International Journal of Lexicography, Vol. 24 No. 4, pp. 463–472
doi:10.1093/ijl/ecr027 Advance access publication 10 September 2011

BUILDING A LEXICAL DATABASE WITH MULTIPLE OUTPUTS: EXAMPLES FROM LEGACY DATA AND FROM MULTIMODAL FIELDWORK

Nick Thieberger: *University of Melbourne (thien@unimelb.edu.au)*

Abstract

The creation of reusable lexical database files, based in fieldwork or arising from historical research, benefits from conformance to established standards which then greatly increases the enduring usability of the lexicon, and its later ability to link to external objects, including media. All linguistic analysis benefits from the close relationship between primary recordings and a textual corpus, but a dictionary can also benefit from links to media in the use of playable example sentences and citation forms of headwords. In this paper several examples will be used to illustrate that not all linguists want to deal with the tools required to take advantage of these methods, so, in some cases, they are better off seeking advice and assistance in advance of building the database or in its later conversion to output formats.

1. Introduction

This article discusses a process for constructing dictionaries that I had hoped would have needed no such description at this time, some thirteen years since Himmelmann's (1998) formative article on language documentation. With all of the activity that appears to be taking place since then (in the form of large funding programs and many research projects) it may have been the case that we now take it for granted that all documentation should include a media corpus, that various data sources can be made to work together, and that outcomes of linguistic work be created in an archival form with derived forms for presentation. However, in my experience it is not the case that most linguistic fieldwork is taking advantage of methods that are now available for the creation of material resulting from fieldwork. This paper is written by a linguist, not a programmer, and is aimed at explaining some of the processes

that linguists can use to structure better dictionaries, suggesting that it is useful to have a service offered by those experienced in the production of dictionaries from lexical databases to which linguists can submit their lexicons for processing.

It is increasingly clear that new research methods require the creation of data in forms that will endure, allowing their reuse in future research. For linguists involved in language documentation there is an even greater need for these methods to be adopted, as we are creating what may be the only records for a language. This has implications not only for the claims we make about the properties of that language, but also for the representation of the people we record and their inscription in a more general understanding of the nature of human diversity. Our records will also provide what may be the only reflection of local identity, available to future generations who will seek a connection to their ancestral language and its associated knowledge systems. For lexicographers building dictionaries of these languages, the challenge is to create lexical databases that can be reused to allow new versions of dictionaries to be created with minimal work, and to allow enrichment of the content over time.

In this paper I discuss two models for the creation of lexical databases, one based in current fieldwork in Vanuatu and resulting in a multimodal dictionary, and the other based in historical materials which are being used for language revitalization. These examples show that, with some work, the same structures used in creating current multimodal dictionaries can be retrofitted to existing dictionaries. Before any of the present methods were available, a similar philosophy to that espoused in this article was being practised in the 1970s onwards by Robert Hsu at the University of Hawai'i. He estimates (pers.com.) that he assisted with the processing of hundreds of dictionaries, having developed Lexware (Hsu 1985), software that used a field-oriented markup very similar to the one that would later be used by Shoebox and Toolbox. Lexicographers worked on the content of their dictionary and provided text files to Hsu for processing, building well-structured lexical databases and not needing to deal with the technology of data conversion.

My ongoing work on the dictionary of South Efate (Oceanic, Central Vanuatu) has relied on a media corpus built incrementally over the course of fieldwork. Once this dataset was established, it was logical to extend the same model to creating a shared data structure and a set of images that could be used by several dictionaries of Vanuatu. This necessitated the conversion from a legacy (MS Word) encoding of one dictionary, and the use of standard formats for the others. The question is: how can we bring existing dictionaries into the new world of interoperating and reusable data structures, thus facilitating links to external documents (including texts and audio-visual media)?

In the same vein, I discuss the process of building a paper-based dictionary for Ngarrindjeri (Pama-Nyungan, Australia) using a standard database program (FileMaker Pro), allowing users to continue with their preferred data

entry system, but with export routines to Toolbox permitting neatly formatted dictionaries to be produced. Each of these models is appropriate to particular contexts, and each results in structured lexical data suitable for subsequent reuse, which, in addition to building a crafted dictionary, has to be a central concern of dictionary construction for small and endangered languages.

A broader issue that arises from this discussion is the need for the provision of advice and support to linguists developing dictionaries, acknowledging that they can work in various ways based on their own experience, as long as they are aware of the standard formats their dictionaries need to conform to in order to be reusable in future.

2. Building a media corpus for South Efate

The advantages of well-constructed data underlying a dictionary include the ability to properly organise textual outputs, including various arrangements of the lexical database (dictionary, topical list, reversal, and so on) and these can be rendered as printed documents or created as standalone dictionaries or as web-pages. Linking to multimedia on computers has been an option since the 1980s but has had slow uptake among linguists. In part this is because of the difficulty of applying appropriate methods in the absence of dedicated tools. Recent work in the paradigm of language documentation has moved in this direction, see for example Cablitz (this volume and Cablitz et al. 2007) who discuss the use of Lexus to create an elaborate multimedia dictionary for Marquesen, based on a collection made using the tools provided by the Max Planck Institute in Nijmegen. LexiquePro is another popular tool that allows links between data types to be instantiated and their website (http://lexiquepro.com) lists a number of lexicons and dictionaries made using the software.

My own dictionary of South Efate is a work in progress with drafts made available periodically as paper dictionaries and as an online version with images and audio. The process for construction of the lexicon described in this paper has allowed new versions to be generated periodically, always maintaining their links to various media files and thus allowing a multimedia rendition of the dictionary. I was concerned to build a reusable corpus of texts (cf. Thieberger 2006) that could be cited in the grammatical analysis and reused in other ways. The textual corpus, linked to the primary recordings, allowed me to refine the transcription based on my growing understanding of the structure of the language, and to corroborate my analysis by immediate reference to the recorded texts (see Thieberger 2004 for a discussion of this method). This corpus of transcripts provided a concordance of some 119,000 word tokens which were then able to be checked against headwords in the lexicon. Texts exported into a Toolbox format retained their time alignment to the primary media and were then interlinearized, feeding the toolbox lexicon with the citation form or lemma. The primary media was located in a repository

(PARADISEC http://paradisec.org.au) that provided persistent identification for the files. Such persistence is integral to the kind of eResearch methods discussed here as it ensures that citations of data can be resolved to the same dynamic data whenever a researcher needs to resolve them, now and into the future.

The lexical database stores—in addition to its definitions, example sentences and categorial information—references to media related to lexical items: names of image files, media files and timecodes within media files. The images would typically aid in identifying the headword (plants, animals and so on), while the media files are playable example sentences, or headwords spoken in isolation specifically for the dictionary presentation. Entering these references can be done incrementally as the database is developed, or can be subsequently automated as will be discussed below.

3. Getting spoken headwords into the lexical database

Recording and inserting spoken headwords into a dictionary is not a trivial task, especially once the dictionary includes hundreds or several thousands of headwords. While the details of the method discussed here are likely to change as new tools emerge, it nevertheless illustrates the underlying principle of creating the data once and then allowing it to be used in multiple outputs. To begin, a list of headwords is extracted from the dictionary and prepared as a text file, printed and provided to a speaker as a script to read. In what is a fairly tedious process for them, speakers are then recorded reading the headwords, typically three times each to allow for varying intonation. These recordings are then time-aligned by pasting the script into tools like Elan or Transcriber, resulting in a new file of the text plus timecodes associated with the start and end of each headword. Exported in a simple format of text with time codes, this can be imported into software such as Audacity, which assigns labels to the audio file based on the timecodes, with each label named as per the headword. Audacity can then automatically segment the media file at these labels and creates individual files named by their labels. Playing these sounds is then a matter of coding in the delivery format, so, for example, each headword could be coded to call a file named as per the headword plus a media extension (perhaps .mp3). So, for the headword *afsak* 'turtle' we could create a tag in html that includes reference to a file named 'afsak.mp3' (as shown in Figure 1), and do this automatically over the whole file of several thousand headwords.

```
<A HREF="afsak.mp3">afsak</A></span><span class="lpPartOfSpeech">n. </span><span
class="lpGlossEnglish">turtle (generic).</span></p>
```

Figure 1: Example HTML tag generated by copying the headword and adding the extension '.mp3'.

```
<span class="filename">NT1-98015-A</span> <span class="start" >5.868</span> <span
class="end ">16.377</span><br/> <span class="text erk"><a
href="javascript:jumpToTimeoffset(5.868,16.377);"> afsak </a></span><br/> </p>
```

Figure 2: Example HTML code to call a segment of a larger sound file.

HTML5 offers an alternative to snipping thousands of mp3 files for use in an online dictionary. In HTML5 it is possible to call timecoded segments from within large media files, so the same procedure would apply as described above for MP3 files, but, in this case the reference is to a media file and timecodes (as shown in Figure 2). The media file has to be transcoded to a suitable format (e.g., OGG, H.264, WebM, MP4) and placed into a server, which then allows all or some of the media to be delivered without the need for streaming server software (like Flash or QuickTime).

In both of the methods just described it would make little sense to manually add audio references to the lexicon, especially when it becomes a significant set of data to be dealing with (in this case there are some 2,800 headwords). By paying attention to filenaming conventions (that is, maintaining consistent and unchanging filenames) and using regular expression processes it has been possible to automate the creation of links based on the structure of the lexical database, copying the headword (because it is identified as being a headword in the structure of the document) into a new field and appending the relevant coding to call an MP3 file, or the timecoded section of the media file in which that word is spoken.

4. Converting legacy dictionaries into reusable lexical databases

Dictionaries created in word-processors, like MS Word, provide little scope for reuse, but, as digital files they still have more possibility for repurposing than do paper-only dictionaries. In a project for the Vanuatu Cultural Centre, several existing 'legacy' dictionaries are to be included in a computer-based representation linking text to sounds and images. As the languages represented are from similar and not too geographically dispersed areas, they will each include names for the same flora and fauna, so a bank of images can be shared. The initial impetus for this work came from Catriona Hyslop's project to record two languages, Vurës and Vera'a, of northern Vanuatu. The project includes several collaborators, including linguists and biologists specialising in marine and terrestrial life who all contributed to a databank of images of plants and animals.

For various practical reasons, the project began before a data management plan was developed, with the result that each researcher used their own ways of identifying their material, employing a kind of folksonomy rather than using any standard systems for metadata which would have provided for far more

va- *gramm.part.* 1. causative prefix. e.g. *va-hani:* feed (lit. cause to eat). 2. **(East: ve-)**

gramm.part. plural prefix. e.g. *va-uranji* (East: *ve-uraji*) : children (lit. plural-child)

vahamauru *v.tr.* save someone (lit. cause to-live).

Figure 3: Example of data in the unpublished Tamambo dictionary by Dorothy Jauncey (in MS Word).

\lx va- \ps gramm.part.\sn 1 \de causative prefix. \xv va-hani \xe feed (lit. cause to eat).

\sn 2 \a (East: ve-)\de \ps gramm.part. plural prefix. \xv va-uranji \xe children

(lit. plural-child) \a (East ve-uraji)

\lx vahamauru \ps v.tr.\de save someone (lit. cause to-live).

Figure 4: A first pass of inserting codes indicating field boundaries in the Word document shown in Figure 3. Subsequent manual editing was required to complete the entry.

functionality — like a simple OLAC system for example. OLAC, the Open Language Archives Community, provides a metadata schema that is used by a number of language archives to provide interoperable descriptions of language records they hold. My role in the project was to bring various existing dictionaries into a format suitable for an installation in the national museum in Port Vila, with spoken headwords and images to add to definitions of as many of the headwords as possible. We decided that a simple method would be to create HTML versions that could run on a browser with a map interface. In order to get the dictionaries into the required format, they would first need to be structured, for example in a Toolbox lexical database. As one of the dictionaries had only been produced as a MS Word document, it needed to be reworked, but because the original formatting (see Figure 3) was fairly faithful to the implicit structure of the dictionary (e.g., using bold outdented formatting for headwords followed by italicized parts of speech) it was possible to convert it into a Toolbox format (as in Figure 4) with a series of regular expression operations.

The lexical database includes scientific names which we thought would allow images from the shared image databank to be linked automatically wherever the names matched. Unfortunately, the use of these names is not standardised, and the names of some plants and animals themselves have gone through various changes over time, so it is still necessary to do a great deal of manual work relating biological names to entries in the database. Nevertheless, the shared image bank is a great resource for dictionaries of Vanuatu and will, we hope, be used more generally by future dictionary projects. Some 5,300 images were available from all of the participants in the project, but only 1,500 had sufficient metadata to allow their contents to be

known without inspecting them. Of those 1,500, many had the Vurës word correlated with the photo identifier in a spreadsheet. It seemed like a solution to including images into the lexical database was to correlate the dictionary headword with the spreadsheet name. Using the relational function of a DBMS to load up the spreadsheet and the headwords we expected a significant correlation to save us the time of making the links by hand. In fact, of the more than 1,000 eligible entries, only fifty links were established. Furthermore, homographs (e.g., *lō* 'seaweed', and 'sun') in the dictionary prevented even those fifty from being accepted without further work. A major problem was the form in which data was stored in the spreadsheet, with queries marked with question marks inside the headword field, and differences in spelling between the two data sets preventing automatic correlation. Filenaming was another problem, as files were named in diverse ways even by the same individuals. Further, files were stored in nested directories with names that were not always unique, so they needed renaming before being put into a flatter file system for retrieval.

4.1 *Ngarrindjeri: from a flat database to a paper dictionary*

Ngarrindjeri is an Australian language from South Australia that was recorded in sixteen historical sources dating back some 170 years as well as being remembered by Ngarrindjeri people today. The small and finite set of written records, none of which gives particularly rich semantic information about lexical entries, was typed into a database and combined with the words known by speakers today. This resulted in 3,684 headwords tagged with their sources (oral, written or both), and each source is provided for each form in the dictionary, as can be seen in Figure 5.

nakun *Verb (trans).* seeing; looking at. *Written source:* K= nakun*;* T= nakkin*;* YA in S= nhakun*;* M= nakkin *;* Y= nakun 'is looking, seeing'. *Variant:* nhakun; nakin. *Etym:* From nak- 'see' + -un 'present tense suffix'. *Note:* This is a well known word. The present tense form of the verb 'seeing' can be pronounced and spelt as 'nakun, nhakun or nakin'. The future tense form is spelt 'nakan' and means 'see you later', see separate entry. *Oral source:* VB= nakun 'seeing' EM= nakun 'seeing' JY= nakun 'seeing' NG= nakun 'seeing' TR= nakun 'seeing' MS= nakun 'looking for' (eg. swan eggs)

Figure 5: A sample entry in the printed Ngarrindjeri dictionary (Gale et al. 2009).

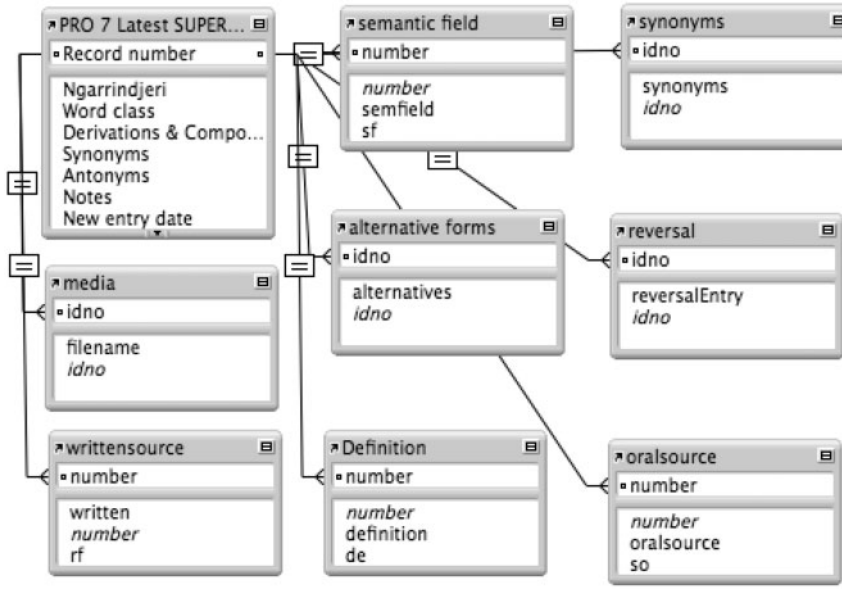

Figure 6: Relationship diagram from the Ngarrindjeri dictionary database showing eight tables related to the main dictionary table.

The Ngarrindjeri dictionary project used a FileMaker Pro database over some years to collate the various available sources. A problem for the project was the poor output that FileMaker Pro generates in its reports, typically providing little control on the output format, and unable to maximize the use of space on a page. I became involved in assisting the export of data from FileMaker to Toolbox. An immediate problem with the data in FileMaker was that it had used repeating fields (the same field can be repeated but does not distinguish its repeated content on export, nor does it formally capture the relationships between apparently related fields using this feature) and allowed more than one kind of information in one field so that it was not easy to export from. On the other hand, FileMaker does provide an easy interface and ways of sorting and presenting material onscreen that make it very useful, even more so for rendering a legacy dictionary like this which has no complexity of entries–no subentries, and no elaborated senses of definitions.

In order to allow the group of users to continue working with the tool that they were used to, I broke out repeated information into related tables, as shown in Figure 6, cleaned up inconsistencies in the data, and provided an XSL script to convert the output of the dictionary into a format that could be opened and read in Toolbox (Figure 7). The group had identified that they wanted to create dictionaries of the type created by the Multi-Dictionary Formatter from Toolbox. This was in preference to using the FileMaker output with its inability to take full advantage of page space and consequent extra expense in printing costs.

```
<!--<?xml version="1.0" encoding="ISO-8859-1"?>-->
<xsl:stylesheet exclude-result-prefixes="fmp" version="1.0"
xmlns:fmp="http://www.filemaker.com/fmpxmlresult"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<!-- <xsl:output encoding="ISO-8859-1" indent="yes" method="xml" version="1.0"/> -->
<xsl:template match="fmp:FMPXMLRESULT">
\_sh v3.0  400  MDF 4.0
\_DateStampHasFourDigitYear
<xsl:apply-templates/>
</xsl:template>
<xsl:template match="fmp:RESULTSET">
<xsl:apply-templates/>
</xsl:template>
<xsl:template match="fmp:ROW">
\lx <xsl:value-of select="fmp:COL[1]/fmp:DATA"/>
<xsl:value-of select="fmp:COL[2]/fmp:DATA"/>
\ps <xsl:value-of select="fmp:COL[3]/fmp:DATA"/>
<xsl:for-each select="fmp:COL[4]/fmp:DATA">
\de <xsl:value-of select="."/></xsl:for-each>
\nt <xsl:value-of select="fmp:COL[10]/fmp:DATA"/>
\an <xsl:value-of select="fmp:COL[9]/fmp:DATA"/>
<xsl:for-each select="fmp:COL[8]/fmp:DATA">
\sy <xsl:value-of select="."/></xsl:for-each>
<xsl:for-each select="fmp:COL[5]/fmp:DATA">
\so <xsl:value-of select="."/></xsl:for-each>
<xsl:for-each select="fmp:COL[6]/fmp:DATA">
\rf <xsl:value-of select="."/></xsl:for-each>
<xsl:for-each select="fmp:COL[7]/fmp:DATA">
\sd <xsl:value-of select="."/></xsl:for-each>
\et <xsl:value-of select="fmp:COL[12]/fmp:DATA"/>
<xsl:for-each select="fmp:COL[13]/fmp:DATA">
\re <xsl:value-of select="."/></xsl:for-each>
<xsl:for-each select="fmp:COL[14]/fmp:DATA">
\va <xsl:value-of select="."/></xsl:for-each>
</xsl:template>
</xsl:stylesheet>
```

Figure 7: Example of an XSL transformation applied to FileMaker Pro output to create a Toolbox file.

5. Conclusions

All software can be used in better or worse ways from the point of view of later recovery of the primary data. Examples of better ways of encoding data include a styled word processor document in which styles have been consistently applied, or a database in which data is recoverable as structurally discrete objects. By using structured data it is possible to create references to external media automatically, based on the form of the headword, rather than the references being manually constructed for each headword. If, as is suggested here, we need to engage linguists who may have low levels of technical skills but who want to produce and are quite capable of producing a fine dictionary, then we also need to provide assistance in training and support for their work.

References

Cablitz, G., J. Ringersma and M. Kemps-Snijders. 2007. 'Visualizing endangered indigenous languages of French Polynesia with LEXUS'. In *Proceedings of the 11th International Conference Information Visualization (IV07)* IEEE Computer Society. 409–414.

Gale, M. with S. Sparrow and the Ngarrindjeri community. 2009. *Ngarrindjeri dictionary*. Raukkan: Raukkan Community Council.

Himmelmann, Nikolaus P. 1998. 'Documentary and descriptive linguistics'. *Linguistics*, 36: 161–195.

Hsu, Robert. 1985. *Lexware Manual*. Manoa: University of Hawai'i Department of Linguistics.

Thieberger, N. 2004. 'Documentation in practice: Developing a linked media corpus of South Efate'. In Peter Austin (ed.), *Language documentation and description*, Volume 2. London: Hans Rausing Endangered Languages Project, SOAS, 169–178. (http://repository.unimelb.edu.au/10187/2199).

Thieberger, N. 2006. *A Grammar of South Efate: An Oceanic Language of Vanuatu*. Oceanic Linguistics Special Publication, No. 33. Honolulu: University of Hawai'i Press.