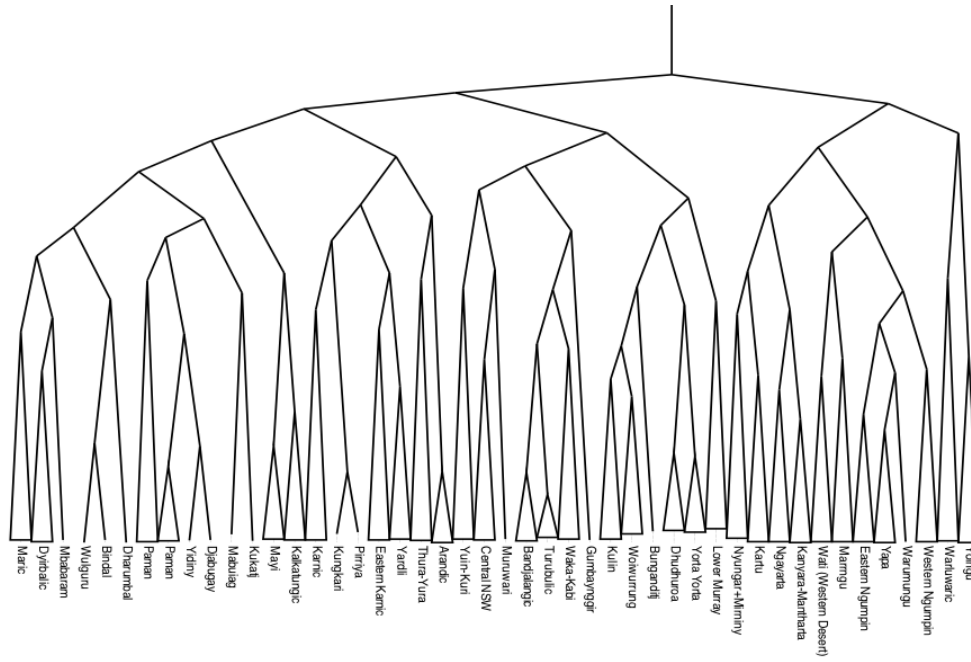


Figure 9. Subgroups of Pama-Nyungan, from Bower & Atkinson (2012)

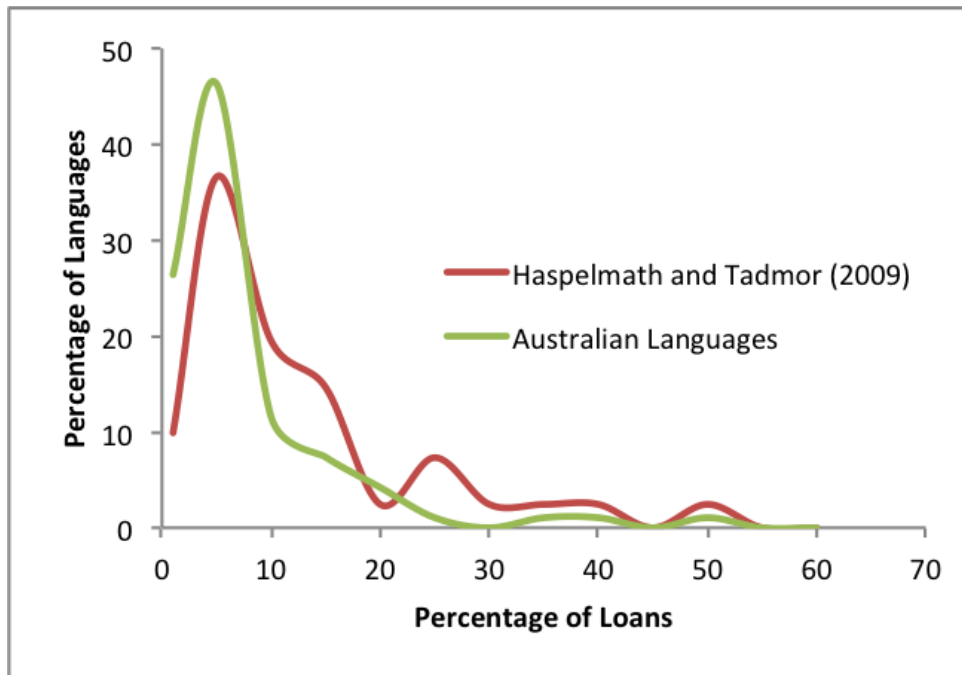


Figure 10. Frequency of loans in 49 Australian languages from across the country

‘dialects’ in different places, that the difference in numbers simply reflects a different standard of what counts as sufficiently different to warrant a different standard language name. Even so, the disparity in lists is substantial, with my list being about 60 percent longer. The big differences in numbers suggest that this is not simply a matter of which varieties are treated as dialects, and that a comprehensive survey has revealed systematic underestimation of the diversity of Australia’s Aboriginal heritage.

8. Conclusions and future directions Australian languages have been claimed to be exceptional in several ways and not amenable to analysis by the historical methods used in other parts of the world (Dixon 1997, Dixon 2002, Mühlhäusler 1996). It is claimed that either too much language contact is in evidence, which obscures the ‘true’ genetic relationships, or the data are simply too poor to be useful for detailed historical study. However, because Australian languages have, for so long, been thought to be problematic for lexical reconstruction, authors have tended to rely on other methods to discuss linguistic relationships (see, for example, Harvey 2008, McGregor & Rumsey 2009). As a result, there is a dearth of lexical reconstruction and much basic research which still needs to be done.

In the last five years, Australian languages have gone from being a historical ‘backwater’ to one of the main testing grounds for new theories of evolutionary linguistics. The Pama-Nyungan lexical phylogenetic tree (Bower & Atkinson 2012) is the second largest published phylogeny (after Austronesian; Gray et al. 2009) and the most comprehensive in terms of number of languages sampled in the family. Some questions, particularly in phylogenetics, can only be answered with data such as this. Very few language families have the comprehensive data coverage to be able to marshal systematic comparisons. The only other equivalently sized family for which this type of research is possible at this stage is Indo-European; other language families have phylogenetic trees (Atkinson et al. 2005, Holden 2002) but not the broad lexical database which enables rapid comparison and reconstruction.³²

A large-scale comparative lexical database has allowed research into Australia’s prehistory that was unthinkable just a few years ago. It is hoped that opening up this resource to other researchers will allow Australian languages to be included more representatively in worldwide surveys, and that more researchers within Australia will be able to take advantage of its holdings for their own work, from detailed etymological comparisons of individual lexical items to broad-scale comparisons, to studying trends in language contact, to semantic change, to making language resources accessible and usable for the speakers of these languages and their descendants.

³²The Austronesian Basic Vocabulary Database contains many entries but only includes Swadesh-type lists, so cannot be used for semantic field comparison. Other etymological databases, such as Starling (<http://starling.rinet.ru/cgi-bin/main.cgi>), are organized by etyma, which gives reconstruction information but makes their use in phylogenetics very cumbersome. It is also difficult to download data for use in other database programs.

References

- Aklif, Gedda. 1999. *Ardiyooloon Bardi ngaanka: One Arm Point Bardi dictionary*. Halls Creek, W.A.: Kimberley Language Resource Centre.
- Alpher, Barry. 1991. *Yir-Yoront lexicon: Sketch and dictionary of an Australian language*. Berlin & New York: Mouton de Gruyter.
- Alpher, Barry. 2004. Pama-Nyungan: Phonological reconstruction and status as a phylogenetic group. In Claire Bowern & Harold Koch (eds.), *Australian languages: Classification and the comparative method*, 93–126. Amsterdam & Philadelphia: John Benjamins.
- Alpher, Barry, Geoffrey O’Grady & Claire Bowern. 2008. Western Torres Strait language classification and development. In Claire Bowern, Bethwyn Evans & Luisa Miceli (eds.), *Morphology and language history: In honour of Harold Koch*, 15–30. Amsterdam & Philadelphia: John Benjamins.
- Amery, Rob. 2000. *Warrabarna Kaurna!: Reclaiming an Australian language*. Exton, PA: Swets & Zeitlinger Publishers.
- Amery, Rob & Joshua Nash (eds.). 2008. *Warra Wiltaniappendi = Strengthening Languages: Proceedings of the Inaugural Indigenous Languages Conference (ILC) 2007*, September 24–27, 2007, University of Adelaide, S.A. Adelaide: Discipline of Linguistics, University of Adelaide.
- Atkinson, Quentin D., Geoff Nicholls, David Welch & Russell Gray. 2005. From words to dates: Water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society* 103(2). 193–219. DOI:10.1111/j.1467-968X.2005.00151.x.
- Austin, Peter. 1981. *A grammar of Diyari, South Australia*. Cambridge: Cambridge University Press.
- Austin, Peter. 1990. Classification of Lake Eyre languages. *La Trobe University Working Papers in Linguistics* 3. 171–202.
- Bird, Steven. 2009. Natural language processing and linguistic fieldwork. *Computational Linguistics* 35(3). 469–474. DOI:10.1162/coli.35.3.469.
- Black, Paul. 1980. Norman Paman historical phonology. In Bruce Rigsby & Peter Sutton (eds.), *Papers in Australian Linguistics* 13, 181–239. Contributions to Australian Linguistics. Pacific Linguistics A-59. Canberra: Pacific Linguistics.
- 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.
- Bowern, Claire. 2003. Laves’ Bardi texts. In Joseph Blythe & R. McKenna Brown (eds.), *Maintaining the links: Language, identity and the land. Proceedings of the Seventh FEL [Foundation for Endangered Languages] Conference*, 137–143. Broome, W.A.: Foundation for Endangered Languages.
- Bowern, Claire. 2009. Reassessing Karnic: A reply to Breen (2007). *Australian Journal of Linguistics* 29(3). 337–348.

- Bowern, Claire. 2010. Historical linguistics in Australia: Trees, networks and their implications. *Philosophical Transactions of the Royal Society B* 365. 3845–3854. DOI:10.1098/rstb.2010.0013.
- Bowern, Claire. 2012a. Nyikina paradigms and refunctionalization: A cautionary tale in morphological reconstruction. *Journal of Historical Linguistics* 2(1). 7–24. DOI:http://dx.doi.org/10.1075/jhl.2.1.03bow.
- Bowern, Claire. 2012b. The riddle of Tasmanian languages. *Proceedings of the Royal Society B* 279. DOI:10.1098/rspb.2012.1842.
- Bowern, Claire. 2015. Data “big” and “small”—Examples from the Australian lexical database. *Linguistics Vanguard* 1(1). 295–303. DOI:10.1515/lingvan-2014-1009.
- Bowern, Claire & Quentin Atkinson. 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* 88(4). 817–845. DOI:10.1353/lan.2012.0081.
- Bowern, Claire, Grace Brody & Patrick Killian. 2014. *Revisiting 'Standard Average Australian'*. New Haven, CT: Yale University MS.
- Bowern, Claire, Patience Epps, Russell D. Gray, Jane Hill, Keith Hunley, Patrick McConvell & Jason Zentz. 2011. Does lateral transmission obscure inheritance in hunter-gatherer languages? *PLoS ONE* 6(9). e25195. DOI:10.1371/journal.pone.0025195.
- Bowern, Claire & Bethwyn Evans. 2015. Editors' introduction: Foundations of the new historical linguistics. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge Handbook of Historical Linguistics*, 1–42. London & New York: Routledge.
- Bowern, Claire, Hannah Haynie, Catherine Sheard, Barry Alpher, Patience Epps, Jane Hill & Patrick McConvell. 2014. Loan and inheritance patterns in hunter-gatherer ethnobiological systems. *Journal of Ethnobiology* 34(2). 195–227. DOI:10.2993/0278-0771-34.2.195.
- Bowern, Claire, Amalia Skilton & Hannah Haynie. 2014. Lexical stability and kinship patterns in Australian languages. Poster presented at LSA [Linguistic Society of America] Annual Winter Meeting, January 2–5, 2014 in Minneapolis, MN.
- Bowern, Claire & Jason Zentz. 2012. Diversity in the numeral systems of Australian languages. *Anthropological Linguistics* 54(2). 133–160. DOI:10.1353/anl.2012.0008.
- Breen, Gavan. 2007. Reassessing Karnic. *Australian Journal of Linguistics* 27(2). 175–199. DOI:10.1080/07268600701522780.
- Busby, Peter A. 1980. The distribution of phonemes in Australian Aboriginal languages. In Bruce E. Waters & Peter A. Busby, *Papers in Australian Linguistics* 14, 73–139. Pacific Linguistics Series A-60. Canberra: Pacific Linguistics.
- Butcher, Andrew. 1994. On the phonetics of small vowel systems: Evidence from Australian languages. *Proceedings of the 5th Australian International Conference on Speech Science and Technology*, vol. 1, 28–33.
- Capell, Arthur. 1956. *A new approach to Australian linguistics (Handbook of Australian languages, Part 1)*. (Oceania Linguistic Monographs 1). Sydney: University of Sydney.

- Cooper, Doug. 2014. Data warehouse: Bronze, gold, STEC, software. *Workshop on the use of computational methods in the study of endangered languages*, 91–99. Baltimore, MD: Association for Computational Linguistics.
- Crowley, Terry & R.M.W. Dixon. 1981. Tasmanian. In R.M.W. Dixon & Barry Blake (eds.), *Handbook of Australian languages*, vol. 2, 394–427. Philadelphia, PA: John Benjamins.
- 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*. Melbourne: J. Ferres.
- Dawson, James. 1881. *Australian Aborigines: The languages and customs of several tribes of Aborigines in the western district of Victoria, Australia*. Melbourne, VIC: G. Robertson.
- Dixon, R.M.W. 1980. *The languages of Australia*. Cambridge: Cambridge University Press.
- Dixon, R.M.W. 1997. *The rise and fall of languages*. Cambridge: Cambridge University Press.
- Dixon, R.M.W. 2002. *Australian languages: Their nature and development*. Cambridge: Cambridge University Press.
- Dixon, R.M.W. 2007. Field linguistics: A minor manual. *Sprachtypologie und Universalienforschung* 60(1). 12–31. DOI:10.1524/stuf.2007.60.1.12.
- Gasser, Emily & Claire Bower. 2014. Revisiting phonological generalizations in Australian Languages. *Proceedings of the Annual Meetings on Phonology 2013*. DOI:<http://dx.doi.org/10.3765/amp.v11i.17>
- Goddard, Cliff. 1985. *A grammar of Yankunytjatjara*. Alice Springs, N.T.: Institute for Aboriginal Development.
- Good, Jeff & Michael Cysouw. 2013. Languoid, doculect, and glossonym: Formalizing the notion ‘language’. *Language Documentation & Conservation* 7. 331–359. <http://hdl.handle.net/10125/4606>.
- Gray, Russell D., Alexei J. Drummond & Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323(5913). 479–483. DOI:10.1126/science.1166858.
- Greenhill, Simon J., Robert Blust & Russell D. Gray. 2008. The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics Online* 4. 271–283.
- Hale, Kenneth L. 1976. Phonological developments in particular Northern Paman languages. In Peter Sutton (ed.), *Languages of Cape York: Papers presented to the Linguistic Symposium, Part B, held in conjunction with the Australian Institute of Aboriginal Studies Biennial General Meeting, May, 1974*, 7–40. (Australian Aboriginal Studies—Research and Regional Studies 6). Canberra: Australian Institute of Aboriginal Studies.
- Hamilton, Philip. 1995. Vowel phonotactic positions in Australian Aboriginal languages. *Proceedings of the Twenty-First Annual Meeting of the Berkeley Linguistics Society [BLS]: General Session and Parasession on Historical Issues in Sociolinguistics/Social Issues in Historical Linguistics*. 129–140.

- Harvey, Mark. 2008. Proto Mirndi: A discontinuous language family in northern Australia. *Pacific Linguistics* 593. Canberra: Pacific Linguistics. <http://hdl.handle.net/1959.13/43719>.
- Haspelmath, Martin. 2004. How hopeless is genealogical linguistics, and how advanced is areal linguistics? *Studies in Language* 28(1). 209–223.
- Haspelmath, Martin & Uri Tadmor (eds.). 2009. *Loanwords in the world's languages: A comparative handbook*. Berlin: De Gruyter Mouton.
- Haynie, Hannah, Claire Bowern & Hannah LaPalombara. 2014. Sound symbolism in the languages of Australia. *PLoS ONE* 9(4). e92852 DOI:10.1371/journal.pone.0092852.
- Hendrie, Timothy R. 1981. Distinctive features matching as a basis for finding cognates. *Working Papers of the Linguistics Circle* 1(1). 32–41.
- Hewson, John. 1993. *A computer-generated dictionary of proto-Algonquian*. Hull, Quebec: Canadian Museum of Civilization.
- 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.
- 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.
- Hunter, Jessica, Claire Bowern & Erich Round. 2011. Reappraising the effects of language contact in the Torres Strait. *Journal of Language Contact* 4(1). 106–140. DOI:10.1163/187740911X558798.
- Hymes, Dell. 1956. Na-Déné and positional analysis of categories. *American Anthropologist* 58(4). 624–638. doi:10.1525/aa.1956.58.4.02a00040.
- Koch, Harold. 1997. Pama-Nyungan reflexes in the Arandic languages. In Darrell Tryon & Michael Walsh (eds.), *Boundary rider: Essays in honour of Geoffrey O'Grady*, 271–302. Pacific Linguistics Series C-136. Canberra: Pacific Linguistics.
- Maddieson, Ian 1986. The size and structure of phonological inventories: Analysis of UPSID. In John J. Ohala & Jeri J. Jaeger (eds.), *Experimental phonology*, 105–123. Orlando: Academic Press.
- McGregor, William & Alan Rumsey. 2009. *Worrorran revisited: The case for genetic relations among languages of the Northern Kimberley region of Western Australia*. Pacific Linguistics 600. Canberra: Pacific Linguistics.
- Mielke, Jeff. 2008. *The emergence of distinctive features*. New York: Oxford University Press.
- Mühlhäusler, Peter. 1996. *Linguistic ecology: Language change and linguistic imperialism in the Pacific region*. London: Routledge.
- Nash, David. 2013. Wind direction words in the Sydney language: A case study in semantic reconstitution. *Australian Journal of Linguistics* 33(1). 51–75. DOI: 10.1080/07268602.2013.787905.
- Nash, David. 2014. The value of scientific names from (Australian) indigenous languages. In Patrick Heinrich & Nicholas Ostler (eds.), *Proceedings of FEL XVIII*, 37–42. Okinawa: Foundation for Endangered Languages.

- O'Grady, Geoffrey N. 1990a. Prenasalization in Pama-Nyungan. In Philip Baldi (ed.), *Linguistic change and reconstruction methodology*, 451–477. Berlin & New York: Mouton de Gruyter.
- O'Grady, Geoffrey N. 1990b. Wadjuk and Umpila: A long-shot approach to Pama-Nyungan. In Geoffrey N. O'Grady & Darrell T. Tryon (eds.), *Studies in comparative Pama-Nyungan*, 1–10. Pacific Linguistics C-111. Canberra: Dept. of Linguistics, Research School of Pacific Studies, Australian National University.
- O'Grady, Geoffrey N. 1998. Toward a Proto-Pama-Nyungan stem list, part I: sets J1–J25. *Oceanic Linguistics* 37(2). 209–233.
- Plomley, N.J. Brian. 1976. *A word-list of the Tasmanian Aboriginal languages*. Launceston, Tas.: N. Plomley in association with the Government of Tasmania.
- Rankin, Robert L., Richard T. Carter, A. Wesley Jones, John E. Koontz, David S. Rood & Iren Hartmann (eds.). 2015. *Comparative Siouan dictionary*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://csd.clld.org>.
- Ridley, William. 1875. *Kamilaroi and other Australian languages*, 2nd ed. Sydney: Government Printer.
- Roth, Walter E. 1897. *Ethnological studies among the north-west-central Queensland Aborigines*. Brisbane: Government Printer.
- Sutton, Peter & Michael Walsh. 1979. *Revised linguistic fieldwork manual for Australia*. Canberra: Australian Institute of Aboriginal Studies.
- Teichelmann, Christian G. & Clamor W. Schürmann. 1840. *Outlines of a grammar, vocabulary, and phraseology, of the Aboriginal language of South Australia, spoken by the natives in and for some distance around Adelaide*. Adelaide, S.A.: Teichelmann & Schürmann.
- Thieberger, Nicholas. 2004. Documentation in practice: Developing a linked media corpus of South Efate. In Peter Austin (ed.), *Language Documentation and Description*, vol. 2. 169–178. London: SOAS. <http://hdl.handle.net/11343/34484>.
- Thieberger, Nicholas. 2011. Building a lexical database with multiple outputs: Examples from legacy data and from multimodal fieldwork. *International Journal of Lexicography* 24(4). 463–472. DOI:10.1093/ijl/ecz027.
- Thieberger, Nicholas & Andrea Berez. 2012. Linguistic Data Management. In Nicholas Thieberger & Andrea Berez (eds.), *Oxford handbook of language documentation*, 90–118. Oxford: Oxford University Press. DOI:10.1093/oxfordhb/9780199571888.013.0005.
- Voegelin, Florence M., Stephen Wurm, Geoffrey O'Grady, Tokuchiro Matsuda & Charles F. Voegelin. 1963. Obtaining an index of phonological differentiation from the construction of non-existent minimax systems. *International Journal of American Linguistics* 29(1). 4–28.
- Walsh, Michael. 1997. How many Australian languages were there? In Darrell Tryon & Michael Walsh (eds.), *Boundary rider: Essays in honour of Geoffrey O'Grady*, 393–412. Pacific Linguistics C-136. Canberra: Pacific Linguistics.
- Zhou, Kevin & Claire Bower. 2015. Quantifying uncertainty in the phylogenetics of Australian numeral systems. *Proceedings of the Royal Society B: Biological Sciences* 282(1815). DOI:10.1098/rspb.2015.1278.

Claire Bower
 claire.bower@yale.edu