Developing Interactive Knowledgebases for Australian Aboriginal Languages — Malyangapa

Peter K. Austin

Department of Linguistics and Applied Linguistics

University of Melbourne. Parkville. Vic 3010 Australia

16 May 2002

1 Background[1]

This paper is a description of an interactive database model which I have developed to describe and analyse data on Australian Aboriginal languages. The approach is exemplified with materials from Malyangapa, an extinct language once spoken in western New South Wales. I have done this for a number of reasons. Firstly, I wish to test the concept of an interactive database for language documentation by applying it to a language for which the corpus is relatively self-contained but rich enough that a valuable documentation project results. Secondly, I wish to test the flexibility of current computer software to implement such a system. Thirdly, I wish to challenge the dominant paradigm in the academic study of Australian Aboriginal languages that has emphasised the writing of descriptive grammars, and to argue that an equally important goal of research should be the production of fully articulated and richly described corpora, especially for languages that are not longer spoken (which means the vast majority of Australian languages), which can provide an enduring legacy for future generations of researchers and language community members. Current developments in computer software, data models, and emerging cross-platform standards of representation such as XML, mean that this task is easily within the reach of all practising linguists.

The Malyangapa interactive knowledgebase is implemented in the SIL Shoebox 5.0 computer program as a series of ASCII text files that contain indexes and pointers between them, with rich linkage between the different data and metadata information. The system is designed to

[1] This paper was originally prepared for the Workshop on Australian Aboriginal Languages organised by Rachel Nordlinger, University of Melbourne, in March 2002. I am grateful to the late Stephen Wurm for access to his fieldnotes and recordings, and to Luise Hercus for sharing with me her detailed knowledge of the indigenous languages of eastern Australia. Needless to say, neither scholar is responsible for the details of my analysis of Malyangapa. Thanks also to SIL for making their powerful Shoebox program available to the linguistic world, and to David Nash and David Nathan for helpful comments on an earlier draft. This paper was written while I was on sabbatical leave from the University of Melbourne, which I thank for giving me the opportunity for an extended period of research leave.

enable users to move between different types of data and to fully explore the relationships between words and sentences, capturing as much information as possible about the Malyangapa materials. References to the original fieldnotes enable all instances to be located and checked. The general design for this interactive knowledgebase is one that has been successfully applied to other Australian Aboriginal language materials, some of which are much more complex and extensive (Austin 2001). It is my hope that it can serve as an example of what is possible using current computer software in terms of providing richly annotated documentation of a closed corpus of materials on a language that is no longer spoken. A planned future phase of this work will see export of the data files in a more widely used format such as extensible markup language (XML) for distribution among interested scholars and community members.

The Malyangapa language was traditionally spoken in far western New South Wales. Materials on the language were recorded by Stephen Wurm in 1957 with Hannah Quayle, born near Yancannia in about 1875, and Alf Barlow. This material consists of 48 pages of fieldnotes (24 double-sided sheets) plus a brief tape-recording, amounting to 386 sentences and containing a vocabulary of 358 items. Luise Hercus also did some recording with George Dutton in the mid-sixties on Malyangapa; Jeremy Beckett had previously worked with him on social and cultural traditions. Luise Hercus also recorded Laurie Quayle, son of Hannah Quayle, checking some of the earlier materials. He passed away in 1976, and with his death the language became extinct. The Hercus and Beckett data has not yet been incorporated into the current data files.

Malyangapa is relatively closely related to the neighbouring Wadikali and Yardliyawarra languages forming with them the Yardli group (Austin and Hercus 2002). There is limited data on these other two languages, and it is planned in future to include this material for comparative purposes.

2 Research on Australian Aboriginal languages

The academic study of Australian Aboriginal languages has been dominated over the past thirty odd years by a research paradigm that saw the primary goal of linguistic research as the writing of descriptive grammars in a more-or-less standard structuralist format covering phonology, morphology and syntax, with some brief mention of sociolinguistics and peculiarities of language use such as special speech styles for taboo contexts[2]. The grammars that emerged from this tradition were typically single-volume compact studies of 300-400 pages, or in the case of truly moribund languages, a 50-100 page sketch following the format of the *Handbook of Australian Languages* (Dixon and Blake 1978-2000). Almost completely lacking have been comprehensive dictionaries and text collections (as pointed out in Austin 1991), and there has been no concern for corpus-based approaches to linguistic research.

[2] This model has been more generally espoused for *all* languages by Dixon, who writes: “if every person who called themselves a linguist settled down to provide a full description of a single previously undescribed language, then he or she would justify the title” (Dixon 1994:229, cf. Dixon, 1997:135ff “What every linguist should do”). What Dixon means by ‘full description’ is a single volume descriptive grammar.

There are a number of problems that arise as a direct result of this research and publication paradigm:

- the resulting grammars are necessarily limited both in terms of their depth and breadth of coverage. There is usually insufficient detail to enable testing of the author's claims, and no concern with *tokens* of language forms or categories that would enable checking of distributions, frequencies, or other token patterns in a corpus. This point has been forcefully made by Heath (1984:5): "The extensive exposition of textual citations and statistics in many chapters of this volume may strike some readers as reflecting a personal fetish of mine. While this may be true, it is a fetish I would defend. ... it gives a more patient (or more skeptical) reader a feeling for the raw data which underlie the analysis and the opportunity to "cross-examine" the author by going directly to the data. It also encourages readers with highly specialised interests, or with a different theory of language, to discover new patterns which I overlooked or did not have space to discuss". He goes on to say: "my concern with documentation reflects my own sad experiences as a reader of other linguists' grammars, which almost never provided me with the information I wanted to undertake my own (re-) analysis of the language in question. It also reflects my experience that most published grammars are based on material obtained in unreliable direct-elicitation (sentence-translation) sessions, and/or utterances which were produced by the linguist with or without "confirmation" from a native informant"[3];
- there is a lack of appreciation among linguists operating in this paradigm that grammar writing is a tertiary level of language documentation — the primary documentation is the audio, video and other recorded media together with the original fieldnotes and transcriptions made by the researcher in collaboration with native speakers[4]. This was pointed out almost thirty years ago by Goddard (1973:86): "most descriptive linguists probably feel that their finished grammars have a greater validity, in some sense, than their raw fieldnotes. But **the field notes are the primary documents, the nearest thing to the actual speech events there is**, and they should always ultimately be deposited in a suitable library or public archive, together with explanatory information on dates of fieldwork, relevant characteristics of informants, changing transcriptional conventions, and indexes. Only if this practice become more general can the present situation be improved, in which numerous cases of possible informant errors, artifacts of elicitation methods, misprints, and miscopyings remain

[3] An example of this problem is the phenomenon found in a large number of eastern Australian languages and described in Austin 1997 whereby a single verb affix has either applicative or causative effect depending on the semantics of the root to which it is attached. Most of the grammars of the relevant languages (which typically fail to describe the split) contain one or two examples of the phenomenon but do not examine it in detail — only by combing through dictionaries (where they exist) and by cross-linguistic comparison is it possible to uncover the semantic range of this covert category. No corpora are available for examination.

[4] The secondary level is the stage between fieldnotes and grammar writing when fieldnotes are reworked and retranscribed, example sentences are selected, analysed and glossed, paradigms are assembled, and the linguist 'works out' the structure of the language.

forever undetected or in doubt because of the impossibility of checking them against the primary documents” [emphasis added, PKA];

- as a result of the death of most of the indigenous languages of Australia, including their almost total extinction in the southern part of the country, research on these languages has entering a new phase where corpora are necessarily static[5] and analysis of the languages will need to draw on traditional philological approaches to extinct languages. Important in such work will be detailed annotations of primary documents, and careful comparisons of all available sources, including pre-modern materials collected by amateurs or less trained observers (see Austin and Tindale 1986 for one example of this; Blake’s publications on the languages of Victoria over the past 10 years are also another instance). A shift of emphasis from grammar writing to corpus documentation will need to take place as a result. Goddard (1973:86) argued the case for this orientation for Americanist linguists a generation ago: “the linguist who has a philological approach looks not only to the past but also to the future; **he must be concerned with minimizing the problems which the documents he produces will cause his successors. This means making explicit in the fullest practicable way all the information about a form or a corpus that a future investigator might seek.** It is impractical, of course, to give full particulars for every form ever cited in print. But it is possible to do more along these lines than most Americanists have been accustomed to in the past.” [emphasis added, PKA]. Goddard (1973:88) argues further that there is a need to focus: “on the fact that there are and will be only a finite number of documents recording the native languages of North America. **It is necessary to make the fullest and most careful use of what there is, and to exercise the greatest diligence in preserving this corpus for the future in the most useful possible form.**” [emphasis added PKA]. Exactly the same arguments apply to current research in Australia.

For these various reasons, the time is ripe for a shift in direction in Australian language documentation back to examination and analysis of primary documentation, including provision of fully detailed data and metadata descriptions for the existing corpora. Fortunately, computer software of various types is now becoming available to make this task easier than it has ever been in the past. This paper is intended as an example of how it is possible to implement corpus documentation in a flexible database format.

3 Software

For this project I have chosen to use the Shoebox 5.0 program developed by the Summer Institute of Linguistics (www.sil.org) which is a general tool for information management, oriented towards linguistic data. Shoebox maintains a series of ASCII files in a standard file

[5] Some southern languages, such a Kaurna or Kamilaroi, have expanding corpora as a result of language revival projects and associated language engineering. This material is however qualitatively different from data collected from actual speakers or rememberers of the languages.

format where each record begins with a data type code (in the form \x) followed by the data and a carriage return. Typically there is a (bilingual) lexicon listing all morphemes and a database of glossed sentences listing sentence forms, morpheme-by-morpheme glosses and free translations. Shoebox provides a number of functions such as semi-automatic filling-in of information (‘interlinearisation’ or ‘glossing’), generating a wordlist (a list of all items in a chosen data field with an associated count and index list of occurrences), generating a concordance (a listing of chosen data items with their preceding and following context), and automatic numbering of a data set. Shoebox also serves as a viewer of lexicon, texts, wordlist and concordance, allowing the user to have multiple windows open on screen. It has a ‘jumping’ function that instantiates hypertext links by search and pattern match (rather than anchors and pointers as in HTML), linking data in different files and allowing the user to move between material in linked fields in related data files.

Shoebox provides export into XML format but this is weakly developed and does not capture the hierarchical structure of glossed texts which is only indirectly represented in Shoebox by virtue of vertical alignment on screen (encoded as spaces within the text data files).

4 Data structures

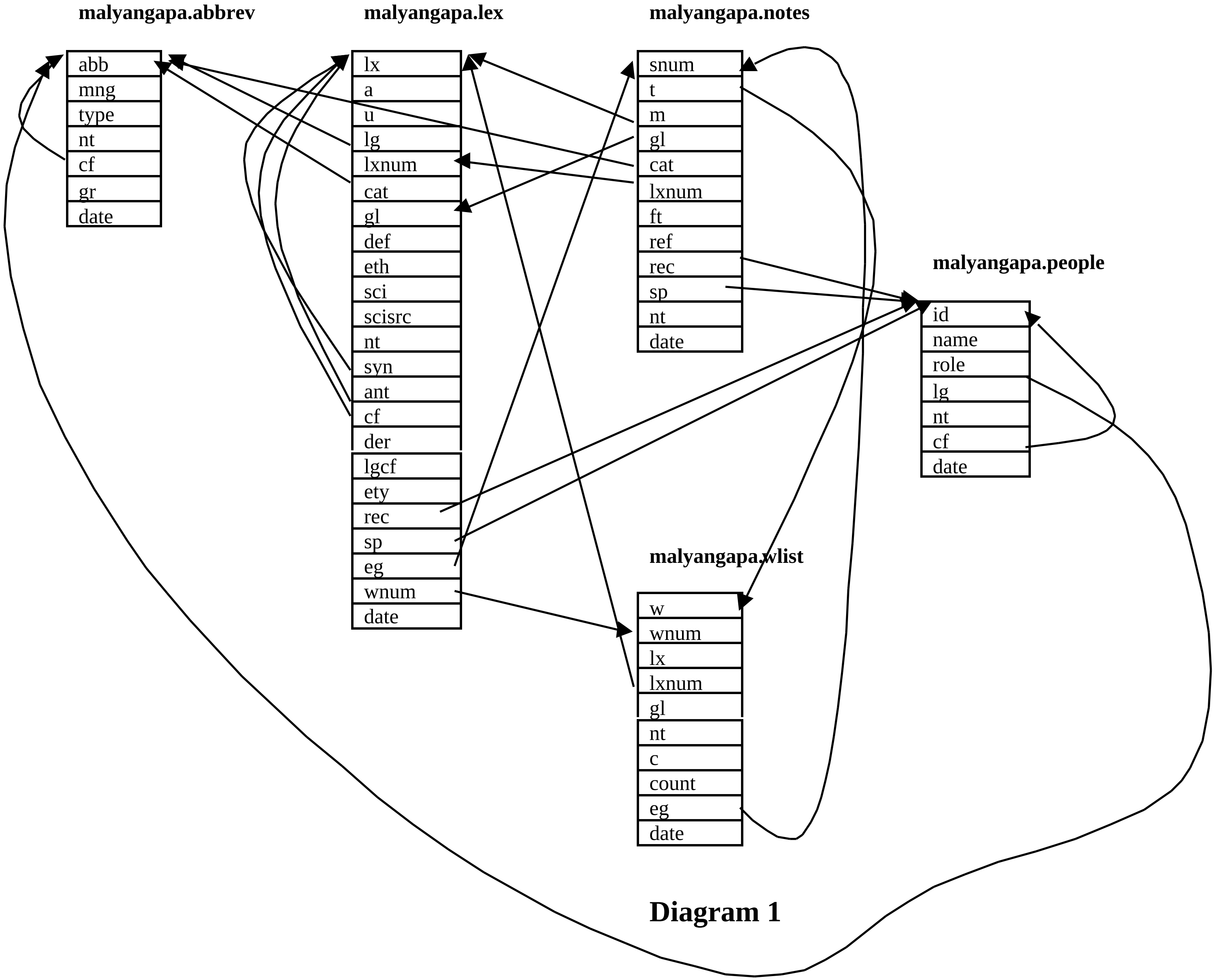
The core of the Malyangapa knowledgebase is a set of files of two types: data files and metadata files. The data files capture knowledge about lexical information and sentence structure. The metadata files deal with metalinguistic terminology (a controlled vocabulary encoded as abbreviations), and information about the people involved in the project (as speakers and recorders). Since there is only one set of notes from Wurm’s fieldwork and all the material is elicited, there is no metadata recorded on such topics as source genre, date and time of recording, transcriptional practices, or other background information[6].

The data and metadata files are linked by exploration pathways (“jump paths”), hypertext links between named fields within one file and fields in a related file. Clicking on an item in a jump path field updates the related file window with the linked data (thus clicking on a morpheme in a glossed sentence updates the lexicon window with that morpheme’s record, or clicking on a sentence number in the wordlist file displays the related sentence in the glossed text window). The exploration pathways are shown in Diagram 1.

Diagram 1 goes here

It is also possible to encode these links as anchors and pointers (as with HTML) by transforming the Shoebox files into marked up data files for viewing by other software such as web browsers. A sample of what this might look like can be seen in the Appendix. I have not yet written the necessary routines for converting the Shoebox files to such a format.

[6] Such metadata is stored for other languages I have been working on (see Austin 2001).



malyangapa.abbrev

abb
mng
type
nt
cf
gr
date

malyangapa.lex

lx
a
u
lg
lxnum
cat
gl
def
eth
sci
scisrc
nt
syn
ant
cf
der
lgcf
ety
rec
sp
eg
wnum
date

malyangapa.notes

snum
t
m
gl
cat
lxnum
ft
ref
rec
sp
nt
date

malyangapa.people

id
name
role
lg
nt
cf
date

malyangapa.wlist

w
wnum
lx
lxnum
gl
nt
c
count
eg
date

Diagram 1

Note: Arrows indicate hypertext links

4.1 Data files

There are two types of language data files: those dealing with lexical information and those dealing with sentence analysis. The lexical material is stored in two files:

- a lexicon that gives lemmas in Malyangapa with their glosses and definitions in English
- a wordlist that lists all occurrences of word forms in the sentences, together with their lemmas, the number of occurrences, and references to the sentences within which the items are found. This wordlist was generated using Shoebox's wordlist and glossing functions

The lexicon has the following structure:

malyangapa.lex

lx the Malyangapa lemma spelled in a practical orthography

a alternative forms of the lemma, used for morphophonemic alternations[7]

u underlying forms of morpheme combinations where the surface form is not simple concatenation of lemmas, or the 'shortest match' principle adopted by Shoebox gives the wrong morphological parse[8]

lg language of the lemma [currently Malyangapa but to be expanded to the other Yarli languages]

lxnum unique numerical identifier for lexical entries

cat morpho-syntactic category

gl gloss, usually a single English word, or a sequence of English words separated by periods [used in sentence glossing and finderlist generation]

def definition of the lemma in English

eth ethnographic information about the referent

sci scientific name for plants and animals

[7] Shoebox uses a simple concatenative morphology model. If the lemma has allomorphs these must be listed in an 'a' field, which Shoebox is set up to look at before using the 'lx' form.

[8] We give two examples of this from the Malyangapa lexicon: firstly, the surface form pulanha at the end of a verb is ambiguous between the suffix lemma -pulanha 'third person dual transitive object pronoun' (X *verb* them two] and the sequence of suffixes -pula 'third person dual transitive subject' plus -nha 'third person singular transitive object' [They two *verb* him/her/it]. Under the lemma -pula the field a contains -pulanha and the field u contains –pula -nha. A second example occurs under yuRinga 'be deaf'; here a lists yuRingarntayi and u lists yuRinga -rnta –yi ['be.deaf -pres -emph'] because there are two other morphemes yuRi 'ear' and -ngarnta 'past' which Shoebox would otherwise associate with yuRingarnta using its 'longest match' principle.

scisrc source for scientific identification

nt any other notes on form, meaning or usage

syn form of any synonymous term

ant form of any antonymous term

cf form of any other related term[9]

der list of any derived forms of the lemma

lgcf cross-reference for lemmas in other related languages [not currently implemented]

ety etymological information about the lemma including reconstructed proto-forms [not currently implemented]

rec initials of person who recorded the entry [cross-reference to **people** file]

sp initials of speaker who provided the entry [cross-reference to **people** file]

eg reference number for example sentence in which the lemma occurs [cross-reference to **notes** file]

wnum reference number for word forms containing the lemma [cross-reference to **wlist** file]. This field gives access to all the word forms (and hence all the sentence examples) actually occurring in the corpus for any individual lemma[10]

date date stamp for entry [generated by Shoebox]

[9] Further sense relations could be distinguished by setting up additional fields (eg. metonymy, hyponymy) if these can be determined from the data. For Malyangapa the material is insufficient to enable this degree of semantic specification.

[10] The wnum reference to the wlist file gives the user indirect access to all occurrences of a lemma in the notes through the lists of wordform sentence numbers generated by Shoebox. The eg field in the lexicon can contain a list of just those occurrences selected by the linguist analyst to illustrate the lemma.

An example of a lexical entry is:

```
\lx    kurntu
\a
\u
\lxnum 086
\lg    Ml
\cat   n
\gl    many
\gl    much
\def   much, many, plenty
\eth
\sci
\scisrc
\nt
\syn   marru
\ant
\cf
\der
\lgcf
\ety
\rec   SW
\sp    HQ
\eg    036
\wnum  113; 114; 115; 116
\date  12/Mar/2002
```

The Malyangapa wordlist is a full list of all wordforms occurring in the sentence examples, together with a count of the number of tokens of each wordform and a list of the example sentence reference numbers in which the wordform occurs. The data in this file is generated using Shoebox’s wordlist function run over the example sentences file (**notes**, described below), and then numbered with a unique wordform identifier (using Shoebox’s number function). Analysis and lemmas of the wordforms was generated by using Shoebox’s glossing (parsing) facility against the lexicon. The relevant wordform identifier numbers were then written back into the wnum field for each lemma in the lexicon. The value of the wordlist is that it gives all occurring forms of lemmas, together with their token frequencies, and hypertext links back to the sentences from which they are extracted.

malyangapa.wlist

w wordforms occurring in the sentence materials

wnum unique numeral identifier for wordforms [generated by Shoebox number function]

lx lemmas for wordform [generated by Shoebox parsing with cross-reference to **lex** file] lxnum lemma identification numbers [generated cross-reference to **lex** file]

gl glosses for lemmas [generated cross-reference to **lex** file]

nt notes

c number of wordform tokens in sentence data [generated by Shoebox wordlist function]

count count of word form occurrences in decimal format [generated by Shoebox wordlist function, can be sort field for frequency analysis]

eg reference number for example sentence in which the wordform occurs [generated by Shoebox, cross-reference to **notes** file]

date date stamp for entry [generated by Shoebox]

An example of a wordlist entry is:

```
\w     ngapanga
\wnum  226
\lx    ngapa -nga
\lxnum 076   -075
\gl    water -loc

\c 3
\count 000003
\eg 071; 207; 306
\date 17/Feb/2002
```

Malyangapa sentence data is stored in a single file that contains surface sentence forms[11], their aligned morpheme-by-morpheme glosses (generated using Shoebox's glossing (parsing) function against the lexicon), free translations of each sentence, identification of speaker and recorder, and any additional notes (often comments on problems of analysis or the surface forms).

Unfortunately, Shoebox does not support media but a nice complement to this file would be scanned images of Wurm's original fieldnotes keyed to the page reference identifier, and links to digitised sound files (Wurm made tape recordings of part of his material). I hope that this can be implemented in the future.

malyangapa.notes

snum unique numerical identifier for each sentence [generated by Shoebox]

t surface form of sentence in a practical orthography

[11] The *t* (for 'text') field is transcribed in a practical phonemic orthography that closely resembles but is not identical to Wurm's transcription. In work on other languages I have a separate field for the original transcription, thus recording it separately from my analysis of the surface sentence form.

m aligned morphemic representation of wordforms [generated by Shoebox glossing (parsing) into lemmas using the lexicon]

gl aligned English gloss of each morpheme [generated by Shoebox glossing]

cat aligned syntactic category of each morpheme [generated by Shoebox glossing]

lxnum aligned lemma number of each morpheme [generated by Shoebox glossing]

ft English free translation of sentence

ref reference to page number and sentence number in fieldnotes

rec initials of person who recorded the sentence [cross-reference to **people** file]

sp initials of speaker who provided the sentence [cross-reference to **people** file]

nt notes on sentence

date date stamp for sentence [generated by Shoebox]

An example of the data in the sentence file is:

```
\snum  386
\t     kata     wanthayi     yiniki
\m     kata     wantha -yi   yiniki
\gl    cockatoo where  -emph that
\cat   n        n      -suff dem
\lxnum 362      071    -034  035
\ft    Where is that cockatoo?
\ref   SW2/1Bs05
\rec   SW
\sp    AB
\nt    Wurm's gloss "Look at that red and white cockatoo"
\date  18/Jul/2001
```

4.2 Metadata files

The metadata files contain non-linguistic background information about the data in the lexicon and sentence collections. For Malyangapa this consists of a list of all the abbreviations and material relating to the recorders and speakers who contributed data.

malyangapa.abbrev

abb unique abbreviation

mng meaning of the abbreviation, usually a short description in lay terms of the functions of syntactic category labels type type of abbreviation, eg. category, language name

nt notes about the abbreviation

cf cross-reference to other related abbreviations

gr cross-references to grammar of Malyangapa [not yet implemented]

date date stamp for item [generated by Shoebox]

An example of an entry in this file is:

```
\abb  vtr
\mng  transitive verb
\type sub-category
\nt   transitive verbs are a sub-category of verbs; they take a
      transitive subject argument in ergative case and a
      transitive object argument in accusative case.
\cf   vi, vdi
\gr
\date 12/Mar/2002
```

malyangapa.people

id unique abbreviation for each person [generally initials of firstname and lastname]

name person’s name

role role of person in language documentation

lg language spoken (if Aboriginal language speaker)

nt notes on individuals

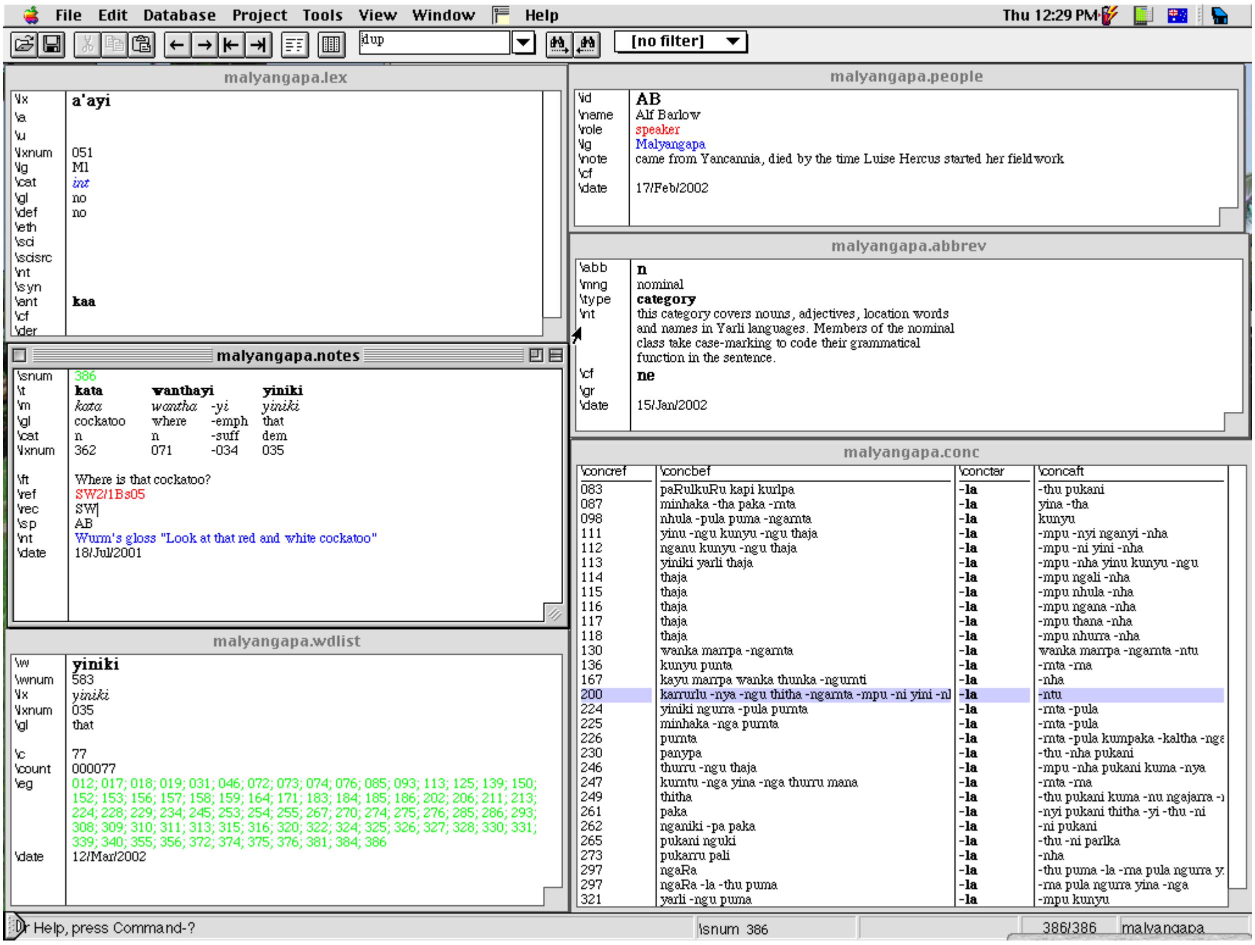
cf cross-reference to other individuals in this file

date date stamp for item [generated by Shoebox]

A sample entry from this file is:

```
\id   LQ
\name Laurie Quayle
\role speaker
\lg   Malyangapa
\note consultant for Luise Hercus, son of HQ
\cf   HQ
\date 17/Feb/2002
```

Here is a screen shot showing a sample of the files as displayed within Shoebox — the windows display lexicon (top left), notes (middle left), wordlist (bottom left), metadata (top right and middle) and a concordance (bottom right). User interaction is by right-click (or option-click on the Macintosh) within the linked data fields (shown as arrows in Diagram 1).



malyangapa.lex

\x	**a'ayi**
\a	
\u	
\xnum	051
\g	M1
\cat	*int*
\gl	no
\def	no
\eth	
\sci	
\scisrc	
\nt	
\syn	
\ant	**kaa**
\cf	
\der	

malyangapa.notes

\snum	386			
\t	**kata**	**wanthayi**		**yiniki**
\m	*kata*	*wantha*	*-yi*	*yiniki*
\gl	cockatoo	where	-emph	that
\cat	n	n	-suff	dem
\xnum	362	071	-034	035
\ft	Where is that cockatoo?			
\ref	SW2/1Bs05			
\rec	SW|			
\sp	AB			
\nt	Wurm's gloss "Look at that red and white cockatoo"			
\date	18/Jul/2001			

malyangapa.wdlist

\w	**yiniki**
\wnum	583
\x	*yiniki*
\xnum	035
\gl	that
\c	77
\count	000077
\eg	012; 017; 018; 019; 031; 046; 072; 073; 074; 076; 085; 093; 113; 125; 139; 150; 152; 153; 156; 157; 158; 159; 164; 171; 183; 184; 185; 186; 202; 206; 211; 213; 224; 228; 229; 234; 245; 253; 254; 255; 267; 270; 274; 275; 276; 285; 286; 293; 308; 309; 310; 311; 313; 315; 316; 320; 322; 324; 325; 326; 327; 328; 330; 331; 339; 340; 355; 356; 372; 374; 375; 376; 381; 384; 386
\date	12/Mar/2002

malyangapa.people

\id	**AB**
\name	Alf Barlow
\role	speaker
\lg	Malyangapa
\note	came from Yancannia, died by the time Luise Hercus started her field work
\cf	
\date	17/Feb/2002

malyangapa.abbrev

\abb	**n**
\mng	nominal
\type	**category**
\nt	this category covers nouns, adjectives, location words and names in Yarli languages. Members of the nominal class take case-marking to code their grammatical function in the sentence.
\cf	**ne**
\gr	
\date	15/Jan/2002

malyangapa.conc

\concref	\concbef	\conctar	\concaft
083	paRulkuRu kapi kurlpa	**-la**	-thu pukani
087	minhaka -tha paka -rnta	**-la**	yina -tha
098	nhula -pula puma -ngarnta	**-la**	kunyu
111	yinu -ngu kunyu -ngu thaja	**-la**	-mpu -nyi nganyi -nha
112	nganu kunyu -ngu thaja	**-la**	-mpu -ni yini -nha
113	yiniki yarli thaja	**-la**	-mpu -nha yinu kunyu -ngu
114	thaja	**-la**	-mpu ngali -nha
115	thaja	**-la**	-mpu nhula -nha
116	thaja	**-la**	-mpu ngana -nha
117	thaja	**-la**	-mpu thana -nha
118	thaja	**-la**	-mpu nhurra -nha
130	wanka marrpa -ngarnta	**-la**	wanka marrpa -ngarnta -ntu
136	kunyu punta	**-la**	-rnta -rna
167	kayu marrpa wanka thunka -ngurnti	**-la**	-nha
200	karrurlu -nya -ngu thitha -ngarnta -mpu -ni yini -nl	**-la**	-ntu
224	yiniki ngurra -pula purnta	**-la**	-rnta -pula
225	minhaka -nga purnta	**-la**	-rnta -pula
226	purnta	**-la**	-rnta -pula kumpaka -kaltha -nga
230	panypa	**-la**	-thu -nha pukani
246	thurru -ngu thaja	**-la**	-mpu -nha pukani kuma -nya
247	kurntu -nga yina -nga thurru mana	**-la**	-rnta -rna
249	thitha	**-la**	-thu pukani kuma -nu ngajarra -
261	paka	**-la**	-nyi pukani thitha -yi -thu -ni
262	nganiki -pa paka	**-la**	-ni pukani
265	pukani nguki	**-la**	-thu -ni parlka
273	pukarru pali	**-la**	-nha
297	ngaRa	**-la**	-thu puma -la -rna pula ngurra y
297	ngaRa -la -thu puma	**-la**	-rna pula ngurra yina -nga
321	yarli -ngu puma	**-la**	-mpu kunyu

For Help, press Command-?

\snum 386

386/386

malyangapa

5 Conclusions

The study of Australian Aboriginal languages is entering a new phase that requires a focus on detailed corpus documentation and analysis. Powerful software tools exist that enable such documentation to be accomplished, and this paper has presented an example of a data model applied to materials on one extinct language that shows the value of such an approach. I have also demonstrated how the Shoebox program can serve as a hypertext viewer of a quite complex set of data and metadata that enables users to fully explore an analysed and annotated corpus. In future research I hope to export this data into a format that allows viewing and interaction by general purpose tools such as web browsers.

6 References

Austin, Peter K. 1991 ‘Australian Aboriginal language studies’, In Michael Clyne (ed.) *Linguistics in Australia: trends in research*, 55-74. Canberra: Australian Academy of Humanities and Australian Academy of Social Sciences.

Austin, Peter K. 2001 Developing an interactive knowledgebase for the Mantharta languages, Western Australia. Presented at Department of Linguistics and Applied Linguistics, University of Melbourne, March 2001, and University of Utrecht, February 2002.

Austin, Peter K and Norman B. Tindale 1986 ‘Emu and brolga, a Kamilaroi myth’. *Aboriginal History*, 9:8–21.

Austin, Peter K. and Luise A. Hercus 2002 ‘The Yarli languages’. Paper presented at International Conference on Historical Linguistics, Melbourne August 2001, to appear in the proceedings.

Dixon, R.M.W. 1994 *Ergativity*. Cambridge: Cambridge University Press

Dixon, R.M.W. 1997 *The rise and fall of languages*. Cambridge: Cambridge University Press

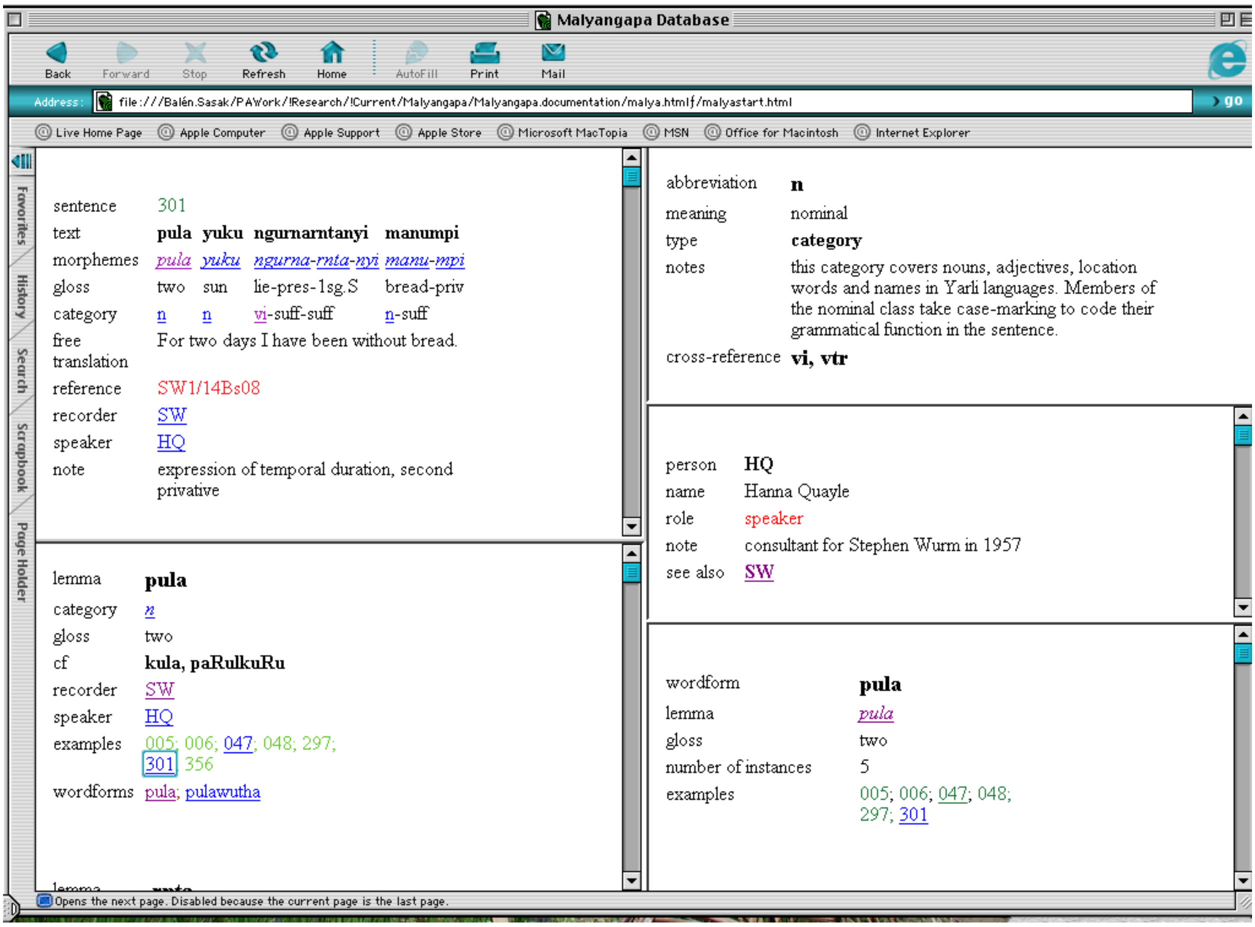
Dixon, R.M.W. and Barry J. Blake 1978-2000 *The Handbook of Australian Languages*. Vol 1-5. Canberra: ANU Press and Melbourne: Oxford University Press.

Goddard, Ives. 1973. ‘Philological approaches to the study of North American Indian languages: documents and documentation’. In Thomas A. Sebeok, ed. *Current Trends in Linguistics, Vol 10.* The Hague: Mouton.

Heath, Jeffrey 1984 *A Nunggubuyu grammar.* Canberra: Aboriginal Studies Press.

7 Appendix — Web browser files

This is a mockup view of the Shoebox database as HTML tables with hypertexts links as viewed by Internet Explorer.



file:///Balén.Sasak/PAWork/!Research/!Current/Malyangapa/Malyangapa.documentation/malya.htmlf/malyastart.html

sentence	301			
text	**pula**	**yuku**	**ngurnarntanyi**	**manumpi**
morphemes	*pula*	*yuku*	*ngurna-rnta-nyi*	*manu-mpi*
gloss	two	sun	lie-pres-1sg.S	bread-priv
category	n	n	vi-suff-suff	n-suff
free translation	For two days I have been without bread.			
reference	SW1/14Bs08			
recorder	SW			
speaker	HQ			
note	expression of temporal duration, second privative			

lemma	**pula**
category	*n*
gloss	two
cf	**kula, paRulkuRu**
recorder	SW
speaker	HQ
examples	005; 006; 047; 048; 297; 301 356
wordforms	pula; pulawutha

abbreviation	**n**
meaning	nominal
type	**category**
notes	this category covers nouns, adjectives, location words and names in Yarli languages. Members of the nominal class take case-marking to code their grammatical function in the sentence.
cross-reference	**vi, vtr**

person	**HQ**
name	Hanna Quayle
role	speaker
note	consultant for Stephen Wurm in 1957
see also	**SW**

wordform	**pula**
lemma	*pula*
gloss	two
number of instances	5
examples	005; 006; 047; 048; 297; 301

Opens the next page. Disabled because the current page is the last page.