### THIEBERGER — Encoding

<0 figures>

### Encoding Vocabularies of Australian Indigenous Languages

Thieberger, N. and Tuohy, C.

This paper discusses a project that is encoding archival vocabularies of Australian Indigenous languages recorded by Daisy Bates (White, 1985; McGregor, 2012) in the early 20th century. With this material we aim to create new research outcomes by exploring the text using novel techniques. We advocate a methodology in which primary research materials should be citable and accessible so that derived analyses can be verified and the primary data enriched by the process of research. This work will allow identification of the languages represented by these vocabularies and the internal variation in those languages. It will also make the data searchable for the first time, and linkable to other datasets, with the long-term goal of creating rich records for Australian Aboriginal languages, and, at the same time, embedding research practice in established international methodologies. This is the first application of the Text Encoding Initiative (TEI) to Indigenous-language material in Australia, and one of the few uses of the TEI with Indigenous languages that we know of (see Czaykowska-Higgins et al., 2014).

Some 23,000 pages of these manuscripts at the National Library of Australia were digitised as TIFF images from a microfilmed copy and named according to NLA conventions, and 4,300 pages have so far been keyboarded, geocoded, and TEI encoded. These pages represent three kinds of material: (1) a handwritten questionnaire, containing some 2,000 prompt words and sentences; (2) a typescript version of each questionnaire; and (3) other manuscript pages. We have identified over 150 speakers and on the order of 40 languages in the collection. One of the outcomes of the project will be to determine the relationships between these wordlists and later sources, and to identify more securely what languages are represented.

The Barr Smith Library has been producing pdf/text versions of some of these manuscripts that increases their availability,1 and the collection has been in the public domain since the 1950s. It has been used in various projects, including Native Title claims, but, in general, the knowledge gained by these processes is not reflected in the collection itself. The promise offered by new DH methods is that any enrichment of the sources should then be reflected in them. We acknowledge that Indigenous people have a strong interest in the Bates vocabularies, but, given the number of speakers involved, the many descendants they would now have, the geographic spread of the sources (covering most of Western Australia), and the resources available to us, we are unable to consult widely in the way that, for example, the Laves project did, working as it was with archival material from a single geographical area (as reported in Henderson, 2008).

The TEI texts will remain the repository of the full data model of the project, representing the corpus of texts in a number of different dimensions. At one level, the TEI texts will constitute a set of facsimile editions by linking page facsimile images to the textual transcriptions of those images. At another level they will include explicit markup of person, language, and place names, to provide further points of entry to the data, and allow linking to external authority files (where possible). This multi-dimensional data model will be built up step by step and encoded in the TEI corpus by a combination of careful manual markup and the application of algorithmic methods.

The manual data-entry work was initially restricted to a typographical encoding of the texts simply as two-column tables with headings. The right-hand column of these tables, generally containing a comma-separated list of Indigenous words, was then marked up by an automated XSLT transformation. This naturally leads to an iterative data cleaning workflow based on a software development ‘build’ process, in which a time-consuming batch process crunches through the documents, gradually enhancing them, performing automated validation checking, and generating visualisations for manual review. We found these data visualisations a potent force for quality assurance. It is 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.

The dataset exists in a number of forms: a paper archive; a set of photographic images in TIFF format, and derived JPEGs; a set of TEI transcriptions that encodes the text and the tabular layout of the forms; and links to the JPEG page images. A gazetteer of places is maintained as a spreadsheet. The workflow is managed via a spreadsheet, and the source files are stored in a bitbucket repository.

One important requirement is the ability to track the provenance of the individual words. A dataset of this scale and type has scope for numerous transcription errors, and it must be possible for a reviewer to query any word, to see the manuscript from which it came, and to view images of the source materials to establish the context in which the term appeared, and thus to check that the contextual metadata in the TEI is correct.

The solution is to begin by building a model of the texts, rather than attempting to directly model the lexicographic knowledge they contain. The model of the texts can then be enriched with the lexicographic knowledge in the form of embedded metadata, and from this a simplified model can then be derived to serve the purpose of linguistic computation. We build the data model by parsing the text, but rather than parsing the text and populating a data model, we embed the parsed version into the text in the form of XML markup. This makes the interpretive act explicit—something that can be questioned and altered. We classified each word as either English or Indigenous, and hyperlinked each to the English 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.

This XSLT script parses the lists into distinct words and inserts the appropriate hyperlinks to relate each Indigenous word to the English word or words 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 at the end of 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, usually written in parentheses, following the corresponding Indigenous word, e.g.,

|  |  |
| --- | --- |
| Kangaroo | Maloo (plains), margaji (hill) |

Sometimes an additional English gloss is written before the corresponding Indigenous term and separated with an equal sign (or hyphen):

|  |  |
| --- | --- |
| Woman, old | Wîdhu; old man = winja |

An XSLT script is easily able to handle all these cases and rapidly produce markup that 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, or the text is written in a more explanatory style, which is not amenable to parsing with a simple script:

|  |  |
| --- | --- |
| Bandicoot | Jiriwardu, bardari and bira - two species like a bandicoot.  Bardari - like a bandicoot, only with long ears and nose. |

In these exceptional cases the easiest thing to do is apply some human intelligence and mark up the text by hand. To support this ‘markup by exception’ style of work, the automated markup XSLT is programmed to simply pass over any text that already contains the TEI <term> and <gloss> markup.

If we find something that looks odd through an analysis of the lexicographic data, how are we able to check whether it is a genuine language oddity or the result of an error? It is essential to be able to track the linguistic data back to the archival source, and that means maintaining links from the derived data that point back to the source questionnaires, and that in turn means having a data model that is able to accommodate all aspects of the data.

The value of the TEI is that it can link those aspects together in a single set of source files. Then the ‘normalised’ version can be extracted as needed from the fully modelled data.

A number of issues will be explored as we develop this project, including the possibility of automated regularising of orthographic differences between word-lists, further permitting comparison of similarities via a Levenshtein distance measure. As the underlying material is in a reusable format, it will be possible to publish electronic editions of the manuscripts and output sets of word-lists for use in schools, and to do comparative analysis of the vocabularies to infer relationships and locations (as in Nash, 2002, and more recently Embleton et al., 2013). We have used XSLT to extract a graph database in RDF-XML format from the TEI source data, and we use SPARQL to query this derived dataset to produce visualisations, and to expose it as Linked Data.

**Funding**

This work was funded by a research grant from the Faculty of Arts at the University of Melbourne and supported in part by ARC grants DP0984419 and FT140100214.

**Note**

1. https://digital.library.adelaide.edu.au/dspace/handle/2440/69252.

**References**

**Czaykowska-Higgins, E., Holmes, M. D. and Kell, S. M.** (2014). Using TEI for an Endangered Language Lexical Resource: The Nxaʔamxcín Database-Dictionary Project. *Language Documentation & Conservation,* **8**: 1–37, http://hdl.handle.net/10125/4604.

**Embleton, S., Uritescu, D. and Wheeler, E. S.** (2013). Defining Dialect Regions with Interpretations: Advancing the Multidimensional Scaling Approach. *Literary and Linguistic Computing,* **28**(1): 13–22.

**Henderson, J.** (2008). Capturing Chaos: Rendering Handwritten Language Documents. *Language Documentation & Conservation,* **2**(2): 212–43, http://hdl.handle.net/10125/4347.

**McGregor, W.** (2012). Daisy Bates’ Documentations of Kimberley Languages. *Language and History,* **55**(2): 79–101.

**Nash, D.** (2002). Historical Linguistic Geography of South-East Western Australia. In Henderson, J. and Nash, D. (eds), *Language in Native Title*. Canberra: AIATSIS Native Title Research Unit, Aboriginal Studies Press, pp. 205–30.

**White, I.** (ed.). (1985). *The Native Tribes of Western Australia*. National Library of Australia, Canberra.