The language documentation quartet

Simon Musgrave
Monash University
simon.musgrave@monash.edu

Nicholas Thieberger
University of Melbourne
thien@unimelb.edu.au

Abstract

As we noted in an earlier paper (Musgrave & Thieberger 2012), the written description of a language is an essentially hypertextual exercise, linking various kinds of material in a dense network. An aim based on that insight is to provide a model that can be implemented in tools for language documentation, allowing instantiation of the links always followed in writing a grammar or a dictionary, tracking backwards and forwards to the texts and media as the source of authority for claims made in an analysis. Our earlier paper described our initial efforts to encode Heath's (1984) grammar, texts (1980), and dictionary (1982) of Nunggubuyu, an Australian language from eastern Arnhemland. We chose this body of work because it was written with many internal links between the three volumes. The links are all encoded with textual indexes which looked to be ready to be instantiated as automated hyperlinks once the technology was available. In this paper, we discuss our progress in identifying how the four component parts of a description (grammar, text, dictionary, media, henceforth the quartet) can be interlinked, what are the logical points at which to join them, and whether there are practical limits to how far this linking should be carried. We suggest that the problems which are exposed in this process can inform the development of an abstract or theoretical data structure for each of the components and these in turn can provide models for language documentation work which can feed into hypertext presentations of the type we are developing.

1 Introduction

In this paper we describe work we have done to encode the quartet of language documentation, the grammar, texts, dictionary, and media recordings. We explore a method using text encoding and automated markup of existing textual documents, those by Heath, who published his description of Nunggubuyu (also known as Wubuy, ISO639=nuy) in three volumes (1980, 1982, 1984). As a matter of deliberate choice, he did not include example sentences in the grammar or in the dictionary, providing instead references to the text collection. In Heath's words:

> These textual citations serve several purposes. When attached to a fully cited Nunggubuyu ex[ample]., they have basically a documentary value –the reader is assured that the ex[ample]. is from a real text, and a reader wanting to know more or having doubts about the analysis can find it and analyse it. […] In this way, we take maximal advantage of the published texts (especially NMET*) achieving a far higher level of documentation than is observable in other reference grammars. (Heath 1984: 4) (*NMET = Heath 1980)

This practice was based on the important principle that all examples should come from spontaneous text and should be viewed in context – but it makes using the description challenging. Heath's work is visionary but constrained by the available technology. It is interesting to note that Heath's practice has changed over the years; in more recent work (e.g. Heath 2017) he includes example sentences in his grammatical description.

Although hypertext was first used as a term in 1965 (Nelson 1965), it was not an easily accessible technology in the 1980s in the way it has become more recently. The technology to make Heath's vision usable now exists in hypertext (and the world wide web as a mode of delivering hypertext) and our project aims to demonstrate the possibility of presenting grammatical materials in this way. As Nelson suggests, hypertext can be a 'linkage structure between documents' that might 'hold the thoughts together between documents' (1982). Building a working version of the links in these documents is a contribution towards demonstrating the value of the project of creating such richly interlinked grammars in future – something that more and more linguists will be wanting to do. We have succeeded in encoding the three volumes and in making an online version of these interlinked texts. A sample set of texts also have links to audio files (see https://rebrand.ly/text163).

2 Why is this worth doing?

The utility of hypertext can be and probably has been overstated (Bevilacqua 1989). However, for the case described here, links between an analytical apparatus (the grammar), the source text – ideally with its recording – and a dictionary, dense linking is a fruitful and scientifically important technique. Why scientifically important? As Heath designed his three volumes, he was providing the evidence for his analysis with citation links. In the decades since his work, the theory of language documentation (Himmelmann 1998) has formalised the necessity of providing well-formed records of language as part of linguistic fieldwork. Included in these desiderata is the ability to cite parts of a corpus of transcripts as the warrant for a claim made in analysis. Verifiability of analytical claims is increasingly important, and, while replicability is typically not achievable in non-science/technology research, it is nevertheless possible to set an analytical apparatus in a rich context allowing others to follow one's reasoning.

In the late 1980s, J. Randolph Valentine (Valentine 1992) produced a set of HyperCard stacks called Rook that presented audio and a transcript of a text in Ojibwe, together with morphemic glosses that, when clicked, linked to the relevant point in a dictionary. Part of speech information linked to a grammar sketch using a lookup table of correspondences. By not using handmade links it was thus extensible, allowing new content to be added and automatically linked. So, while the instance of Rook no longer works, it nevertheless offers a useful model for us in our work on the Heath materials. The ease with which such links can be created and used has improved over the past 40 years since Heath's work was published. Nevertheless, it is still the case that few grammatical analyses instantiate links in this way. In the 1990s, grammars by Morey (published as Morey 2005) and Thieberger (published as Thieberger 2006) included links to media for example sentences (also Thieberger 2001). In each case a great deal of manual work was required as there were no tools available for citation of primary media. The most popular method now is to use the free software Elan that became available in 2002, and to create a media/text corpus of Elan files. However, even then, the citation is to a timed section of media and associated utterance unit, it is only, to maintain the musical analogy, a duet between the text and the media.

In recent work, Matter (2020) presents a system for generating linked descriptive materials of the same kind which we are developing. His system uses Cross Linguistic Data Formats[1] to store input data and the presentation uses Markdown as the basic technology. To date, only a small sample of the output is available for inspection.

3 Current tools and where they leave us

When working from legacy materials, there are two stages of processing to be accomplished: from the existing material to some kind of structured format that will be the preserved version of the text, and then from that format to the presentation format. The assumption is that what is the structured format in this process, intermediate in our workflow, will be the starting point for those working in this way in the future. A different kind of structure is appropriate for each part of the grammatical description, and we will briefly introduce these, aside from the media component (for which standards are well-known and ubiquitous).

3.1 Dictionary

For dictionaries, the Lexical Markup Framework (LMF) has been an ISO standard (24613) since 2008. However, the main uses of LMF are in natural language processing and in creating machine readable dictionaries and its relation to the work practices of descriptive linguists is not intuitively clear. We have chosen to work instead with a version of the Lexical Interchange FormaT (LIFT) developed by the Summer Institute of Linguistics (Hosken 2006). Although this format

[1] https://cldf.clld.org/

has not been used widely, and has not been updated since 2011, it has two advantages for our purposes: it has some flexibility and it has clear relations to the lexicons created by other SIL tools (such as Toolbox and FLEx) which remain popular with field linguists.

3.2 Texts

The typographical presentation of interlinear glossed text is easy to understand but it represents a complex data structure (Bow, Hughes and Bird 2003). We adopt Xigt which is the model proposed by Xia et al (2016) and note the work done in the ODIN project (Lewis et al 2009) to identify IGT on the web. The attraction of the Xigt model is that elements on each tier have an index, and alignment between tiers is accomplished by referencing these indexes. For example, alignment of a free translation line would reference a text unit of some kind, and alignment of a group of morphemes would reference an element in the word tier. We have made some minor additions to the model as set out in Xia et al (2016), for example to allow for punctuation within text units and to include references to media sources. There are no references to media in the text collection (or indeed in any other part of Heath's description); this is not surprising given the technological possibilities in the 1980s. We have to manually retrieve time codes from the available media files and, compared to current best practice, this represents a difference in workflow rather than a difference in method (although forced alignment is showing promise as an automated means of aligning media and transcripts).

3.3 Grammar

Discussions of what constitutes a descriptive grammar (e.g. the contributions in Ameka, Dench & Evans 2006) concentrate on the content of the grammar but we are not aware of similar discussion of the formal properties of the genre. Most linguists would have intuitions about the typical layout of a grammar, and at least one project tried to regularise the process (Comrie & Smith 1977), but this remains a relatively unexplored area. Given this situation, we take the most widely accepted approach to encoding text in humanities research, the Text Encoding Initiative Guidelines, as a starting point. Basic units such as chapters, sub-sections and paragraphs are not at issue, but the process of adding explicit markup to a grammar raises some questions. For example, Heath (1984) uses tables to present permitted intervocalic consonant clusters (Table 2-3, p21-22), paradigms of inflected forms (e.g. Table 7-1, p272) and what he calls root forms with associated verbs and their glosses (Table 12-1, p424). It is not clear to us that the similarity of presentation in these cases reflects similarity in underlying data structures and therefore whether encoding the tables as tables is the best approach for each instance. Questions such as these will certainly not be resolved in encoding a single instance of the genre.

A more immediately relevant question is to what extent can we automatically identify potential links to other parts of the quartet in the text of the grammar. This is straightforward for the references to text examples which Heath gave in a standard format (see further discussion below) and this is also the case for internal cross-references to other parts of the grammar as Heath also formatted these consistently. Identifying potential links to dictionary entries depends on being able to identify Nunggubuyu forms in the text, and, fortunately, in the grammar, Heath consistently places slashes around most such forms, for example: "in which case the postposition /-magi/ must be used to mark the form as Evitative" (1984: 340). For other forms we have to manually mark the Nunggubuyu forms, but there would be potential to use NLP methods to identify non-English strings.

4 Language description as a dense network

As text encoding of linguistic grammars has not been a feature of our field, we want to explore what it offers as a means to identify entities within the quartet that can be linked to or from, such as example sentences (in the grammar or the dictionary) pointing back to textual corpora, themselves linked via time-alignment to primary media, or terms (lemma/word/morpheme) in the corpus linked to a dictionary and items in the corpus linked to relevant points in the grammar.

Our strategy is to identify items in the textual materials which can reliably be recovered by search procedures and which can then be coded with information which will allow links to be generated. Examples of the kind of targets for internal references (important in navigating a document of 666 pages) are references to chapters, sections, and tables:

"Chapter 1" can be rendered as <ref type="int-c" n="1"/>Chapter 1.</ref>

"§1.2" can be rendered as <ref type="int-s" n="1.2"/>1.2 </ref>

"TABLE 2-1" can be rendered as <ref type="int-t" n="2-1"/>TABLE 2-1</ref>

Aside from internal cross-references, Heath explicitly encoded links from dictionary entries and from grammatical description to examples in Nunggubuyu Myths and Ethnographic Texts (Heath 1980). These links have a standard form which can be searched easily, line 3 of unit 5 in text 63 is encoded as 63.5.3. There are 5,765 such links from the dictionary and 5,404 from the grammar (these figures are slightly low as there are a handful of references of the type 63.4-7). Internal cross-references between words in the dictionary are being generated automatically (there are currently 8,282 such links) as are reverse links from texts to dictionary entries. The encoding of texts is a work in progress, and, to date, we have 280 text units encoded with 3,520 links to the dictionary. This allows us to estimate the number of such links per text unit and then to estimate that the complete text collection will have around 15,000 such links. However, both of these last two numbers are probably too low as the correspondence between citation forms in the dictionary and forms used in the texts is not entirely consistent. Developing 'fuzzy' matching in generating such links is an area of future work.

We have not yet begun to instantiate links between forms mentioned in the grammar and corresponding dictionary entries (there are approximately 10, 000 of these) and we have not mentioned links to media in this discussion (although these will not be so numerous), but even on the basis of our work to date, it is clear that the complete hypertext description will include around 45,000 links between the different components. This represents dense interlinking, but it is not without precedent in computational approaches to humanities materials. Willard McCarty's An Analytical Onomasticon to the Metamorphoses of Ovid contains approximately 60000 links in 1200 lines of Latin poetry (McCarty 2017).

Heath did not provide explicit links from texts back to the grammar. We believe them to be an essential part of our project, but implementing these links raises several problems which we have yet to solve. First, there is the question of the level of granularity of a linked target. A sub-section in the grammar is a possible target, but many sub-sections are several pages long. Page references might therefore be a better solution, but this would mean relying on an artefact of print presentation and this would not be consistent with our approach which sees hypertext presentation and print presentation as two possible outputs from a dataset. We are currently working with sub-sections as targets, and rely on the reader to locate relevant information within that sub-section.

Second, there is the question of what is the correct anchor for such links. Heath's references into the texts resolve at the level of lines within text units, but the lines are another artefact of print presentation as is evident in places where Heath breaks words across lines. This suggests that the heading of the text unit could be the anchor, but this in turn raises a third question: linking to a text unit may be many-to-one (several grammatical points may be exemplified in one text unit) and how should this be handled for reverse links? A possible solution would be to have a drop-down menu listing the available reverse links when the user's mouse hovers on the text unit heading. We have aimed to rely only on the possibilities of standard HTML5, the solution just mentioned would need to be implemented with JavaScript (or similar technology).

We are not automatically generating links from a dictionary entry to text examples and, although the possibility is tempting, we suggest that doing so could be problematic. For example, it would be straightforward to automatically generate a link from a dictionary entry to every instance of a morpheme in texts, and to the extent that a morpheme was associated with a grammatical phenomenon (e.g. a relative clause marker), from grammatical description to every text instance. But is it useful to have links to every example of a common morpheme representing a pervasive grammatical phenomenon? While we are sure that it is desirable for the interested user to be able to access that information somewhere, we are much less sure that a grammatical description is the right place. Writing a grammatical description has a curatorial aspect to it, and presenting the most relevant examples, rather than all possible examples, is important. An alternative could be to provide a service to generate KWIC views of texts based on a dictionary entry, sending the query to be dealt with in corpus software, for example.

Links to media are provided from text units. We located Heath's recordings (they are archived at the Australian Institute of Aboriginal and Torres Strait Islander Studies) and aligned some with the corresponding text, something that was not possible at the time Heath wrote the work. There are, of course, corrections in the text so that it is not an exact transcript, but the links work as can be seen in an online example (https://rebrand.ly/text163).

Another advantage of the coded document is that we can mark Nunggubuyu terms and provide two versions, one using Heath's orthography, and one using the current community orthography. We do this by using the <choice> tag as follows (with no implication that one form is a correction of the other, this is simply a way of making TEI work for us in this context).

```
<choice>
   <orig>/waraywaray/</orig>
   <corr>/warraywarray/</corr>
</choice>
```

Nunggubuyu forms are easily distinguished from English in texts and in the dictionary, and as we noted above, the correspondence between the original and the current orthographies is consistent. Therefore it is a simple task to encode every Nunnggubuyu form in the textual material with information that can allow the presentation to be in the orthography chosen by the user.

5 Lessons from this work

Here we discuss what we have learned about what needs to be done for similar work to be made tractable

The minimal requirement for encoding links to and from a grammar would be that there be consistent markup in the source document. As we have seen, this consistency can simply be in the way in which references are created, for example, marking sections by using the character §, or always referencing textual lines with three-part identifiers (e.g., 3.4.1). A more fundamental issue that underpins any links, is what are the primary objects being addressed. In media, these are timecoded chunks, in a dictionary it could be headwords or senses, in a grammar it is tables, sections, and chapters, with internal items identified as required for a particular work (e.g., phonological rules), and in a text collection it can be any unit which has a unique identifier in the encoded format.

The linking between grammatical description and text examples which we are instantiating was set out explicitly by Heath. We envisage that a future workflow could exploit tagging of text units to assist in building such links: the researcher would add tags to sections of text indicating that this was a good example of some grammatical feature or construction. This would then allow easy generation of sets of examples which could illustrate the description of the phenomenon. Working along such lines would make explicit the inductive linking of data and description so that a grammar would (at least in part) grow as an integral part of the linguist's workflow.

We used the TEI as a mechanism to identify for processing points within the documents that would act as anchors and targets, and to identify Nunggubuyu text in order to allow a switch between Heath's orthography and the current orthography. The TEI is not a simple set of tags, it is the accreted result of many years of ad-hoc decisions by its creators and includes many idiosyncratic tags and tag hierarchies. Despite the length of time it has been available (since 1994), there are few useful presentation systems available for TEI documents. For those who decide to persist with using the TEI, a major benefit is the real time validation against schemas provided by the use of software editors like, for example, Oxygen[3].

We see our work as being a part of the development of rich data ecosystems associated with linguistic research and especially with language documentation and description. In such

[3] https://www.oxygenxml.com/

an environment, there would be more than just the published material: there would be the *assemblage* of all records created, from which are created *collections*, and then the *corpus*, and from that the *published texts* [4]. So the work we are doing provides a means for navigating all of these, but only if they are accessible.

We believe that the Heath material which we have now made available as a densely interlinked web resource already demonstrates the value of our approach. Such presentation will be accessible to future linguists provided they are prepared to work with formats (and tools) which are based on the underlying content of grammatical description and its abstract structures. As we are showing, paths from such formats to the hypertext presentation are not hard to create. Most of the labour has been in moving from the existing form-focused presentation to explicit markup and that work has inevitably involved a large amount of handicraft. We hope that the benefits of the type of presentation we are developing will be sufficient to encourage our colleagues to prepare their data in well-structured formats. Further, we look forward to new techniques for representation of that work that will enable a generalisable navigation of interlinked documents such as those we have created in this project.

Acknowledgments

Thanks to Brett Baker, Kate Horrack, Peter Nyhuis, and the Australian Institute of Aboriginal and Torres Strait Islander Studies.

References

Ameka, Felix K, Alan Charles Dench & Nicholas Evans (eds.). 2006. *Catching language: The standing challenge of grammar writing.* Berlin ; New York: Walter de Gruyter.

Bevilacqua, Ann F. 1989. 'Hypertext: Behind the Hype' *American Libraries*, Vol. 20, No. 2: 158-162

Bow, Cathy, Baden Hughes, and Steven Bird. 2003. Towards a general model of interlinear text. Paper presented at EMELD 2003, Michigan State University.

Comrie, Bernard & Norval Smith. 1977. *Lingua descriptive studies: questionnaire. Lingua.* 42(1): 1-72.

Heath, Jeffrey. 1980. *Nunggubuyu myths and ethnographic texts.* Canberra: Australian Institute of Aboriginal Studies.

Heath, Jeffrey. 1982. *Nunggubuyu dictionary.* Canberra: Australian Institute of Aboriginal Studies.

Heath, Jeffrey. 1984. *Functional grammar of Nunggubuyu.* Canberra: Australian Institute of Aboriginal Studies.

Heath, Jeffrey. 2017. *A grammar of Najamba (Dogon, Mali)*. Language Description Heritage Library (MPI). Electronic publication. DOI: 10.17617/2.2397771

Lewis, W. D. & Xia, F. 2009. Parsing, Projecting & Prototypes: Repurposing Linguistic Data on the Web, in *Proceedings of The 12th Conference of the European Chapter of the Association of Computational Linguistics (EACL),* Athens, Greece, March 2009.

Matter, Florian. 2020. Integrating grammatical description, text collection and dictionary: Language documentation and description for the digital age. https://osf.io/3vm6u/wiki/home/

McCarty, Willard. 2017. The Analytical Onomasticon: An auto-ethnographic vignette. Unpublished paper. http://www.mccarty.org.uk/essays/McCarty,%20The%20Analytical%20Onomasticon.%20An%20auto-ethnographic%20vignette.pdf (27 September, 2020).

Morey, Stephen. 2005. *The Tai languages of Assam : a grammar and texts*. Canberra : Pacific Linguistics.

Musgrave, Simon & Nick Thieberger. 2012. Language description and hypertext: Nunggubuyu as a case study. In Sebastian Nordhoff (ed.), *Electronic Grammaticography* (Language Documentation & Conservation Special Publications 4), 63–77. Honolulu: University of Hawai'i Press.Nelson,

Theodor H. 1965. Complex information processing: a file structure for the complex, the changing and the indeterminate. *Proceedings of the 1965 20th national conference*, 84–100. ACM.

Nelson, T.H. 1982. *Literary Machines 93.1* Sausalito: Mindfulness Press.

Thieberger, Nicholas. 2001. As we may link: time-aligned concordances of field recordings. A working model. Online paper presented at

[4] This distinction is made by Jane Simpson in this blog post https://www.paradisec.org.au/blog/2018/02/texts-and-more-texts-corpora-in-the-coedl/

Computing Arts 2001: Digital Resources for Research in the Humanities

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

Thieberger, Nick. 2009. Steps toward a grammar embedded in data. Epps, Patricia & Alexandre Arkhipov. (eds.) *New Challenges in Typology: Transcending the Borders and Refining the Distinctions.* Berlin; New York, NY: Mouton de Gruyter Mouton. 389-408.