From Small to Big Data: paper manuscripts to RDF triples of Australian Indigenous Vocabularies

Nick Thieberger
School of Languages and Linguistics,
University of Melbourne, Parkville,
Vic 3010, Australia
thien@unimelb.edu.au

Conal Tuohy
322 Henson Rd, Salisbury, Queensland
4107, Australia
conal.tuohy@gmail.com

Abstract

This paper discusses a project to encode archival vocabularies of Australian indigenous languages recorded in the early twentieth century and representing at least 40 different languages. We explore the text with novel techniques, based on encoding them in XML with a standard TEI schema. This project allows geographic navigation of the diverse vocabularies. Ontologies for people and placenames will provide further points of entry to the data, and will allow linking to external authority. The structured data has also been converted to RDF to build a linked data set. It will be used to calculate a Levenshtein distance between wordlists.

1 Introduction

Of the several hundred languages spoken in Australia over millennia before European settlement, less than fifty are currently learned by new generations of Aboriginal children. Records of the languages that are no longer spoken everyday are thus extremely valuable, both for those wanting to relearn their own heritage language, and for the broader society who want to know about indigenous knowledge systems. In this paper we discuss current work to encode vocabularies collected by Daisy Bates for a number of indigenous Australian languages, mainly from Western Australia in the early 1900s. These papers have been in the public domain since 1936 and as a collection of manuscripts held in Australian state libraries. We outline the process of creation and naming of the digital images of this paper collection, then show how we have encoded parts of this material, and created novel views based on the encoded formats, including page images with facsimile text. As the project develops we expect to build a model that can be applied to further sections of the collection that are not as well structured as the vocabularies. This work is offered as one way of encoding manuscript collections to provide access to what were otherwise paper artefacts[1].

2 The task

The complex problem of using historical records of Australian languages has benefited from the cooperation of a linguist (NT) with a technology expert (CT). The dataset has been constructed according to the TEI Guidelines[2], to embody both a (partial) facsimile of the original set of typescripts and a structured dataset to be used as a research collection. This material will be open to reuse, in particular providing access for indigenous people in remote areas to vocabularies of their ancestral languages. The model will also be an exemplar of how a text and document-based project, typical of humanities research, can benefit from new methods of encoding for subsequent reuse. For more on the content of the collection see [3].

By processing the wordlists and making them accessible online, we have prepared material that will be of use to indigenous Australians today, as well as creating an open research dataset which may be linked to and from other data. We believe that new digital methods can enrich the metadata and the interpretation of primary records in this collection. There are some 23,000 images on microfilm, and the first task has been to rename all files. Analysis of the texts identified three types of document, the 167 original questionnaires (around 100 pages each), 142 typescript versions of those questionnaires (each made up of varying numbers of pages), and 84 handwritten manu-

[1] The current alpha version is at http://bates.org.au/
[2] http://www.tei-c.org/Guidelines/P5/ (Accessed 2016-09-17).

Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 19–23,
Honolulu, Hawai'i, USA, March 6–7, 2017.

scripts that could be either questionnaires or additional material.

Any given word in a typescript comes from a predictable location in the associated questionnaire, and so can be assigned an identifier to allow targeted searching. Thus links can be automatically established to display a typescript image page and a questionnaire image page for any target word.

The JPEG files of typescripts were sent to an agency for keyboarding. The XML was subsequently enriched as seen in the snippet in Fig.1.

```
<listPerson>
 <person>
  <persName role="speaker">
   Kulaiji
  </persName>
 </person>
 <person>
  <persName role="speaker">
   Ijala
  </persName>
 </person>
</listPerson>
<listOrg>
 <org>
  <orgName type="tribe">
   Barduwonga
  </orgName>
 </org>
 <org>
  <orgName type="tribe">
   Burdurad,a
  </orgName>
 </org>
</listOrg>
```

Figure 1. Template XML used for keyboarding the vocabularies

This template further references pages in the typescript, thus allowing any term to resolve to the image of its source page. All files are stored in a bitbucket[3] repository allowing us to work on them collaboratively at a distance and to track file versions.

At the end of the first stage of work we are able to visualise the wordlists in various ways, including a geographic map (Fig. 3), a list of all words and their frequencies, and a list of wordlists and the number of items they contain, in addition to being able to search the whole work for first time. Sorting and arranging the words helps in the correction of errors that inevitably occur in the process of dealing with large numbers of vocabularies and exporting the lists in an RDF format allows us to generate the statistics about frequency of terms, and to identify the coverage of particular lists.

3 Design decisions for encoding the dataset

The scale of the Bates dataset requires outsourced transcription, but it is difficult to outsource the full (lexicographic) semantics, that is, capturing the meaning of added entries and examples. This is even more the case as the documents include a great deal of variation, both in their spellings and in their contents, so it is not necessarily easy to interpret them semantically. We focused the outsourced transcription task on a superficial (typographic) encoding of the documents. The encoding captured the tabular layout (i.e. the text is divided into rows and cells), local revisions (i.e. rows added to the table), and pagination. The right-hand column of these tables, generally containing a comma-separated list of indigenous words, was then marked up by an automated process (an XSLT transformation). To explicitly encode the lexicographic data in the forms, we needed to tag each of the words, classify it as either English or indigenous, and hyperlink each indigenous word or phrase to the English words or phrases to which it corresponds.

Given the size of the encoding task, it was essential to minimise the amount of manual work, and reduce the scope for human error, by automating the markup process as much as possible.

The typing did not include identification of the relationships between the words in the lexicon, recognising that it is preferable to use transcribers to capture source texts with a high level of accuracy, but conceptually at a superficial level, and then to add those semantics later, automatically, or using domain experts. We provided our keyboarders with document layout (i.e. pages and tables), rather than linguistic categories (terms and translations).

As an example of the automatic addition of semantic information, we decided to recover the lexicographic semantics implicit in the text by programmatic means, inserting explicit metadata (markup) in the text to record these inferred semantics. This had the additional advantage that the automated interpretation could be revised and re-run multiple times, and the output checked each time. We see the visualisation of the results

[3] http://bitbucket.org/

that is permitted by this work as contributing to the repeated process of correction of the data.
The tabular layout itself implies a relationship between a prompt term (in the left hand column of the questionnaire), and one or more terms in the right hand column. The right hand column contains one or more terms in an indigenous language, but in addition it may contain other English words, typically in brackets, or separated from the indigenous words by an "=" sign. (e.g. Sister joo'da, nar'anba = elder (57-033T[4])).

```
<row>
  <cell>Snake</cell>
  <cell>
   Burling, jundi (carpet),
   binma, yalun
  </cell>
</row>
```

Figure 2. A sample row of content

The left-hand column of the questionnaire form was pre-printed by Bates, for example, in Fig. 2 the printed word was "Snake". The right hand column was to be filled in with the local language term. In this case the recorder wrote *Burling, jundi (carpet), binma, yalun.* Our aim is to identify which of the words are intended to represent indigenous words, and which (like "carpet") are actually additional English words which specify a refinement of the original term. In this case, the word *jundi* is specifically for "carpet snake", whereas the other words may refer to snakes more generically.

The next step is to pass these XML documents through a series of small transformation programs, each of which makes some interpretation of the text and enhances the XML markup in one way or another. The cumulative effect of the transformations is to produce a final output document in which the English and indigenous words and their lexicographical relationships are marked up explicitly using hyperlinks.
For example, a few steps along in the transformation pipeline the same row will have been enhanced with punctuation characters parsed into <pc> markup:

```
<cell>Snake</cell>
<cell>Burling<pc>, </pc>
jundi <pc>(</pc>carpet
<pc>)</pc><pc>,</pc> binma
<pc>,</pc> yalun</cell>
```

Once the punctuation characters "(", ")", "=", and ";" are picked out, a subsequent transformation classifies the residual words into different types, based on the surrounding punctuation. The TEI element <seg> (segment) is used to assign a type, which is either item or parenthetical:

```
<seg type="item">
  Burling</seg><pc>,</pc>
<seg type="item">
  jundi</seg> <pc>(</pc>
<seg type="parenthetical">
 carpet</seg><pc>)</pc>
<pc>,</pc>
<seg type="item">binma</seg>
<pc>,</pc>
<seg type="item">yalun</seg>
```

These "lexical" and "grammatical" transformations set the stage for final transformations to make a guess as to what the text actually *means*; which of the words are English and which are indigenous, and how they interrelate:

```
<cell><gloss xml:id="snake"
xml:lang="en">Snake</gloss>
</cell>
<cell>
 <term ref="#snake"
 xml:lang="nys">
  Burling</term>,
 <term ref="#snake-carpet
 #snake" xml:lang="nys">
  jundi</term>
 (<gloss type="narrow"
 xml:id= "snake-carpet"
 xml:lang="en">carpet
 </gloss>),
 <term xml:lang="nys"
 ref="#snake" >binma</term>,
 <term ref="#snake"
 xml:lang="nys">yalun</term>
</cell>
```

Note how the term *jundi* is linked to both the "Snake" and the "(carpet)" glosses, whereas the other terms are linked only to "Snake". Note also that the words "Snake" and "carpet" are both now explicitly identified as English and the language words are identified as being in a particular Australian language.

The intermediate TEI documents (containing automatically-inferred term/gloss markup) will contain errors in many places, due to inconsistency and ambiguity in the source documents.

[4] http://bates.org.au/images/57/57-033T.jpg

Those errors became most apparent in the word lists and maps generated in the first phase outputs of the project, as shown in Fig. 3.

4 Markup: automation and "markup by exception"

The transformation of the base XML files is via an XSLT script that parses the lists into distinct words, and inserts the appropriate hyperlinks to relate each indigenous word to the English word(s) to which it corresponds. Some indigenous words have multiple corresponding English words, separated by commas or sometimes semicolons:

```
Ankle    Kan-ka, jinna
werree, balgu
```

Occasionally, the word "or" is used before the last item in a list:

```
Blood Ngooba or yalgoo
```

Sometimes the right hand column contains additional English language glosses, generally to indicate a narrower or otherwise related term. Most commonly, these additional English glosses were written in parentheses, following the corresponding indigenous word:

```
Kangaroo    Maloo (plains),
margaji (hill)
```

Sometimes an additional English gloss is written before the corresponding indigenous term, and separated with an equals sign (or occasionally a hyphen):

```
Woman, old    Wīdhu; old man
= winja
```

An XSLT script is easily able to handle all these cases, and rapidly produce markup which is semantically correct. However, as the forms were filled out by many different people, inevitably there are some inconsistencies in the text which can lead the XSLT script into error. Sometimes, for instance, the indigenous words are in brackets, rather than the English words. Sometimes the text is written in a style which is just not amenable to parsing with a simple script:

```
Bardari - like a bandicoot,
only with long ears and
nose.
```

```
Bira - also like a bandi-
coot, but short and thick
body, little yellow on back.
```

In these exceptional cases the easiest thing to do is apply some human intelligence and mark up the text by hand.

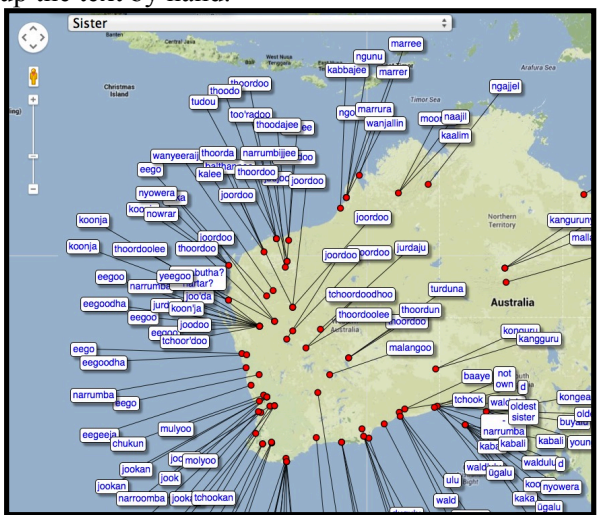

Figure 3 shows the range of equivalents for the word 'sister' mapped geographically

This naturally leads to an iterative data cleaning workflow in which a time-consuming batch process crunches through the documents, gradually enhancing them, performing automated validation checking, and finally generating visualisations for humans to review. We found the data visualisations to be a potent force for quality assurance. It is often very easy to spot interpretive errors made by the automated parser, and that correction can feed back, either as a refinement of the automated process, or as a manual correction of the document markup, leading to a gradual improvement in the data quality.

5 Conversion to Linked Data

The TEI XML contains a great deal of detail about the questionnaires as texts, such as how the word lists were formatted, punctuated, and paginated, and although this is essential in order to be able to read the questionnaires as texts, for proofing, it is also helpful to be able to abstract away from that contingent information and deal only with the vocabularies as linguistic data. For this purpose, once the TEI XML has been automatically enhanced to include the explicit lexicographic semantics, a final XSLT extracts information from each of the TEI documents and re-expresses it as an RDF graph encoded in

RDF/XML, using the SKOS[5] vocabulary for the lexicographical information, and the Basic Geo (WGS84 lat/long)[6] vocabulary for the geospatial location of each vocabulary. The distinct RDF graphs are then merged to form a union graph, by saving them into a SPARQL Graph Store.

Each vocabulary is represented as a SKOS:ConceptScheme, which in turn contains a SKOS:Concept for each distinct concept; either a concept identified by Bates in her original questionnaire, or a concept added during the original interviews. In addition, a special SKOS:ConceptScheme (called "bates") represents the original blank questionnaire, and functions as a hub in the network of concepts. Each concept in the "bates" vocabulary is explicitly linked (as a SKOS:exactMatch) to the corresponding concept in every one of the indigenous vocabularies.

The concepts in the "bates" vocabulary have labels in English, whereas the corresponding concepts in the other vocabularies are labelled with indigenous words. Many of the concepts in the indigenous vocabularies have multiple labels attached, representing the synonyms recorded in the questionnaires.

Once the RDF graphs are loaded into the SPARQL Store, the union graph can be easily queried using SPARQL. We use SPARQL queries to produce a map of each word, histograms of the frequency of vocabularies containing a given concept, and of the varying conceptual coverage of the different vocabularies. We can also extract the indigenous words in a form convenient for further processing, including computing Levenshtein Distance between vocabularies, to support automated clustering of the vocabularies.

6 Next steps

Once all the typescripts have been keyboarded we will be in a position to edit the whole collection for consistency. As noted, each wordlist was compiled by different people, and was then typed under Bates's supervision, so having access to both the manuscript and typescript will enable an edition that captures the content more accurately than is currently the case. In phase two, we will implement a framework that allows these images and text to be presented together, and then extend the model into other parts of the Bates collection that also includes words and sentences in Aboriginal languages with the potential that we can attract volunteers (crowdsourcing) to work on transcribing and correcting the content. We will also be in a position to generate similarity measures between vocabularies and from them a multidimensional scaling view of the distance between the vocabularies as in [2], and more recently [1].

7 Conclusion

With the work undertaken so far it is clear that the process of encoding has led to a deeper understanding of the target material. It has provided novel visualisations and helped us to appreciate the context of the original material. While the more usual approach to archival lexical material has been to extract lexical items into a relational database or spreadsheet, the data could not be coerced into such a form now without a significant amount of interpretation and loss of contextual information.

It would be a mistake to focus immediately on the lexicographic data embedded in the forms, and neglect other aspects of the forms. We have no access to the original language speakers; for us the questionnaires are themselves the data, and we should therefore record the questionnaires, not just the data "contained in" the questionnaires. Further, by maintaining the links back to the primary records we allow users to situate the encoded material in its source. Premature data reduction to cells in a database risks ignoring useful information. The data modelling task aims to capture the data along with all possible context. The use of TEI, rather than, say, a relational database enables that conceptually open-ended, exploratory and iterative data modelling.

References

[1] Embleton, Sheila, Dorin Uritescu and Eric S. Wheeler. 2013. Defining dialect regions with interpretations: Advancing the multidimensional scaling approach *Lit Linguist Computing* 28: 1 13-22.

[2] Nash, David. 2002. Historical linguistic geography of south-east Western Australia, pp. 205-30 in *Language in Native Title*, ed. by John Henderson & David Nash. Canberra: AIATSIS Native Title Research Unit, Aboriginal Studies Press.

[3] Thieberger, Nick. fc. Daisy Bates in the digital world. In Peter Austin, Harold Koch, & Jane Simpson (eds) *Language, Land and Story in Australia*. London: EL Publishing.

5 Alistair Miles, Sean Bechhofer, eds. 2009. SKOS Simple Knowledge Organization System Reference. https://www.w3.org/TR/skos-reference/

6 Dan Brickley, ed. 2004. Basic Geo (WGS84 lat/long) Vocabulary. https://www.w3.org/2003/01/geo/