Steps toward a grammar embedded in data

Nicholas Thieberger

1. Introduction[1]

Inasmuch as a documentary grammar of a language can be characterized – given the formative nature of the discussion of documentary linguistics (cf. Himmelmann 1998, 2008) – part of it has to be based in the relationship of the analysis to the recorded data, both in the process of conducting the analysis with interactive access to primary recordings and in the presentation of the grammar with references to those recordings.

This chapter discusses the method developed for building a corpus of recordings and time-aligned transcripts, and embedding the analysis in that data. Given that the art of grammar writing has received detailed treatment in two recent volumes (Payne and Weber 2006, Ameka, Dench, and Evans 2006), this chapter focuses on the methodology we can bring to developing a grammar embedded in data in the course of language documentation, observing what methods are currently available, and how we can envisage a grammar of the future.

The process described here creates archival versions of the primary data while allowing the final work to include playable versions of example sentences, in keeping with the understanding that it is our professional responsibility to provide the data on which our claims are based. In this way we are shifting the authority of the analysis, which traditionally has been located only in the linguist's work, and acknowledging that other analyses are possible using new technologies. The approach discussed in this chapter has three main benefits: first, it gives a linguist the means to interact instantly with digital versions of the primary data, indexed by transcripts; second, it allows readers of grammars or other analytical works to verify claims made by access to contextualised primary data (and not just the limited set of data usually presented in a grammar); third, it provides an archival form of the data which can be accessed by speakers of the language. The great benefit for typologists is that they can access annotated corpora from fieldwork-based research in order to address topics not considered by the original analyst. Well-structured linguistic data will allow a number of outputs to be created, including a grammatical description writ-

ten as a book, but with other potential ways of visualizing the data. The citable archival form of the data can be derived in more ephemeral multimedia or online forms suited to 'mobilising' the data (Nathan 2006). For example, archival files are typically very high resolution and are accordingly very large. To deliver them via the web or on a CD or DVD attached to a grammar requires that they be converted to a lower resolution. Similarly, it is now common to create lexical databases from which a dictionary can be derived. This separation of underlying forms of the data from delivery forms is central to the methods discussed in this chapter.

Writing a grammar of a previously undescribed language is a major undertaking, typically the endpoint of fieldwork-based linguistic research, of which the methodology has recently undergone significant changes with the introduction of new tools for digital recording, transcription and analysis. The process of recording such a language has, until recently, relied little on an empirical dataset and more on the genius of the researcher – observing and writing notes while they live in a community of speakers. The resulting work, the grammar, is a crafted collection of these observations and, more often than not, it is only fieldnotes and written texts that are recorded, with perhaps a few tapes to confirm phonological claims (cf. Dixon 2006). The innovative approach discussed in this chapter does not supersede the linguist's role, it enhances the scientific basis of the linguist's work by providing ready access to primary data. The decisions about what to record, how to record it, and all the normal elicitation and experimental techniques that make up fieldwork today are still implemented by the linguist, but it is in the methods of recording, naming and transcribing the field materials and consequent access to data that novel outcomes can be achieved. An opportunistic corpus cannot answer all questions asked by a grammar writer and there will always be a need to elicit forms, especially for paradigms. Such forms do, of course, need to be marked as being elicited rather than naturally occurring so that their status is clear.

The typical grammar of the past few decades makes no reference to the source of its data nor to how to access further data on the language than is included in the grammatical description. For example, was the data all elicited or was it recorded and transcribed? If it was recorded, then who was it recorded with – are the speakers old or young, male or female? If texts are the source of example sentences, then where in the text does the example come from? Where is the data itself stored?

A sample of some thirty grammars from that period found one (Heath 1984, discussed below) that provided sufficient data to allow verification

of the author's analysis with textual data readily available to investigate features of the grammar not addressed by the grammar writer. None of them provided recordings and none provided links from example sentences to the context of the sentence, neither by provision of the media, nor by the more arduous path of spelled out timecodes to media files that may be available somewhere in the world (but not collocated with the grammatical analysis). It is appropriate to focus on the past few decades because it has been possible to provide access to textual and dynamic media recordings for most of that time, but it has not been part of normal linguistic methodology to take advantage of this possibility. Thus it is not a criticism of any one grammar to observe that it does not consider the corpus on which it is built to be a relevant part of its construction so that the whole corpus or a suitable presentation version of it should be provided to the reader. Rather, one has a sense of wonder that field linguistics as a discipline has kept going as long as it has in willful ignorance of the availability of new methods for recording, transcribing, concordancing, annotating, and presenting the data on which it bases its generalizations. These are methods with which it should be completely engaged, relishing the opportunity to access recorded textual material instantly, and to account for any small inconsistency in the data by reference to that data rather than by sidestepping one or two seemingly aberrant forms in the transcripts because of the difficulty of locating the primary media in an analog tape. Not only do these methods improve the linguist's work, but they also create the kinds of records that speakers of the languages can reasonably expect to result from fieldwork, and that funding bodies are increasingly coming to demand of publicly funded research.

2. The art of grammar writing in recent literature

Two recent works on grammar writing summarise the state of the art, but neither considers the possibilities offered by new technologies for access to primary data. Thus, in a collection of work which details many aspects of grammar writing, Ameka, Dench, and Evans (2006) briefly discuss the issue of new technological methods for accessing data, but conclude that it means that data should be made available by a digital archive (Evans and Dench 2006: 25). Archiving data ensures its longevity; however, it is the relationship of the grammar to the data that ideally forms the basis of the analysis engaged in by the grammar writer. Archived data provides the foundation for this relationship, and archiving is a necessary but not suffi-

cient activity to ensure both that linguistic analysis is embedded in the data, and that there can be long-term access to the data. Mosel (2006), in the same volume, hopes that every grammar would include a text collection which:

> consists of annotated digitalized recordings of different language genres (e.g. myths, anecdotes, procedural texts, casual conversation, political debates and ritual speech events), accompanied by a transcription, a translation and a commentary on the content and linguistic phenomena. (Mosel 2006: 53)

New technologies provide the means for creation of digital records in the course of linguistic fieldwork and analysis, and this requires a change in linguistic methodology, as discussed below.

Another recent collection of papers on grammar writing (Payne and Weber 2006) makes no reference to the potential of a new kind of grammatical description interoperating with its source data via the use of new technologies, despite two chapters touching on technology in grammar writing (Weber 2006a, 2006b). Weber's (2006b) discussion of the linguistic example likens a grammatical description to a museum of fine art, with galleries exhibiting the features of the language, so there could be a gallery of relative clauses, a gallery of noun classes and so on. In these galleries the example sentences form the exhibits. He points out, however, that a museum is not a warehouse – suggesting that data collection on its own, as advocated by some proponents of language documentation, simply results in a warehouse of recorded material, in contrast to a museum in which each item needs to be provided with interpretive material.

Similarly, I suggest, a grammatical description must be based on a corpus (the warehouse) of catalogued items (time-aligned transcripts, recordings, and so on), but use examples to illustrate given points within the description. To go further with the analogy, a museum provides a catalogue of the huge warehouse collection that underlies the few items displayed in a particular gallery. In most grammars to date, the example is the only language data provided, and there is no catalogue of the rest of the data, nor an indication of the relationship of the example to that data. As Weber notes, we must "keep in mind that some day the examples may be appreciated more than the author's fine words giving some clever analysis or theory" (Weber 2006b: 446). In a similar vein, Dorian observes that:

> The only real certainty about data is that you never have quite as much of it as you'll someday wish you had. If you have an inconvenient or puzzling datum that you don't know what to do with, it's wise to put it in an extended footnote or into an appendix, where it remains accessible and at least marginally on the record, for your own future use or others'. (Dorian n.d.: 18)

How much more useful will it be when grammars of little-known languages provide not just decontextualised and unsourced examples, but annotated corpora linked to media? In this way the grammar provides a point of entry to a set of data and, it is to be hoped, allows new analyses to be made by other researchers who are able to locate unusual examples that eluded the initial recorder.

The data on which a grammatical description of a language are based are necessarily a partial set of recordings, observations and elicited forms. They are partial because we are sampling whatever we can within a short timeframe, all the more so when the work is part of a Ph.D. and so constrained to the period of a student's candidacy. Nevertheless, the analysis of this set of data can be replicable if it is clear which elements of the data form the basis of each analytical claim. Thus, an example sentence, the archetypal 'proof' used in a contemporary grammar, has to bear a considerable burden; not only will it provide the authority for the current claim, but it will inevitably then be reused in other work. It is decontextualised first by the original linguist who has some knowledge of the language and the frame from which the example was taken, but is then decontextualised further by other researchers for whom it is an example of a phenomenon, regardless of the context from which it was taken. These example sentences then take on a life of their own as tokens of authenticity of a particular theoretical point, a problem that could be ameliorated if the example was properly provenanced to source data in a digital repository.

Why is there this lack of engagement with the possibilities offered by new technologies for grammatical descriptions? A major reason is the lack of very simple-to-use tools and hence the slow uptake of existing tools among the community of researchers. I suggest that several other reasons conspire to keep data out of grammatical descriptions. They include: inertia, the reluctance of academia to change the way it has been conducting itself; a so-called 'theoretical linguistics' that is satisfied with minimal data selected to prove foregone conclusions and which hence requires no large datasets (see Beaugrande 2002); a perception that the creation of a corpus on which to base claims is too time consuming and therefore cannot be part of a normal fieldwork-based linguistic investigation; a general antipathy to

technology among humanities researchers (but linguists have often been more technically engaged than the typical humanities scholar); a fear of presenting evidence on which claims are made and hence exposing one's analysis to scrutiny; and, related to the last point, a desire to be the unquestioned authority whose word must be accepted in the absence of any corroborating evidence. The recent increasing interest in the use of corpora (witness the use of the World Wide Web as a corpus from which Natural Language Processing applications opportunistically harvest material with commercial implications) and in language documentation methodology (see Himmelmann 1998, several contributions in Gippert, Himmelmann and Mosel 2006, EMELD conferences 2001–2006,[2] Austin (ed.) 2003, 2004, 2005, 2006, Woodbury 2003), has led to an opening up of discussion of a new analytical approach for field linguistics.

3. What could an embedded grammar be?

How could a grammar embedded in data be conceived? Grammars can be characterized as a range of types, beginning with those for which the source data is not mentioned through to a complex dataset in which the grammatical description forms just a part of the whole representation of the language. An ideal embedded grammar would allow the reader to move between the apparatus of the grammar and the source data, using either as a point of entry. So, from watching a video of the performance of a story and seeing its orthographic representation on the screen, the reader may want to find out about the meaning of a particular word. They will be able to link to and view the dictionary entry. They may want to know about the role of a participant and link to a discussion of arguments, or of case roles, or of the pronominal system, depending on the context of the departure point in the text.

Constructing such links is hugely time consuming, but allowing them to fall out naturally from a well-structured set of data is a goal that is worth pursuing. However, such data structures are not yet developed, so there are no tools made easily available to the ordinary linguist. In the meantime, we operate on the assumption that the more explicitly we can structure the data we create, the more likely it is to survive and be interoperable with emerging systems. Thus the basic desiderata for a grammar embedded in data must also conform to the principles of portability set out in Bird and Simons (2003). The process of creating such data will be discussed below.

At the more recognizable end of the spectrum of grammars, those that have been produced more recently usually include a number of example sentences. Examples are the main currency of a descriptive grammar and the usual form of argumentation is to provide some analysis of a phenomenon and then the example that illustrates its existence and its usage. However, if examples are not provenanced and their status not made clear to the reader (Is the sentence elicited? Was it part of a larger text? Who was it spoken by?) then they provide poor data for others who may want to test the analysis.

A first step towards an embedded grammar could be to provide links from the grammar to sources of data, establishing that examples exist in a corpus of texts, as shown by Heath in his Nunggubuyu Grammar (1984), Dictionary (1982) and Texts (1980). Cross-references in these volumes constructed by hand (that is, without using a computer to generate links in the way that we now can do) allow the reader to locate contextual information for example sentences. Heath notes that:

> The standards of accuracy and documentation which I have set for myself in preparing this volume have been high, though I may not have lived up to them uniformly. In essence this is a corpus-based grammar, and my ideal has been to account for all or nearly all instances in the texts of each morpheme or other feature under consideration. (Heath 1984: 4)

Heath accordingly presents references to many examples of any morpheme discussed in his grammar, with the result that, as noted by Musgrave (2005: 113), the complexity of referencing in Heath's typescript presentation makes it difficult to read, with a dense use of superscripts and visual references that could today be replaced by less obtrusive hyperlinks. Heath (1984: 5) discusses the need for documentation because of his own "sad experiences as a reader of other linguists' grammars, which have almost never provided me with the information I wanted to undertake my own (re)analysis of the language in question." Heath's work cited but did not provide taped recordings, but an early example of the provision of textual material with a recording is Brandenstein's (1970) text collection, which included a 45 rpm vinyl disk containing the stories which were provided in interlinear format in the book. He clearly thought it essential for a modern comprehensive linguistic work that a record should be attached to each copy of the book (Thieberger 2008: 327). Last century, when we were dealing with analog data, it could be linked from texts but it took far more effort than is required for digital recordings today. A further, but unpub-

lished, example of textual material accompanying a description is Nicholas Reid's (1990) dissertation on Ngan'gityemerri, which included four audio-cassettes.

Citation of data from primary media requires that the media have persistence, both in location and naming, so that when the speakers want access to recordings they can find them, and when a researcher wants to listen to an example sentence to find out if there is a feature there that the original work ignored, they can do that. However, such persistence is difficult, if not impossible, for the individual researcher to achieve (hence the need for linguistic digital repositories such as the DOBES archive,[3] PARADISEC,[4] or AILLA).[5] The researcher may also create a website or a multimedia package in which the media is available for public use. Websites created to provide access to primary data typically do not have longevity, and issues related to delivery of media over the web mean that high-resolution primary data is not suited to web-delivery. Multimedia packages as they are currently being made have a very short life, usually relying on software that is updated every year or so, leaving orphaned earlier versions unreadable. Musgrave (2005) presents three case studies of multimedia representations of language,[6] each of which includes little grammatical information, concentrating instead on texts linked to lexicons and media files. The presentation of linked texts, media and lexica is to be welcomed; however, it seems that, of the three case studies, it is only for Nahuatl (Amith n.d.) that the underlying data is produced in an archival form. This is important when creating multimedia products, as otherwise there is a risk that the only available data for the language will be in a form that cannot easily be archived and thus may not be available in the long-term. From the perspective of writing a grammar embedded in data and conforming to the principles of portability (mentioned earlier), these three projects are not descriptions presenting primary data but are *representations* of the data, an important distinction that must be understood in order to create well-formed primary data in an archive, with derived versions being used in multimedia representations.

Morey's (2004) grammar of Tai includes a CD-ROM which contains the text of the grammar, with example sentences linked to audio files and to details about the characteristics of the speaker. In this work the links are all produced by hand and each link is to a single media file (each target utterance corresponds to a single file). Morey, in a postscript to the work (and echoing Weber's sentiments quoted above) notes that, "in a hundred years time, though aspects of my analysis may not have stood the test of

time, I am confident that the richness of the corpus of texts will ensure that this work is still useful" (2004: 402).

With similar sentiments, I built a corpus of South Efate texts on which my analysis (Thieberger 2006a) of the language was based.[7] The field recordings were not segmented into utterance-length files as a matter of principle – if they were providing contextualised data for my analysis, it would defeat the purpose to cut them into potentially decontextualized units, and, from an archival point of view, there should not be too many objects (files) to keep track of and to describe. The focus was to develop a means by which a collection of transcripts could maintain links back to the media they transcribed, allowing any subsequent mentions of parts of the texts (as in example sentences) to be cited to that media file. This then is a second step toward an embedded grammar: using a media corpus as the source, and providing the media files with names that will endure over time (also known as *persistent identification*), as provided, for example, by their being lodged in an archive. This means that any reference made to the media files will be resolvable by readers of the grammar into the future, as per Morey's and Weber's desiderata quoted earlier. The paper grammar, in the form of a book, has references to citable data, and may include a DVD of data (that is, an mp3 or similarly compressed form of the file suitable for delivery) derived from the archival form that is stored in an archive with a persistent identifier (a name that will not change).

What was required to create a corpus with examples cited to the level of words or sentences? To summarise, in my work with South Efate I created archival digital media files, paying attention to accepted standards in filenaming (in order that files have persistent identification and thus persistent location over time) and file formats (so that the files can be read over time), and depositing these files with a digital repository (with sufficient metadata to describe the content of the file). These files were then transcribed with time-alignment (using stand-off timecoding) (the process outlined here is discussed in more detail in Thieberger 2004, 2006b) facilitating links to be instantiated between any textual chunk and its media representation. Any sentence in some twenty hours of data can be clicked and heard, and accessed by a concordance (a listing of each word in the data in context). While the software for transcribing with time-alignment has been available since the late 1980s (see MacWhinney 1996), it was necessary to devise a method to access the audio and video data as a corpus in order to facilitate its analysis, as no similar approach had been developed when it was needed for this research. Happily this is no longer the case and tools are becoming easier to use and more widely adopted by lin-

guists (in fact, my students now ask how it could have been possible to do this work in the past without these tools).

I wrote a tool called *Audiamus*[8] to allow me to work on the transcripts of all media files (produced, for example, by Transcriber) as a corpus, and to search and browse through any part of them while maintaining the ability to access the primary media associated with them by mouseclick. *Audiamus* also allows me to quote relevant chunks of the data and to paste these quotes directly into my manuscript and to export it to a number of formats. Thus the references between the textual and media representations of the data are maintained through the process of data analysis. A problem with the current model is that data are not normalized, that is, the same data appear in various parts of the corpus (in the transcript, texts, dictionary and example sentences) and so any change to one version needs to be reflected in the others.

A third step into the future of grammar/data entangling will be the creation of grammatical descriptions whose structure is encoded according to an accepted standard, perhaps using elements determined by the Text Encoding Initiative (Burnard and Sperberg-McQueen 2002) or some similar schema to be determined by linguistic practice as it develops. This would create a document in which all structural elements were explicitly declared – normal and generic elements such as chapters, paragraphs and so on, as well as elements more particular to a linguistic grammar. While some parts of a grammar could use the same markup standards as other texts (essays, theses and so on), there are some particular datatypes in linguistic descriptions that will need standards agreed to by the community who will use them. Work has been done to identify some of these datatypes and strategies for encoding them, in particular paradigms (typically used in representation of pronouns) (Penton and Bird 2004) and interlinear glossed text (IGT) (Hughes, Bird, and Bow 2003, Bird, Bow, and Hughes 2003). An immediate benefit of adopting a standard encoding for these datatypes would be to increase their archivability as underlying data rather than as images of data on a page. This would also mean that a typological search for all members of a given class (first person pronouns for example) could be executed via the structure of the data. Another example using interlinear glossed text would allow searches over form/meaning pairs within texts, currently difficult if not impossible to do due to the lack of an explicitly encoded relationship between lines of interlinear glossed text. A similar focus on conceptualising the structure of a grammar can be found in Good (2004) and Nordhoff (2007), but neither has yet resulted in a workable

framework, let alone replicable methods or tools for construction of interoperating grammars and data.

A further motivation for building an explicit structure for a grammatical description is that it will provide automatically comparable data across grammatical descriptions in addition to an archival form of the text. At the moment our expectations of an archival linguistic text are very low, as there is a long way yet to go in convincing the linguistic community of the necessity to create functioning digital versions of textual documents in addition to paper outputs. A well-formed and schema-based XML version of a grammar will allow remote (that is, internet-based) searching over specific types of data, of the kind that typologists would like to do but are currently finding difficult due to the lack of a standard encoding schema for grammars. At present a grammar, if it is available on the internet, is an undifferentiated collection of textual material, most likely to be a pdf file, and with no distinguishing characteristics that allow it to be identified as a grammar by a search engine except that its title may include the word 'grammar'. If there were a schema associated with the document type 'grammar' then a search on that type would locate only grammars, and only those with a predictable structure able to be searched remotely. The remaining issue of an agreed ontology for linguistic terms is currently being investigated.[9]

With the lack of data structure in current written grammatical descriptions it is not possible to automate links between various parts of a linguistic corpus and annotation set. However, a schema-based grammatical description could provide a normalised data structure in which a change to an item in one place (for example, the lexicon) would be reflected in all mentions of that form throughout the corpus, texts and grammar.

A schema for the representation of IGT has been developed, following both the work discussed above and an XML-based presentation system developed by Michel Jacobson at LACITO.[10] To begin, this project established an XML schema for a minimal five tier interlinear text, with provision for the text line, morphemic and gloss lines, a free translation, and a reference to a media file (together with a small set of metadata fields to assist in discovery of the text). This is a schema-based XML system for making explicit the relationship between parts of interlinear texts together with links to the source media, and has resulted in an online tool (called EOPAS, see Schroeter and Thieberger 2006) for presentation of interlinear texts, which allows searching and concordancing linked directly to the media. EOPAS also provides the environment into which linguists can upload files in a Toolbox[11] export format for online viewing and playback

of media linked to transcripts. The resulting tool is portable, allowing other initiatives to capitalise on the work and potentially develop it in different directions. The ultimate aim of this approach is to allow new perspectives on the data itself, provided by contextualised access to primary data, and then to allow new research questions to be asked, and richer answers to be provided, all in a fraction of the time that it would have taken with analogue data.

We can envisage a set of XML-based tools that operate over a richly annotated corpus, using schemas for various datatypes to understand how to read and interact with them. Thus when a snippet of IGT is located in a grammar, the browser uses the schema definition of IGT to present it and to instantiate the link to media encoded in it. The schema also allows us to browse all IGT in the grammar and to search for texts in which a target structure occurs, as we can be confident that the formats in use are all validated against a schema.

A fourth step to an embedded grammar would be one in which each hypothesis could be computationally tested by the reader – assuming that the particular theoretical model of analysis employed is computationally tractable. For example, PARGRAM (Butt, Flickinger, and Oepen 2002) implements a Lexical Functional Grammar while Bender (n.d.) and Bender et al. (2002) describe a system they call the LinGO Grammar Matrix which includes an HPSG computational grammar, and Bender (n.d.) and Bender et al. (2004) summarise various grammar engineering projects and the development of a proposed model called Montage (see also Good 2004) in which they "hope to bring the hypothesis-testing power of grammar engineering to linguists engaged in primary linguistic documentation" (Bender n.d.: 15). From the point of view of primary grammatical descriptions of endangered languages, written as most are in a fairly eclectic and certainly not mono-theoretical model (as discussed for example in Dryer 2006), it is unlikely that this kind of grammar will appear anytime soon. It is more likely that it will be derivative of a more narrative form of grammar, one that I hope is of at least the second or third steps described earlier.

4. Building documentation into normal field methods

A major objection to creating even the most minimal dataset as the basis for a grammatical description is that it takes too long and so detracts from the analysis (this is especially the case for a student writing a Ph.D. thesis). As long as it is not rewarded by the linguistic community, the preparation

of archival data linked in to a grammatical analysis will only be carried out by those who have sufficient technological skills, who recognize its value to posterity and are committed to the scientific method in their own work. We clearly need more accessible tools and methods, but there are sufficient tools now being used by new researchers and the methods are established well enough that it does not add significantly to the linguist's workload to create time-aligned transcription, adequate filenaming, metadata, and to archive the primary materials (if we accept that creation of a representative corpus of recorded material is a minimal outcome of our work, as argued by proponents of language documentation, e.g. Himmelmann 1998, Woodbury 2003). In fact, it has a number of benefits that reflect a substantial return on the initial investment, including an improved analysis and an ability to keep working on the same dataset over time because of the ease of access to the material. The process of creation of good field records reinforces one's knowledge of the language and enriches the analysis, first because it requires interaction with the data through the initial transcription and annotation, and second because it provides instant access to the media and so allows us to locate examples and to correct mistranscriptions incrementally as we work with the data. As Beaugrande (2002: 105) points out, we are 'seeing' phenomena in language which only become visible through the technology. For example, my Ph.D. thesis is available in an eprints repository,[12] which means that the text can be searched via Google or a similar search engine. Robert Eklund, studying ingressives in the world's languages, located a reference to the ingressive in the thesis and emailed a query. He had doubts that it actually was a pulmonary ingressive, especially as the example included [s], a phone he thought cannot be produced ingressively, at least not without considerable acoustic changes. The segments in my sound files were easily located and within ten minutes he had a clip of the sound so that he could confirm that it was indeed a pulmonary ingressive (see Eklund 2007).

Another example of the research benefit of investing effort in creating an accessible dataset is from the pre-verbal benefactive phrase in South Efate, which can be represented by a single morpheme in the form of a possessive pronoun, or by a complex prepositional phrase (Thieberger 2006c). The possessive pronouns in some cases differ minimally in form from the subject pronouns (e.g., *ag* '2sgS', *gag* '2sgPOS').[13] In the initial transcriptions, and before it was clear to me that this was a benefactive construction, both I and my South Efate colleague had occasionally mistranscribed these forms. The resulting transcript thus provided ambiguous data about the form of the pronoun in this position. The ability to instantly

access the recordings and to confirm that the pronoun is a possessive form allowed the analysis to be more robust and the transcript to be incrementally corrected. Thus, with an improved analysis we are able to update the transcription via the digital media, rather than being constrained by a once-only analog transcription system in which the media are no longer accessed and the transcript is taken to be the primary authority.

Another motivation for creating digital data using these methods is that it is easier to repatriate or to return copies of this material to the speakers. Recently, while at my fieldsite in Erakor village, I was able to respond to a request from the widow of my main language teacher to provide an audio CD of fifteen tracks, including stories and songs from a range of villagers, some now deceased. It took a few hours to select and locate the data and then produce ten copies of the CD with cover notes. I subsequently placed a collection of narratives, songs and hymns into an ITunes[14] installation on the village school's computer so that others in the village can burn their own selections to CD. This was only possible because of the well-structured nature of the data and metadata and illustrates the way in which new methods can make us more responsive to community requests for information; especially valuable at a time when academia is increasingly being called to account by its funding sources.

The process of discovery of linguistic facts proceeds best by constant interaction with the corpus. Noting that certain features exist in the data, by means of tagging the textual annotation, allows those examples to be retrieved later. The nature of our language learning varies from individual to individual but typically includes achieving plateaux of understanding of structures in the language, which then become an operating framework. New data can challenge these frameworks and allow us to supersede previous working hypotheses if we can interact with the data in real time. New insights can thus be confirmed and incorporated into the corpus and the analysis.

5. Conclusion

Having established a methodology for recording and linking media and transcriptions as part of normal fieldwork, our corpus will allow all derived work, including texts, example sentences or dictionaries to be verified by reference to primary recordings. An academic grammar of a language that is created using the methods described in this chapter will have multiple possible outputs in addition to the grammar, and the data created using

these methods will be available for others to access in the future. A book written using these methods will allow the reader to access media, either on an accompanying disk or via web delivery. The underlying data will be safely stored in an archive that guarantees longevity of the unique recordings made as part of the grammar-writer's effort.

Notes

1. An early version of this paper was presented at the Australian Linguistic Society in July 2008 and has benefited from discussion with members of that audience. It has also been improved by comments from Pattie Epps, David Nash and an anonymous reviewer. Work reported on in this chapter was supported by the following Australian Research Council grants: SR0566965 – Sharing access and analytical tools for ethnographic digital media using high speed networks; DP0450342 – New methodologies for representing and accessing resources on endangered languages: a case study from South Efate.
2. http://www.emeld.org/events/index.cfm
3. http://www.mpi.nl/DOBES
4. http://paradisec.org.au
5. http://www.ailla.utexas.org/site/welcome.html
6. The three exemplars are: Amith, Jonathan D. (n.d.), Nahuatl Learning Environment <http://nahuatl.ldc.upenn.edu/>; Csato, Eva A. and David Nathan (2003), Spoken Karaim (CD-ROM); Nikolaeva, Irina and Thomas Mayer (2004), Online Documentation of Kolyma Yukaghir, <http://ling.uni-konstanz.de/pages/ home/nikolaeva/documentation/intro.html>.
7. South Efate is an Austronesian Oceanic language of central Vanuatu. My fieldwork began in the mid 1990s and has so far resulted in a grammar, dictionary and texts in the language, as well as an archived corpus of recordings.
8. http://www.linguistics.unimelb.edu.au/thieberger/audiamus.htm
9. See, for example, preliminary results at http://www.linguistics-ontology.org/
10. http://lacito.vjf.cnrs.fr/archivage/index.htm
11. Toolbox is the best current software for entering and working with structured interlinear data (http://www.sil.org/toolbox).
12. http://repository.unimelb.edu.au/10187/1318
13. South Efate orthography uses /g/ to represent the velar nasal.
14. ITunes is commercial but free software produced by Apple Inc. (http://www.apple.com/itunes/).

References

Ameka, Felix, Alan Dench and Nicholas Evans (eds.)
2006 *Catching Language: The Standing Challenge of Grammar Writing*. Berlin: Mouton de Gruyter.

Amith, Jonathan D.
n.d. Nahuatl learning environment <http://nahuatl.ldc.upenn.edu/>

Austin, Peter (ed.)
2003 *Language Documentation and Description.* 1. London: SOAS.
2004 *Language Documentation and Description.* 2. London: SOAS.
2005 *Language Documentation and Description.* 3. London: SOAS.
2006 *Language Documentation and Description.* 4. London: SOAS.

Beaugrande, Robert de
2002 Descriptive linguistics at the millennium: corpus data as authentic language. *Journal of Language and Linguistics.* 1 (2): 91–131.

Bender, Emily M.
n.d. Grammar engineering for linguistic hypothesis testing. Manuscript.

Bender, Emily M., Dan Flickinger, Jeff Good and Ivan A. Sag
2004 Montage: Leveraging advances in grammar engineering, linguistic ontologies, and markup for the documentation of underdescribed languages. In *Proceedings of the Workshop on First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation*, LREC 2004. Accessed at http://faculty.washington.edu/ebender/papers/Montage_LREC.pdf on 14/9/2008.

Bender, Emily M., Dan Flickinger, and Stephan Oepen
2002 The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, *Taipei, Taiwan*. J. Carroll, N. Oostdijk, and R. Sutcliffe (eds.), 8–14.

Bird, Steven, and Gary Simons
2003 Seven dimensions of portability for language documentation and description. *Language* 9 (3): 557–582.

Bird, Steven, Catherine Bow, and Baden Hughes
2003 Towards a general model of interlinear text. In *Proceedings of the EMELD Language Digitization Project Conference 2003: Workshop on Digitizing and Annotating Texts and Field Recordings*. Accessed at http://www.emeld.org/workshop/2003/bowbadenbird-paper.pdf on 14/9/2008.

Brandenstein, Carl Georg von
1970 *Narratives from the North-West of Western Australia in the Ngarluma and Jindjiparndi Languages.* (Volumes1–3 + audio-disc.) Canberra: Australian Institute of Aboriginal Studies.

Burnard, Lou, and C. Michael Sperberg-McQueen
2002 TEI Lite: an introduction to text encoding for Interchange. Accessed at http://www.tei-c.org/Guidelines/Customization/Lite/teiu5_split_en.html on 23/3/2008.

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer
2002 The Parallel Grammar Project. *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*. 1–7.

Csato, Eva A., and David Nathan
2003 Spoken Karaim. (CD-ROM)

Dixon, Robert M.W.
2006 [Dixon accepts Bloomfield Award] SSILA Newsletter Number 236: April 10, 2006: [6].

Dorian, Nancy
n.d. Getting it right: some conscious and unconscious obstacles to getting a language on the record realistically. (Manuscript.)

Dryer, Matthew S.
2006 Descriptive theories, explanatory theories, and basic linguistic theory. In *Catching Language: The Standing Challenge of Grammar Writing*. Felix Ameka, Alan Dench, and Nicholas Evans (eds.), 207–234. Berlin: Mouton de Gruyter.

Eklund, Robert
2007 Pulmonic ingressive speech: a neglected universal? In *Proceedings of Fonetik 2007, May – 1 June 2007*, Stockholm, Sweden, *TMH-QPSR,* 50 (30): 21–24.

Evans, Nicholas, and Alan Dench
2006 Introduction: catching language. In *Catching Language: The standing challenge of grammar writing,* Felix Ameka, Alan Dench and Nicholas Evans. (eds.), 1–39. Berlin: Mouton de Gruyter.

Gippert, Jost, Nikolaus P. Himmelmann, and Ulrike Mosel. (eds.)
2006 *Essentials of Language Documentation*. Berlin: Mouton De Gruyter.

Good, Jeff
2004 The descriptive grammar as a (meta)database. Paper presented at the EMELD Workshop, June 2004. Available at http://www.emeld.org/workshop /2004/jcgood-paper.html.

Heath, Jeffrey
1980 *Nunggubuyu Myths and Ethnographic Texts.* Canberra: Australian Institute of Aboriginal Studies.

1982 *Nunggubuyu Dictionary.* Canberra: Australian Institute of Aboriginal Studies.
1984 *Functional Grammar of Nunggubuyu.* Canberra: Australian Institute of Aboriginal Studies.

Himmelmann, Nikolaus P.
1998 Documentary and descriptive linguistics. *Linguistics.* 36: 161–195.
2008 Reproduction and preservation of linguistic knowledge: Linguistics' response to language endangerment. *Annual Review of Anthropology* 37.

Hughes, Baden, Steven Bird, and Catherine Bow
2003 Encoding and presenting interlinear text sing XML technologies. In *Proceedings Australasian Language Technology Workshop.* Alistair Knott and Dominique Estival (eds.), 105–113. Melbourne, Australia. Accessed at http://eprints.unimelb.edu.au/archive/00000455 on 23/3/2008.

MacWhinney, Brian
1996 The CHILDES System. *American Journal of Speech-Language Pathology.* 5:5–14

Morey, Stephen
2004 *The Tai Languages of Assam: A Grammar and Texts.* Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies.

Musgrave, Simon
2005 Multimedia presentation of grammatical description: design issues. In *Proceedings of the Australasian Language Technology Workshop.* Timothy Baldwin, James Curran and Meno van Zaanen (eds.), 113–119. Accessed at http://www.alta.asn.au/events/altw2005/cdrom/pdf/ALTA200517.pdf on 14/9/2008.

Mosel, Ulrike
2006 Grammaticography: the art and craft of writing grammars. In *Catching Language: The Standing Challenge of Grammar Writing.* Felix Ameka, Alan Dench, and Nicholas Evans (eds.), 41–68. Berlin: Mouton de Gruyter.

Nathan, David
2006 Proficient, permanent, or pertinent: aiming for sustainability. In *Sustainable Data from Digital Fieldwork,* Linda Barwick and Nicholas Thieberger (eds.), 57–68. Sydney: Sydney University Press.

Nikolaeva, Irina, and Thomas Mayer
2004 Online Documentation of Kolyma Yukaghir. Accessed at http://ling.uni-konstanz.de/pages/home/nikolaeva/documentation/intro.html on 14/9/2008.

Nordhoff, Sebastian
2007 Growing a grammar with Galoes. Paper presented at the DobeS workshop, June 14, 2007.

Payne, Thomas, and David Weber (eds.)
2006 Perspectives on grammar writing. *Studies in Language* 30:2.

Penton, David, and Steven Bird
2004 Representing and Rendering Linguistic Paradigms. In *Proceedings of the Australasian Language Technology Workshop*, 123–130. Sydney, Australia. December 2004. Accessed at http://www.alta.asn.au/events/altw2004/publication/04-26.pdf on 14/9/2008.

Reid, Nicholas
1990 Ngan'gityemerri: a language of the Daly River region, Northern Territory of Australia. Ph.D. diss., Department of Linguistics, Arts, Australian National University.

Schroeter, Ronald, and Nicholas Thieberger
2006 EOPAS, the EthnoER online representation of interlinear text. In *Sustainable Data from Digital Fieldwork*. Linda Barwick and Nicholas Thieberger (eds.), 99–124. Sydney: Sydney University Press.

Thieberger, Nicholas
2004 Documentation in practice: developing a linked media corpus of South Efate. In *Language Documentation and Description*, Peter Austin (ed.), 2: 169–178. London: SOAS.
2006a *A Grammar of South Efate: An Oceanic Language of Vanuatu.* Oceanic Linguistics Special Publication, No. 33. Honolulu: University of Hawai'i Press.
2006b Computers in field linguistics. In *Encyclopedia of Language & Linguistics, Second Edition*. Keith Brown (Editor-in-Chief), 780–783. Volume 2. Oxford: Elsevier.
2006c The benefactive construction in South Efate. *Oceanic Linguistics* 45 (2): 297–310.
2008 Language is like a carpet: Carl-Georg von Brandenstein and Australian languages. In *Encountering Aboriginal Languages: Studies in the History of Australian Linguistics*. William M. McGregor (ed.), 321–335. Canberra: Pacific Linguistics.

Weber, David
2006a Thoughts on growing a grammar. In *Perspectives on Grammar Writing*. Thomas Payne and David Weber (eds.), *Studies in Language*, 30 (2): 417–444.
2006b The linguistic example. In *Perspectives on Grammar Writing*. Thomas Payne and David Weber (eds.), *Studies in Language*, 30 (2): 445–460.

Woodbury, Anthony C.
2003 Defining documentary linguistics. In *Language Documentation and Description*. Peter K. Austin (ed.), 1: 35–51. London: SOAS.

University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
THIEBERGER, N

Title:
Steps toward a grammar embedded in data

Date:
2009-06-05

Citation:
THIEBERGER, N. (2009). Steps toward a grammar embedded in data. Epps, P (Ed.). Arkhipov, A (Ed.). NEW CHALLENGES IN TYPOLOGY: TRANSCENDING THE BORDERS AND REFINING THE DISTINCTIONS, (2009), 217, pp.389-408. Walter de Gruyter.

Publication Status:
Published

Persistent Link:
http://hdl.handle.net/11343/26089