Vol. 11 (2017), pp. 157–189
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24731
Revised Version Received: 06 Mar 2017

Putting practice into words: The state of data and methods transparency in grammatical descriptions

Lauren Gawne
Nanyang Technological University, SOAS University of London

Barbara F. Kelly
The University of Melbourne

Andrea L. Berez-Kroeker
University of Hawai'i at Mānoa

Tyler Heston
University of Hawai'i at Mānoa

Language documentation and description are closely related practices, often performed as part of the same fieldwork project on an un(der)-studied language. Research trends in recent decades have seen a great volume of publishing in regards to the methods of language documentation, however, it is not clear that linguists' awareness of the importance of robust data-collection methods is translating into transparency about those methods or data citation in resultant publications. We analyze 50 dissertations and 50 grammars from a ten-year span (2003–2012) to assess the current state of the field. Publications are critiqued on the basis of transparency of data collection methods, analysis and storage, as well as citation of primary data. While we found examples of transparent reporting in these areas, much of the surveyed research does not include key information about methodology or data. We acknowledge that descriptive linguists often *practice* good methodology in data collection, but as a field we need to build a better culture with regard to making this clear in research writing. Thus we conclude with suggested benchmarks for the kind of information we believe is vital for creating a rich and useful research methodology in both long and short format descriptive research writing.

1. Introduction Language documentation and description are closely related tasks, often performed as part of the same fieldwork project on an un(der)-studied language. However, since Himmelmann (1998) we have been encouraged to consider that documentation and description are methodologically different, and that data collected with documentary methods can enable verification of descriptive claims based upon them. We have an increasingly rich set of resources for how to undertake language

E-ISSN 1934-5275

documentation work. This body of literature is discussed in §2 below. We find that it is not clear that linguists' awareness of the importance of robust data-collection methods is translating into transparency about those methods, or the whereabouts of data, in resultant publications.

We believe that such transparency is important, because clear methodological description is a hallmark of reproducible linguistic research (Berez 2015). *Reproducibility* is the aim to provide accountability in research by facilitating access to the underlying data and methods so that other researchers may also reach the same conclusions (Gezelter 2009). While the term arose in computer science (Buckheit & Donoho 1995) it has applicability across a range of sciences, particularly where it is not always feasible to replicate the original conditions of the initial data set. The documentation and description of language is a good example of research that is not easily replicable. This is because it can be difficult or impossible to recreate the context of an utterance that is part of a spontaneous conversation or dialogue, and because a language documentation project is often founded on the long-term trust and collaboration between researchers and speakers. Taking into account the particularities of the language documentation and description research paradigm, our aim in this paper is to understand the state of the art in writing reproducible research in language documentation and description.

Making clear the research methods in data collection allows others to reproduce successful experiences, and makes it easier to assess trends and preferences across the field. Making clear the source of data used in description through transparent and consistent citation of primary materials allows others to assess an existing analysis or make use of documentation corpora to address other topics.

Given the centrality of data and method transparency to the contemporary research focus on data replicability across the sciences (Gezelter 2009), we conducted a survey into the state of the art in methodology and data citation within the genre of descriptive grammars. We did this to see what areas the field is already doing well in, and where our efforts need to be improved, so that we can build structures to support better practice.

In this paper we present a survey of 50 published grammars and 50 grammar-based PhD dissertations, examining how explicitly the authors discuss data collection methodology and cite their data. The publications surveyed were selected from a ten-year period beginning five years after Himmelmann (1998) encouraged the use of language documentation to provide verification for language description. We find that while there are some examples of strong methodologically-driven writing, the majority of authors do not include key documentary metadata or methodological information. The result is that it is often difficult or impossible to verify or reproduce descriptive linguistic claims.

The field of linguistics sits at the intersection of traditional disciplines: we have an interest in asking questions about the nature of humanity, like many other areas of the humanities; we use data to answer questions about society, as do other social sciences; and we are building large datasets that can address not only our own questions, but those of other researchers, as do many researchers in the "hard sciences". Linguistics can benefit from sitting at the juncture of these disciplines. We can also learn from discussions that are happening in data sciences (e.g., Renear et al. 2009, Pröll & Rauber 2013, Starr et al. 2015) about how to ensure that the language data that we develop in our research is as useful as possible in our own research and to future users.

In what follows, we do not seek to lay blame upon individual authors, but rather we wish to demonstrate where, collectively, our field could improve and to give some support regarding how this might be achieved. §2 provides some background on the notion of reproducibility in the language documentation and description literature. In §3 we present the methods of our own study, the publications we examined, and the variables of methodology and data transparency we coded for. §4 presents the results of our survey. We conclude in §5 and §6 with a discussion focusing on how we can encourage researchers working within language documentation and description to take their good research practice and make it clearer in their written research outputs.

2. Background

Recent years have seen a surge in the literature on good fieldwork methodology, giving researchers a range of resources that can help them be more effective and considerate. In this section we do not aim to provide an exhaustive review of this literature, but to simply demonstrate that the documentary linguist of today is provided with a wealth of information with regards to how to conduct fieldwork and collect data. Nonetheless, she is offered comparatively little advice on how to develop practices in both documentation and description that will facilitate clear reporting of methodology and citation of data in any subsequent writing.

In the last decade or so we have seen a range of edited books and monographs on linguistic fieldwork, including Gippert, Himmelmann & Mosel (2006), Crowley (2007), Bowern (2008), Chelliah & De Reuse (2011) and Thieberger (2012). We also have the journal *Language Documentation & Conservation*, which published its tenth volume in 2016, and the (more or less) annual publication of *Language Documentation and Description*. All of these publications cover a wide range of practical and theoretical issues around language documentation. Although the authors of these works have differing opinions on some of the details, it is possible to gain an idea of the general consensus in the field with regards to best practice for the tools and methods used in documenting languages.

There are also a number of volumes on the practice of writing descriptive grammars. Within the last decade or so this includes Ameka, Dench & Evans (2006), Payne & Weber (2007), Dixon (2009), Aikhenvald (2014), and Nakayama & Rice (2014). These volumes provide ample advice about grammar writing, but are, on the whole, not explicit about the need to ensure that research methodology is included as a feature of the grammar, nor are they explicit about the need to ensure that examples used have clear citations to underlying data. Mosel (2006:27, 2014:154) overtly mentions including "fieldwork methods" as a feature of the front matter of a grammar, although does not provide specification as to what should be included. Beyond this, any indication that linguistic methodology should be included in a descriptive grammar remains implicit.

Austin (2013:4) argues for a "broad approach to observation and documentation of the methods, processes, and outcomes of language documentation project". He refers to this process as "meta-documentation", that is, the documentation of documentation projects. While we agree that the expectations as to what should be made note of are currently underspecified, we find the terminology "meta-documentation" to be unnecessary. "Meta-documentation" implies that this is a process that is unique to language documentation, even though in reality accounting for one's methods is a basic expectation in most branches of the social sciences and hard sciences. The details of what constitutes a methodology and research context may vary by discipline, but they exist to make research transparent for assessment of success and aid in replicability and reproducibility.

While there has not been a lot of discussion regarding the transparency of research methods in publications, the field of language documentation and description has been engaged in a long and open discussion regarding the role of primary data in linguistic theory more generally. Himmelmann's (1998) position paper on language documentation is clear on this point: "[Language] documentation [...] will ensure that the collection and presentation of primary data receive the theoretical and practical attention they deserve" (1998:164; see also Himmelmann 2006; Woodbury 2003, 2011; Thieberger 2004, 2009, 2014; Thieberger & Berez 2012, among others).

We believe that it is important to have an expectation that authors of descriptive grammars will make clear the methods by which the data that forms the core of the work is collected, managed and analyzed. As part of this good practice we also believe that it is important to be able to resolve individual examples used in a descriptive grammar to the primary data in a corpus. This ties in with a number of other discussions that are happening around the nature of the language documentation and description workflow, such as the attempts to gain more professional recognition for the building of linguistic corpora (Haspelmath & Michaelis 2014; Thieberger et al. 2016).

In Bird & Simons's (2003) seminal article on portability for linguistic data in the digital age, four of the seven dimensions discussed pertain directly to reproducible research: citation, discovery, access, and preservation. While citation is primary to our focus here, it is only useful in achieving reproducibility if the primary data are preserved in such a way that other researchers are able to discover and access them. Thieberger (2006, 2009) provides one of the strongest endorsements of the benefits of reproducibility in grammar writing. While he outlines the general principles for linking descriptive analysis to primary materials, he notes that the tools and standards for such work are still under-developed. A more specific data structure for reproducible grammar writing is outlined in Maxwell (2012), including XML data structures and a series of parsing engines and tokenizers.

The language documentation literature continues to develop more specific expectations about methodology in language documentation. However, there is still a gap in the literature on descriptive writing, in genres such as grammars. In the following section we present our methodology for analysis in the current study.

3. Methodology In this section we discuss the publications that together formed the dataset that we used in this survey, and the features that we analyzed. Our variables section is quite extensive, as we illustrate the variables we code for with examples of good practice drawn from the dataset. This section can therefore also provide a good provisional checklist for a researcher embarking on writing a linguistic description drawing on primary documentary data.

3.1 Data We surveyed 50 published grammars and 50 PhD dissertations. All of the surveyed items were published or awarded during the ten-year period from 2003 to 2012. As mentioned in the introduction (§1), analysis was based on publication within a ten year period beginning five years after Himmelmann's (1998) call for documentation data to underpin descriptive linguistics. Given the sometimes lengthy publication timeline in academia, we figured that five years was enough time for researchers to take heed of Himmelmann's message, and begin to implement the lessons learned in their own work.

We defined grammars as book-length publications that use field-collected materials such as recordings and fieldnotes to describe the grammar of a language. For published works this is almost always a broad grammatical description. Dissertations occasionally have a more targeted focus on a particular grammatical feature. In our corpus of 50 dissertations, ten focus on a specific aspect of the language, such as Schnell's (2010) work on animacy and referentiality in Vera'a, Nichols's (2011) investigation of aspect in siSwati, and Salffner's (2010) discussion of the role of tone in the phonology, lexicon, and grammar of Ikaan. These works are still based on fieldwork and have a strong descriptive component, often including a sizable broad description of other grammatical features of the language. All published grammars and dissertations selected for this survey are written in English. Throughout this discussion we refer to "published grammars" and "dissertations" separately, and "grammars" when referring to all items surveyed.

The grammars come from a range of publishers (shown in Table 1), and degree granting institutions (shown by country in Table 2), covering a range of languages (countries of languages shown in Table 3). No PhD dissertation was additionally included here as a published grammar, even though some authors had gone on to publish their dissertations. No duplicate authors were included (for example, if they had published grammars on more than one language).

Published grammars were selected in such a way as to provide a balance of year of publication, publisher, location of topic language, and accessibility of books for the purposes of the present article. The distribution of published grammars is heavily skewed toward publishers with strong traditions of printing descriptive grammars (Table 1). The category of "other major publishers" in Table 1 includes publishers with global distribution; among them are Brill, John Benjamins, Cambridge University Press, and Routledge. The "other minor publishers" category includes smaller institutional presses; among them are Academia Sinica, University of Hawai'i Press, and small commercial publishers that do not necessarily specialize in peer reviewed academic publication.

Table 1. Distribution of publishers

Publisher	Number of grammars
Mouton de Gruyter	20
Lincom	10
Pacific Linguistics	6
SIL	3
Other major publisher	5
Other minor publisher	6
Total	50

For the dissertations, the distribution of institutions skews heavily toward the USA, Australia, and other developed countries, as shown in Table 2. This is for three reasons. The first is that institutions in these countries have a long history of linguistics training that includes the possibility of writing a descriptive grammar as a dissertation. The second is that dissertations submitted in these countries are most often written in English. The third is that institutions in these countries are leading the way in ensuring that dissertation research becomes publicly available, for example, through institutional repositories and services like ProQuest[1] in the United States: a practice we wholeheartedly support. Dissertations are historically not as readily available as published grammars, although this is changing as more universities create digital catalogs of completed dissertations.

Table 2. Distribution of PhD award institutions

Country	Number of dissertations
USA	21
Australia	11
Canada	3
UK/Europe	12
Other	3
Total	50

The languages described in the grammars come from a range of families and geographic locations. Table 3 shows the countries of the languages represented in the published grammars and dissertations. Of course, not all languages conform to geopolitical boundaries; in the case of languages spoken in more than one country, we have chosen the country mentioned by the author. This table indicates that the dissertations and published grammars surveyed display a variety of work across countries and areas with high levels of linguistic diversity.

[1] http://www.proquest.com/

Table 3. Distribution of target language location

Country	Number of publications
India	7
China	7
Papua New Guinea	6
Australia	6
Vanuatu	6
Nepal	5
Russia	5
Indonesia	5
Peru	5
USA	5
Brazil	4
Ethiopia	4
Cameroon	4
Nigeria	4
Malaysia	3
Mexico	2
Canada	2
Taiwan	2
Solomon Islands	2
Other location	16
Total	100

In the category of "other location" in Table 3 are grammars of languages from Tanzania, Swaziland, Sierra Leone, Democratic Republic of Congo, Mali, Benin, Venezuela, Colombia, Bolivia, Pakistan, Bhutan, New Caledonia, East Timor, Vietnam, Turkey, and Latvia.

We attempted to ensure a spread of dates within the ten-year range for both published grammars and dissertations. However, we also attempted to ensure a spread for other factors such as language area, publisher, and institution, which influenced the distribution based on date. The year of publication for dissertations and grammars is shown in Figure 1.[2] Dissertations were easier to locate in later years as online repositories have become more common. The full list of publications we used in our survey are given as appendices to this paper. Appendix A gives a bibliographic list of published grammars, by author. Appendix B gives a bibliographic list of dissertations, by author. We do not provide the coding results from our analysis in this bibliography, because we wish to reflect on the state of the discipline as a whole, and not single out individuals. In the next section, while outlining the features that we coded for, we do take the opportunity to highlight and celebrate positive examples of researchers who make data and methods transparency a feature of their published research.

[2] All graphs in this paper were created using R (R Core Team 2013).

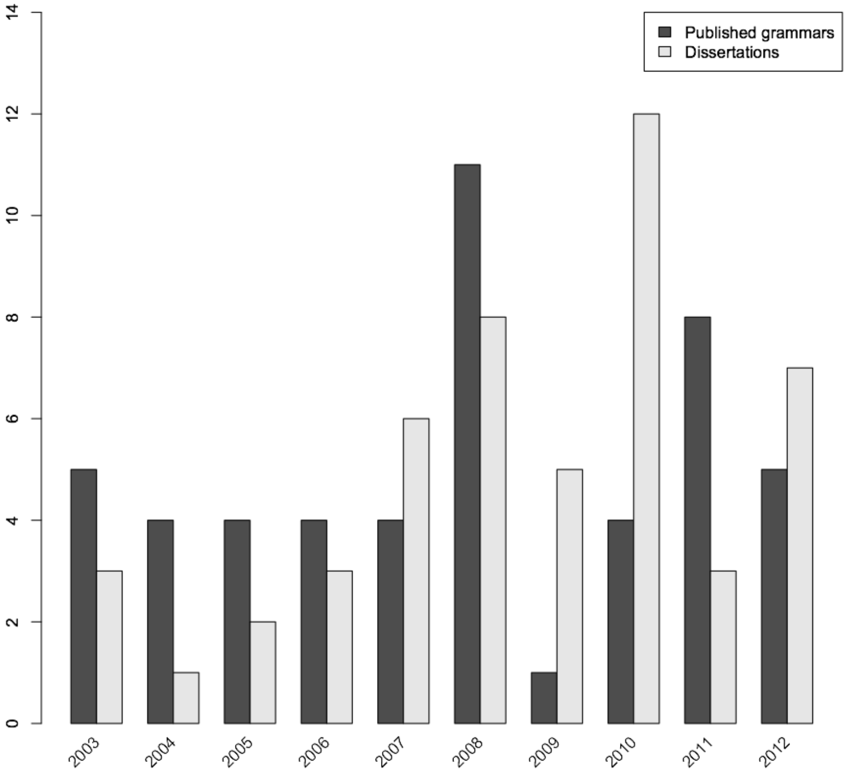

Figure 1. Year of publication for published grammars and dissertations

3.2 Variables Grammars were coded for a number of variables, which we describe below, giving examples where these are useful. The variables include description of data collection methods, information about participants, data collection equipment, data collection tools, speech genres collected, data analysis software, time collecting data, data archiving, sources of data, and where the data are now. We also looked at the citation conventions that were used in numbered examples. These categories were decided upon based on our experience of the existing literature on language documentation, as discussed in §2 above. The variable set was tested with an initial subset of the survey sample and refined into categories below. As we discuss in §5, we do not consider this to be an exhaustive set of features to include in a robust methodology or data citation standard, but we do consider them to be primary features, and easy for researchers to address.

3.2.1 Description of data collection methods (1–5) This variable tracks whether the author explicitly describes the methods used in data collection, and how detailed their description of methodology is. As there is such a broad range of how much detail authors include in their methodology in a monograph-length publication, we used a five-point Likert scale as a qualitative rating of methodological descriptions in grammars.[3] The five points of the Likert scale are given with their defined expectations in Table 4.

Table 4. Likert scale for scope of methodological description

1	No discussion of methodology
2	Methodology mentioned in passing, often in the introduction
3	Methodology as a small, but distinct, subsection of another chapter(s)
4	Methodology as a major subsection, with several pages of detail
5	Methodology as its own discrete chapter, or a significant section of a larger chapter

Grammars varied greatly in the amount of methodological information given, some offering cursory information, and others providing rich descriptions of methodological practice. Notable published grammars with richly described methodology include van der Voort (2004), Coupe (2007), and Thieberger (2006). Notable dissertations with richly described methodology include Mihas (2010), Vallejos Yopan (2010), and Hyslop (2011). Note that length of the methodological section alone is not a good indicator of quality. Vallejos Yopan's dissertation contains a stand-alone methodology chapter of 34 pages, and is certainly outstanding, while Brotchie (2009) provides but a few pages, which are nevertheless targeted and useful to the reader.

Methodological information is often found in the introduction chapter, or as a subsection of a second chapter that focuses on ethnographic information and background literature. In a number of cases the authors provide information regarding time spent collecting data or participant description in the acknowledgments, but do not include this information in any formal methodology section. While such information may provide insights regarding methodological practices it was not counted in in our survey. Unlike journals, monographs do not have the same space constraints for discussing methodology. As such, we decided it was reasonable to expect that methodological information should be found in the body of the work, rather than in the front matter. Consistent and clear presentation of methodology in an early section of a grammatical description would assist in making methodological processes more transparent and comparable.

3.2.2 Information about participants in the study (YES/NO) Although this variable was coded on a binary scale, there is considerable variation in the amount of detail that researchers give about participants. We set the bar relatively low here, so as long as the researcher told us the name (or alias) and gender of participants, this variable was coded as a YES. It is possible that researchers only mentioned participants who were central to the documentation process, while also including data from a wider range of speakers. Ideally participant ages, places of residence, and other

[3] A Likert scale is a graded set of points that are used to rate qualitative option data. This allowed us to gain an idea of the variation in methodologies without focusing specifically on individual features.

languages spoken would also be included. Aikhenvald (2003), for example, gives detailed, paragraph-long biographies of all of her participants and their families.

We coded this variable as YES if the information appeared in the body of the grammar, or an appendix. We did not code as YES if information about participants appeared only in footnotes or acknowledgments. This was for two reasons. First, monographs do not have the kind of space constraints that journal articles face, so this information could easily be presented in the body of the text. The second reason is that this survey is about ascertaining to what extent our discipline is attempting to make methodologies transparent. Having to turn to the acknowledgments for participant information does not, in our opinion, meet this benchmark. We do, however, encourage researchers to include basic participant information for key participants in the methodology section of a grammar, and include information on all participants in an appendix.

3.2.3 Mention of data collection equipment (YES/NO) Data collection equipment refers to the physical equipment or hardware used to record language, and other equipment for general documentation. This includes, but is not limited to, video cameras, audio recorders, microphones, still cameras, and GPS units. We coded data collection tools as a separate variable, discussed in the next section. We coded as YES grammars whose authors simply mention if their work was done with, for example, a cassette recorder or video camera, but some researchers even provide model numbers and details of equipment setup. Coupe (2007) gives particularly detailed information:

> The Khar, Khensa and Waromung data were recorded in the field using a Sony TCM 500EV analogue cassette tape recorder and a Sony ECM-FO1 microphone [...] Most of the Mangmetong data was recorded using a Sony MiniDisc MZ-R50 and a Sony ECM-ZS90 condenser microphone. (Coupe 2007:20)

Mihas (2010) discusses the equipment used through the whole documentation workflow:[4]

> Our meetings were recorded using a Sony PCM D-50 Linear PCM digital recorder, a Panasonic HDC-HC100P/PC digital camcorder and Sony Electret Condenser Stereo Microphone ECM-MS908C, with the elicited audio and video data transferred to the laptops Dell B130, Dell Mini, Dell Latitude E6410, and WD Western Digital 'My Passport Essential' portable hard drive. (Mihas 2010:31)

3.2.4 Mention of data collection tools (YES/NO) Data collection tools in the grammars were taken to be any stimulus or experimental data collection tools that linguists

[4] While we do not believe it is necessary to mention models of storage equipment as part of the language documentation workflow, striving for more information rather than less is certainly preferable. There is also a utility in doing so that is not immediately tied to the reproducibility of research results, in that it normalizes good practice in data storage, and provides evidence of best practice to support researchers writing research grant budgets.

regularly deploy during fieldwork. Citing the work that colleagues put into creating and sharing tools helps build recognition of those resources. It also allows readers to see if there are data that may be useful in comparability studies, particularly in cases of commonly used resources like Swadesh lists (Swadesh 1971), *Frog, where are you?* (Mayer 1969), or the MPI *Field Manual* tools.[5] Finally, it can inspire colleagues to try new methods in their own research. Baird (2008) provides a good example of listing the types of tools used in her work on Klon.

> I used several well-known props to elicit stories and utterances. One speaker provided the storyline to accompany Mercer Mayer's children's book *Frog, where are you?*, which consists solely of pictures. Another speaker provided narration to the *Pear story* video clip, which doesn't contain any speech. Three elicitation tools produced by linguists at the Max Planck Institute were used as prompts to elicit specific linguistic phenomena. There were: short video clips depicting people performing everyday activities; animated video clips, known as *Motion land*; and photos, which were used in a photo matching game, where the participants matched photos by describing them to one another, without being able to see the photo being described or each other. (Baird 2008:6)

Gaby (2006:15) mentions in the introductory section on methodology that "[p]rompted natural speech arose from presenting consultants with a visual (usually video) stimulus to describe (cf. Appendix 2)". Her Appendix 2 gives a list of tasks, the name they are given in the dissertation discussion, and a description and citation of who created them. These are three of the sixteen tasks described:

> *AbsoluteTime* Data collected during pilot trials of a stimulus designed by the author and Lera Boroditsky. The task comprises a set of cards depicting temporally linked objects or events. Consultants are asked to place the cards in sequence.
>
> *CausedPositions* Set of video stimuli depicting caused topological relationships, published as Hellwig and Lüpke 2001.
>
> *DahlTMA* Questionnaire relating to tense, mood, and aspect, included as an appendix to Dahl 1985. (Gaby 2006:638)

Of course, not all researchers will draw upon existing tools for their research. Moreover, some tools may be trialed and precluded from use if they do not produce useful materials. Mentioning failed tools can help other researchers consider possible pitfalls when doing their own research.

3.2.5 Speech genre collected (YES/NO) Descriptive grammars often draw on a range of speech genres, but rather than attempting to build an exhaustive list here, we coded for whether researchers make overt mention of the genres that they had collected in their field work. Kratochvíl (2007:21–23) gives a detailed list of materials in

[5] http://fieldmanuals.mpi.nl/

the corpus he used as the basis of the descriptive grammar, making note of different genres including fables, traditional stories, meetings, conversations, and elicitation tasks. Hellwig's (2011:8) overview of the Goemai corpus notes that it includes "conversations, different kinds of narratives, descriptive texts, procedural texts, speeches, riddles, proverbs, and songs."

3.2.6 Mention of data analysis tools or software (YES/NO) We looked for mention of tools used in the processing, analysis, and presentation of the language materials in the corpus. Grammars in the survey were coded as YES even if the tools are only mentioned by name, and not fully cited. Lidz (2010) presents her discussion of Praat (Boersma & Weenink 2015) in the introductory chapter:

> The phonetic analysis software used was Praat, which is available by free download from: http://www.fon.hum.uva.nl/praat/. Praat version 4.3.19 was used on a laptop operating under Windows XP, with view range set to 0–5000 Hz, a bandwidth of 260 Hz, a window length of 0.005, and a dynamic range of 40dB. (Lidz 2010:12)

3.2.7 Mention of time spent collecting data (YES/NO) This variable tracks whether authors mention how much time they spent doing fieldwork. Even those authors who work with existing documentation need to make the fact they did not participate in any fieldwork/data collection a feature of their methodology. Some authors describe the length of their fieldwork in terms of how long they were actually in the field, while others give the span of an entire project that includes time spent in the field and time spent elsewhere. De Sousa (2006:61) notes that for his dissertation "[f]ive fieldtrips were conducted between August 2002 and November 2005; the time spent in the field amounts to fourteen months accumulatively". Sava (2005:4) provides a detailed itinerary of fieldwork with speakers of Ts'amakko, noting that "fieldwork was carried out during four periods: in June-July 1999, August 2000, March-August 2001, and April-July 2003". Again, this information occasionally was mentioned in the acknowledgments, but this was precluded from the analysis.

3.2.8 Whether data have been archived (YES/NO) For this variable, we coded as YES instances in which an author noted whether the materials have been archived, and where. By archiving we specifically mean a repository that has strategies for long-term preservation of materials and allows for future access. Salffner (2010) discusses where the materials are archived, and also their accessibility.[6]

> All data has been archived with the Endangered Languages Archive of the Hans Rausing Endangered Languages Project at the School of Oriental and African Studies, University of London and will be openly accessible in due time. (Salffner 2010:41)

[6] True to this, the Ikaan corpus is almost totally open-access today http://elar.soas.ac.uk/deposit/0034 visited 13/09/2015

Klamer (2010:36) discusses the content of the archived material relating to her grammar:

> The transcribed texts and elicited materials are archived as Toolbox files. These files contain the recorded text materials in an annotated form, with glosses and translations in Indonesian and English. The digital recordings are stored in the form of mini-DV tapes, with copies in MPEG files. Electronic copies of the recordings will be deposited at the IMDI archive that is maintained at the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands. (Klamer 2010:36)

3.2.9 Source of data While most grammars are based on the author's fieldwork, some authors do use other data sources supplementarily. We coded for these based on four broad possible data sources:

- Own fieldwork
- Other published sources
- Other unpublished sources
- No mention of source

Individual grammars could be coded for more than one category, for example if the description is based on both original fieldwork and unpublished archive notes from a previous researcher.

3.2.10 Where the data are now We believe that archiving data is an important part of language documentation work. However, it is not the only thing that can be done with data from a documentation project. Data can be returned to specific community members, or held by a community language group. This variable tracks where data are located, although very few of the grammars that were coded NO for the archiving variable made it clear if the data were located anywhere else (such as with speakers on storage devices that do not constitute a long term archive). Below we list the different types of places that researchers mentioned their data were archived. Some grammars were coded with more than one of these options.

- ARCH: archived in an institutional repository
- WILLARCH: author states they will archive corpus in near future
- ONL: online (a website or other non-archive internet-based storage)
- COMM: data held by language community
- PUBD CORP: sizable text corpus with grammar, or in separate volume
- UNST: not stated

3.2.11 Citation conventions used in numbered examples

Descriptive grammars of any substance are replete with numbered examples of language use, drawn primarily from recordings and fieldnotes usually collected or created by the author, and supplemented by other data types. We used a five-point Likert scale, shown in Table 5, to grade how clearly authors cite data sources and whether those citations are resolvable to the dataset.

Table 5. Likert sale for grading citation format

1	No citation convention used
2	Minimal ad hoc reference to speaker, or title of text
3	Minimal ad hoc reference to speaker and title of text
4	Resolvable to corpus, but no indication of corpus location
5	Fully resolvable to corpus, with time-codes, explanation of convention, and fully archived corpus.

McCracken (2012) provides a detailed explanation of her citation structure:

> Each example drawn from natural discourse is marked with a unique marker indicating the text it was drawn from, the speaker, and a time-code so that the reader may listen to the example (see Appendix B for these codes and a description of each text). Examples are also drawn from elicited translations, grammaticality judgments, and wordlists; these are used as little as possible, and are marked with the date and the speaker. Speakers are identified by their initials throughout the text (see Appendix C). (McCracken 2012:46)

We can see this citation structure in use in an example from her discussion of verb structures (McCracken 2012:305):

(1) « Ka ivi waga-ji ? »
ka ivi waga-ji
LK be.where.SPC boat-1DU.INCL.POSS
'And where is our boat?' (Yal-01082010-MFD_0016)

In McCracken's Appendix B, the recording is described as "TEAMBOUEON Marie-France tells a legend about the octopus and the rat" (McCracken 2012:571), and in her Appendix C, MFD is again named as "TEAMBOUEON Marie-France" (McCracken 2012:573). Although McCracken refers to time-codes in the quote above, in her Appendix B she makes clear that these are actually references to the Toolbox line numbers, so this example would be from the 16th line of interlinearized text from the story (which would also have a time association in the Toolbox file).

Mushin (2012) gives this explanation of her citation conventions:

> I reference examples from my own text corpus by the date of recording, the text number from that particular day (usually a number between 1

> and 4), and the speaker's initials. So the reference "1.9.01.2.DG" would be from the second recorded text on September 1, 2001 that was spoken by Doreen George. Data that come from elicitation sessions are referenced as "Mushin (year) field notes". (Mushin 2012:12)

Mushin also gives a clear explanation of how she cites earlier work on Garrwa, and notes that her materials are archived at the Australian Institute of Aboriginal and Torres Strait Islander Studies.

4. Results In this section we discuss the results of our survey for each of the categories coded for in §3.2.

4.1 Description of data collection methods As described in §3.2, grammars were coded for their overall attention to methodological features on a five-point Likert scale, with 1 indicating the absence of any discussion of methodology, and 5 indicating an entire section or chapter devoted to a description of the methods used in data collection. The median of both data sets was 2, indicating that while there are some exemplary grammars, as a field we are not providing a great deal of information regarding methodology.

To provide an objective test of the efficacy of our Likert rating, we also counted the frequencies of individual methodological feature descriptions. These are summarized in Figure 2, with the features listed along the X axis, and the number of published grammars and dissertations they appear in on the Y axis.
Overall, authors of dissertations outperformed authors of published grammars in their discussions of methodology for every variable. We see this as a healthy sign for the future of grammar writing, and discuss the implications in §5.

Figure 2 also serves as a summary of results for the next seven features that were coded. We briefly give the numerical results for each of these categories, noting trends observed in the grammars surveyed.

4.2 Information about participants in the study Participants are described in 31 dissertations and 24 published grammars.

4.3 Mention of data collection equipment Data collection equipment is mentioned in 22 dissertations and in only nine published grammars. These are mainly mentions of specific models of audio recorder, video camera, and/or microphones used in data collection.

4.4 Mention of data collection tools Data collection tools are mentioned in 27 dissertations and in nine published grammars; these are usually mentions of an elicita-

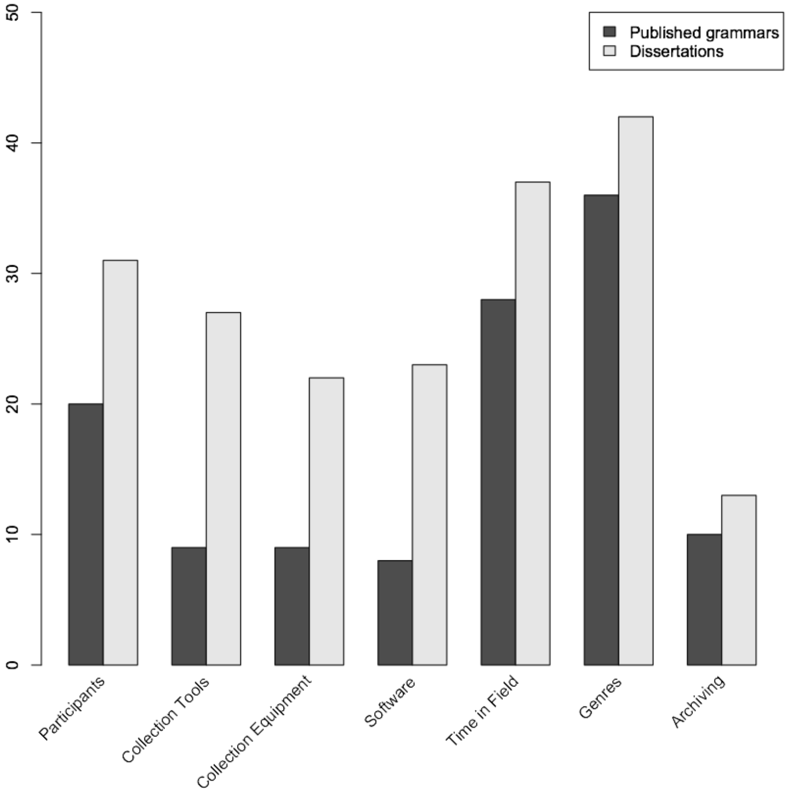

Figure 2. Methodological features mentioned

tion stimulus.[7] Only five published grammars mention both data collection equipment and tools.

4.5 Speech genre collected Speech genre is mentioned in 42 dissertations and in 36 published grammars. This was the most common feature of the language documentation process mentioned by authors in each category.

4.6 Mention of data analysis tools or software 23 dissertation authors mention analysis tools, as do eight published grammar authors. The majority of references are to Praat as well as software relevant to language-documentation workflows, such as Toolbox (SIL International 2013) and ELAN[8] (Wittenburg et al. 2006).

[7] It is possible that we missed some mentions of collection tools, as oblique references to stimulus materials such as *Frog, where are you?* (Mayer 1969) may only be included at points in the text where examples from these stimuli are discussed.

[8] http://tla.mpi.nl/tools/tla-tools/elan/ Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands

4.7 Mention of time spent collecting data In grammars the time period referenced invariably relates to the time that the author spent at the field site. This is mentioned in 37 dissertations and in 28 published grammars.

4.8 Whether data have been archived Data archiving is mentioned in 14 dissertations and in 12 published grammars. This is almost always a reference to the author's own data archive, or the author's plans to archive. More information on authors who mention plans to archive is given in the section "where the data are now," below.

4.9 Source of data While most grammars are based on the author's fieldwork, some authors do use other data sources. These are summarized in Table 6. Totals exceed 50 because some authors drew on both their own data and other data types. Still others made no mention of data source(s) at all (and no mention of data collection); although it may be presumed that they were working with their own data, we coded them as "no mention of source."

Table 6. Source of data

	Dissertations	Published Grammars
Own fieldwork	50	40
Other published sources	6	11
Other unpublished sources	2	5
No mention of source	0	7

It is perhaps unsurprising that dissertation authors are more likely to make their data source clear, and that they draw on their own data collection. This is because dissertation grammars are about illustrating the author's competency in documentation and description.

4.10 Where the data are now A summary of data locations in grammars is given in Table 7. Totals exceed 50 for both dissertations and published grammars, as some authors mention multiple options (e.g., that the data were archived, and a copy was also left with the community). One published grammar mentioned that some data were already archived and some would be archived in the future, and is therefore listed twice here.

In the category "will archive", authors indicate that they have plans to archive the data, and occasionally mention the target repository. Of the two dissertations listed as "will archive", no digitally discoverable archive of the language could be found when this paper was written, four years after the survey end-point.[9] Of the three published grammars where the author stated they would archive the data, two

[9] As of January 2017, they do not appear in Open Language Archives Community (OLAC) search, which includes 60 archives. The researchers' websites were also checked, and general searches also made. It is possible that the materials are archived with offline repositories, which raises future problems regarding their accessibility.

Table 7. Described location of data in grammars

	Dissertations	Published Grammars
Unknown	35	33
Archived	12	10
“Will archive”	2	3
With community	6	2
Online	0	4
Sizable text corpus with grammar	1	5

have made archived materials available. The Teiwa materials for Klamer (2010) are archived with The Language Archive,[10] and the Urarina materials for Olawsky (2006) are archived with the Endangered Language Archive.[11] The third author stated that some narratives had already been archived, and the remainder would be deposited, but the collection still only consists of two narratives. Although we hope that these researchers eventually archive their materials, we still see this as a positive step; authors know archiving is an important part of language documentation work, and that the expectation of archiving is becoming normalized within this field.

4.11 Citation conventions used in numbered examples A five-point Likert scale was used to group publications by the richness of their citation style. This scale is described above, with 1 indicating no citation of data sources and 5 indicating citations fully resolvable to an archived or retrievable data source. The distribution of publications is given in Figure 3.

The vast majority of grammars do not provide citations for examples, making it impossible to verify or confirm examples. Somewhat encouraging is that dissertations out-perform published grammars with regard to citation to resolvable sources, whether they are archived (rated 5 on our scale) or not (rated 4 on our scale).

5. Discussion and recommendations Our survey demonstrates what we impressionistically thought to be true: that while there is a lot of good *practice* in language documentation in terms of proper care for methods and data, good practice is often not made explicit in the subsequent write-up. The result is that descriptive and documentary linguistics may have a growing body of literature on how to do research, but this also needs to carry across to ensuring that all researchers make their research methodology clearer. In this section we discuss the implications of the results of the survey for our field. In no way do we place the onus for greater transparency on the grammar authors who did not meet our retroactively-applied criteria; instead we see this as an opportunity for discipline-wide discussion on how to increase reproducibility in grammar writing. We are optimistic that as a discipline we can effect a positive

[10] https://hdl.handle.net/1839/00–0000-0000–001E-2C34–6@view visited 22nd February 2017.

[11] https://elar.soas.ac.uk/Collection/MPI85732 visited 22nd February 2017.

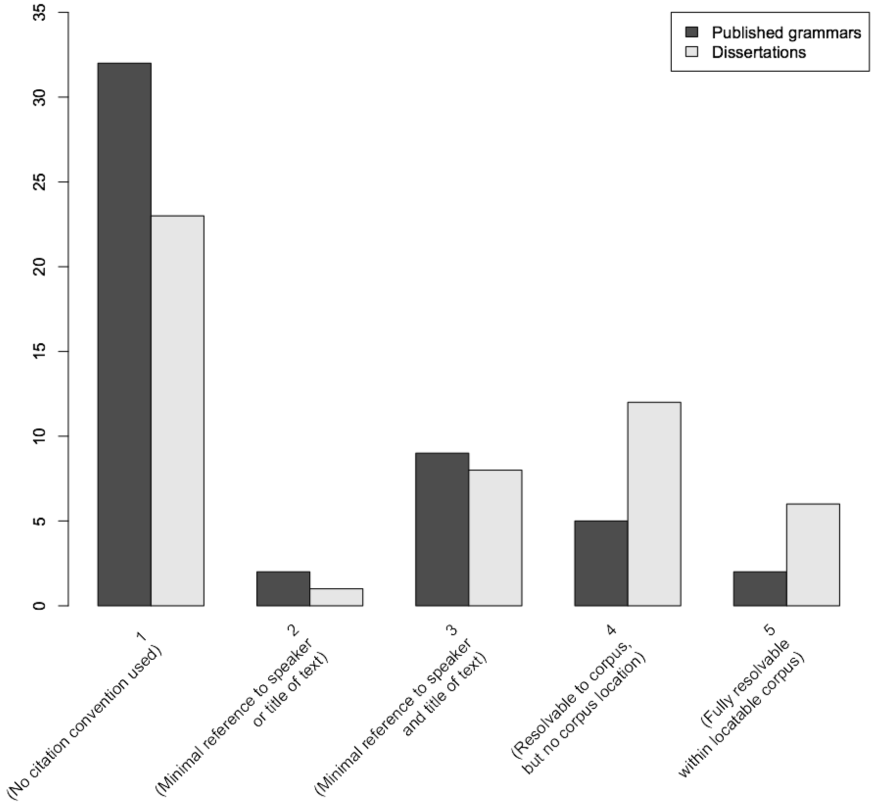

Figure 3. Citation conventions

change in the attitude that we have towards the genre of written grammars, and can aspire to a higher standard of reproducibility.

Our optimism stems from our findings that the dissertations outperformed the published grammars in every metric we assessed. We have considered two possible reasons for this. First, published grammars, although published in the same ten-year span as the dissertations we surveyed, are often based on research that started much earlier than the date of publication, usually five to ten years or more. Dissertations, however, are usually completed within a five year span, and thus dissertation authors are able to employ more current methods and expectations throughout the writing process. Therefore, we could expect that published books may lag behind dissertations in terms of expressing methodological innovations.

The second reason may be that descriptive linguists have come to treat the dissertation grammar and the publication grammar as different genres. Dissertations may contain more overt methodological description because their function is to demonstrate that the author has mastered the discipline; authors who write grammars some time after finishing the PhD may feel less pressure to demonstrate this kind of mastery.

We did observe cases, however, in which a single author developed the dissertation grammar into a published grammar; while these are not included in the study it is our impression that most of these works retain much of their original methodological content. We see it as a positive sign that newly-minted descriptive linguists are showing a general trend towards greater transparency, and we hope that they maintain this practice in their future work as well.

The benefits of increased transparency of methods and data are clear. High quality grammars will serve not just as models of good linguistic description, but also as models of good methodology, allowing other researchers to more efficiently adapt methods to their own situations. Additionally, researchers can cite successful, published methods as precedent in grant applications and IRB/ethics protocols.

Ensuring that example sentences in descriptive grammars have transparent citations linking to primary data reasserts the primacy of the documentary materials. Although it is not possible to replicate the exact context of an utterance in conversation in the same way that it is possible to replicate an experiment in physics, and to a lesser extent psychology, we can strive to allow others to reproduce our analysis by providing access to the original recordings. Thus further research may grow out of the same data used to create a descriptive grammar, including reconsidering examples based on factors not considered in the original analysis. This is not only helpful for the progress of linguistic science, but can be of great assistance to the original authors in deepening their own understanding of the language, as even we ourselves can happily attest.

While we hope that this article has inspired those researchers currently writing a descriptive grammar to consider the way they have written their methodology and cited examples, we believe that we also need to approach a change in the culture of the discipline. We therefore bring this discussion section to a close with a few suggestions that we think can, with only minor changes, make major differences in the way the writing of descriptive grammars is approached. These are: including more information in grammars; better training our students; and fostering a broader cultural shift in descriptive linguistics.

5.1 Including more information In order to make grammars more transparent, we make two recommendations. The first is that descriptive grammars should include an easily-locatable chapter or major section of a chapter towards the front of the volume outlining the methodologies employed in data collection, management, and analysis. We suggest that the parameters we coded for serve as a bare minimum of information. For every grammar it is expected we should know basic information about the participants (including the researcher), the equipment and tools used to collect the data, the tools used to analyze it, the time spent collecting data, the data genres collected, whether the data has been archived, and where else it is stored. The methodological variables we included in our study are, in our recommendation, a minimal set of information to include in descriptive grammars, but the list is by no means exhaustive. There are other methodological features that we are sure that other researchers would be more interested in, and we hope that this survey can be the start of an ongoing discussion about what constitutes a clear, useful, and replicable methodology. These could include information about the IRB/ethics protocols under which data has been collected, whether written or oral consent is documented, and whether the database extends across different members of the speech community, including across age and gender. We would also like to see more overt discussion of the broader linguistic context in which fieldwork was conducted, e.g., what the contact language was, whether it was a language of common use for either the researcher or the community members, and how extensively it was used in elicitation.

The second is that for every example of phrasal length or longer, the author should provide a citation that will allow the reader to not only locate the larger data set of recordings and/or fieldnotes in an archive, but also to resolve back to the particular datum within the set. There are a range of possible formats for citation, and the author should take care to clearly explain the citation format in the front matter of the volume. In reading these grammars, the best examples of citation involve a number of parameters. The first is that the citation is not too obtrusive on the written page, with just a short line of text coming after the translation of the example sentence. The second is that the code is clearly explained in the front of the grammar, and used consistently. The third is that it is easily resolvable within the collection where materials are archived, that is it is consistent with the file-naming conventions already in use. The fourth is that the citation has a sufficient level of granularity that the reader can find the example at the point in the recording or in the line of transcription where it occurs.

An expectation of explication of methods and data is not limited to monograph-length grammatical descriptions, but should also extend to other works as well. In a shorter publication the need for transparent data citation is even more pressing, as there is less space for multiple examples, although space limits in journal articles may require brevity on the topic of methodology, sometimes constrained to even a footnote or passing reference. We are extending our survey beyond grammars to journals across a range of theoretical and areal specializations (Berez-Kroeker et al., in prep). Within the broader field of linguistics, descriptive linguists are in a unique position to examine and reflect on methodological transparency in published work. As a sub-field, we have two decades of discussion around methodological best practice which enables us to be leaders in demonstrating that knowledge in written outputs, and to be leaders in data transparency within linguistics and the humanities more broadly.

5.2 Educating our students As educators, we need to make these expectations more overt when training graduate students. As Pawley (2014) notes, graduate training is the perfect opportunity to introduce people to best practice in the field. Classes in data management in linguistics are few and far between (Dailey & Henke 2016, 2017), but could be added as a regular part of the graduate curriculum. In 2013 the Department of Linguistics at University of Hawai‘i at Mānoa introduced changes aimed at improving data transparency to their PhD requirements (Berez 2015). Students whose dissertations are based on the collection of original data are now required to include a data archiving plan in their dissertation proposal, and to provide proof of archiving before the final dissertation can be approved. Additionally, examples in dissertations must include citations to resolvable resources. These requirements normalize the expectation that a descriptive grammar is based on an archived collection of data. In our experience, students have been quick to realize the value of reproducibility and to accept the workflow it entails.

5.3 Fostering a culture shift Beyond the graduate experience, there are a number of mechanisms we can use to encourage better data and methods transparency. The first is to encourage journals and publishers to adopt a policy encouraging the citation of data, and clear methodologies. Secondly, we can draw upon existing standardizing mechanisms within linguistics publishing to normalize these expectations. For example, the Unified Style Sheet for Linguistics,[12] the Generic Style Rules for Linguistics,[13] or even the Leipzig Glossing Rules[14] could include styling models for data citation (see Berez-Kroeker et al. 2017 for discussion on the need for citation formats). As individuals we can also have a positive effect in this area during the peer-review process by requesting more transparent methodological description. Even though it may not be feasible to request that all examples in a completed work be given citations to corpora, we can suggest it be included in future work.

There are also a number of other bodies and organizations that can assist in promoting a change in expectations. By including an expectation of transparent methodology and the citation of examples to resolvable archives, linguistics awards can normalize the expectation of good practice, and help highlight great examples of work already being done. We see this as a positive contribution that could be made by the Association for Linguistic Typology in their awards for grammatical description work, including the Georg von der Gabelentz Award[15] for best published grammar, The Pāṇini Award[16] for best grammar dissertation and the Joseph Greenberg Award[17] for best typological dissertation. Similarly, the Linguistic Society of America awards,[18] including the Leonard Bloomfield Book Award and Best Paper in Language Award, could include consideration of methodology and data citation. DELAMAN has introduced the Franz Boas Award[19] for an outstanding documentary work and the development of an archived corpus by an early career researcher, which is intended to help encourage archiving, the foundation of good practice in data citation. The National Science Foundation-funded project *Developing Standards for Data Citation*

[12] http://www.linguisticsociety.org/resource/unified-style-sheet

[13] http://www.eva.mpg.de/linguistics/past-research-resources/resources/generic-style-rules.html

[14] https://www.eva.mpg.de/lingua/resources/glossing-rules.php

[15] http://www.linguistic-typology.org/awards.html#Gabelentz

[16] http://www.linguistic-typology.org/awards.html#Panini

[17] http://www.linguistic-typology.org/awards.html

[18] http://www.linguisticsociety.org/about/who-we-are/lsa-awards

[19] http://www.delaman.org/delaman-franz-boas-award/

and Attribution for Reproducible Research in Linguistics has been working to build consensus about these issues across the discipline more broadly.[20]

6. Conclusion Language documentation methodology is increasingly centered around the development of annotated and archived digital corpora. Researchers have shown a general willingness to embrace new tools and methods in both data collection and analysis. We acknowledge that descriptive linguists often *practice* good methodology in data collection. Our survey has illustrated that we still need to improve how we *write* about the methods that we employ.

By and large we have encountered positive reactions when discussing the need for increased transparency with our colleagues. Nonetheless, we understand that some people may be reluctant to lay bare their fieldwork process. During earlier presentations of this work (Berez 2015, Gawne et al. 2015, Berez-Kroeker 2016), we had discussions with researchers who worried that being required to make their methodology more transparent places them in a position of vulnerability, open to criticism for not spending enough time in the field, not working with enough people, etc. We do not wish to dismiss these concerns, and as fieldworking linguists who have all undertaken documentation work in a variety of research contexts ourselves, we are aware of the uniquely personal realities of this kind of data collection. However, we also believe that in order for language documentation and description to advance as a field, we need to also be more upfront about the fact that descriptive grammars are not objective records of language, but subjective analyses of corpora by individuals (see Thieberger 2014). Greater transparency does not mean diminished respect for each other and the contexts that we work in.

Another concern is that the additional work of transparency requires more from us in terms of time and resources. That is certainly true, but we would argue that the rewards to linguistic science justify the cost to any one researcher, and that like Thieberger & Berez's (2012:91) metaphor of the pains of data management as "the firm foundation on which a house is built", extra effort in the near term pays off in the long term. In reality, the time cost of writing about one's methodology is a fraction of the time spent writing a descriptive grammar. Most of this methodology can even be pre-written and included as metadata in the archived collection, or in a corpus article in the model of Salffner (2015). Data citation is time intensive only when it is post hoc; if tracking the location of examples in the larger dataset has been part of the linguist's workflow from the earliest stages of data analysis, citation is a trivial matter, no more difficult than citing a quote from a book. It is for this reason that we expect to see an evolving, rather than immediate, change in the state of data citation, and why we also think it is important to set clear expectations of transparency, so that scholars beginning their careers or new projects can set up their workflows accordingly.

Language documentation can lead linguistics in this respect. We already have expectations and practices for digital data management that are only now starting to be

[20] https://sites.google.com/a/hawaii.edu/data-citation/welcome; see also Berez-Kroeker, Holton, Kung & Pulsifer (2017).

normalized in other areas of linguistics and the humanities more broadly. Thanks to initiatives like the Open Language Archive Community[21] and the Digital Endangered Languages and Music Archiving Network[22] we already have structures in place to make wide-scale data citation and methodological transparency possible. It is time that we shared the good practice that we are doing in our documentation in the descriptive grammars that we write as a product of that work.

References

Aikhenvald, Alexandra Y. 2003. *A grammar of Tariana, from northwest Amazonia*. New York: Cambridge University Press.

Aikhenvald, Alexandra Y. 2014. *The art of grammar: A practical guide*. Oxford: Oxford University Press.

Ameka, Felix K., Alan Charles Dench & Nicholas Evans. 2006. *Catching language: the standing challenge of grammar writing*. Berlin: Mouton de Gruyter.

Austin, Peter K. 2013. Language documentation and meta-documentation. In Jones, Mari C. & Sarah Ogilvie (eds.), *Keeping languages alive: Documentation, pedagogy and revitalization*, 3–15. Cambridge: Cambridge University Press.

Baird, Louise. 2008. *A Grammar of Klon*. Canberra: Pacific Linguistics.

Berez, Andrea. 2015. Reproducible research in descriptive linguistics: Integrating archiving and citation into the postgraduate curriculum at the University of Hawai'i at Mānoa. In Harris, Amanda, Nick Thieberger & Linda Barwick (eds.), *Research, records, and responsibility: Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures*, 39–52. Sydney: University of Sydney Press.

Berez-Kroeker, Andrea L. 2016. Reproducible research and the Americanist tradition in linguistics. Keynote address presented at the 19th Workshop on American Indigenous Languages, Santa Barbara, CA, May 7–8 2016.

Berez-Kroeker, Andrea L., Gary Holton, Susan Smythe Kung, Geoff Nathan, Peter L. Pulsifer, Anthony Woodbury, Keren Rice, Stanley Dubinsky & David Beaver, Shobhana Chelliah, Ruth Duerr, Richard Meier & Nick Thieberger. 2017. Symposium & panel discussion: Data citation and attribution for reproducible research in linguistics. Paper presented at the Linguistic Society of America Annual Meeting, Austin, January 5 2017. https://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/43564/.

Berez-Kroeker, Andrea L., Gary Holton, Susan Smythe Kung & Peter Pulsifer. 2017. Developing standards for data citation and attribution for reproducible research in linguistics: Project summary and next steps. Poster presented at the Annual Meeting of the Linguistic Society of America Annual Meeting, Austin, January 6 2017. https://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/43565.

Berez-Kroeker, Andrea, Lauren Gawne, Barbara F. Kelly, Tyler Heston, Anthony Woodbury, Keren Rice, Stan Dubinsky, Shobhana Chelliah, Gary Holton, Peter Pulsifer, Susan Kung & Nick Thieberger. In prep. Transparency of data and methods in linguistics: Position statement on data citation and attribution.

[21] http://www.language-archives.org/

[22] http://www.delaman.org/

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79. 557–582. Boersma, Paul & David Weenink. 2015. Praat: Doing phonetics by computer (computer program). http://www.praat.org/.

Bowern, Claire. 2008. *Linguistic fieldwork: A practical guide*. New York: Palgrave Macmillan.

Brotchie, Amanda. 2009. *Tirax grammar and narrative: An Oceanic language spoken on Malakula, north central Vanuatu*. Melbourne: University of Melbourne. Doctoral dissertation.

Buckheit, Jonathan B. & David L. Donoho. 1995. WaveLab and reproducible research. In Antoniadis, Anestis & Georges Oppenheim (eds.), *Wavelets and statistics*, 55–81. New York: Springer.

Chelliah, Shobhana L. & Willem J. de Reuse. 2011. *Handbook of descriptive linguistic fieldwork*. London: Springer.

Coupe, Alec. 2007. *A grammar of Mongsen Ao*. Berlin, New York: Mouton de Gruyter.

Crowley, Terrence. 2007. *Field linguistics: a beginner's guide*. Oxford: Oxford University Press.

Dailey, Meagan & Ryan Henke. 2016. Data management in the job market and graduate education. Paper presented at the 2nd Workshop on Data Citation and Attribution in Linguistics, Austin, April 8–10 2016. https://scholarspace.manoa.hawaii.edu/handle/10125/40024.

Dailey, Meagan, & Ryan Henke. 2017. Data citation, attribution, and employability. Poster presented at the Linguistic Society of America Annual Meeting, Austin, January 6 2017. https://scholarspace.manoa.hawaii.edu/handle/10125/43571.

Dixon, Robert M.W. 2009. *Basic Linguistic Theory*, vol. 1. Oxford: Oxford University Press.

Gaby, Alice R. 2006. *A grammar of Kuuk Thaayorre*. Melbourne: University of Melbourne. Doctoral dissertation.

Gawne, Lauren, Barbara F. Kelly, Andrea Berez & Tyler Heston. 2015. Putting practice into words: Fieldwork methodology in grammatical descriptions. Paper presented at the Fourth International Conference on Language Documentation & Conservation, Honolulu, February 26– March 1 2015.

Gezelter, Dan. 2009. Being scientific: Falsifiability, verifiability, empirical tests, and reproducibility. *The OpenScience project*. http://www.openscience.org/blog/?p=312.

Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (eds.). 2006. *Essentials of language documentation*. Berlin: Mouton de Gruyter.

Haspelmath, Martin & Susanne Maria Michaelis. 2014. Annotated corpora of small languages as refereed publications: A vision. *Diversity linguistics comment*. http://dlc.hypotheses.org/691.

Hellwig, Birgit. 2011. *A grammar of Goemai*. Berlin: Mouton de Gruyter.

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

Himmelmann, Nikolaus P. 2006. Language documentation: What is it good for? In Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 1–30. Berlin: Mouton de Gruyter.

Hyslop, Gwendolyn. 2011. *A grammar of Kurtöp*. Eugene: University of Oregon. Doctoral dissertation.

Klamer, Marian. 2010. *A Grammar of Teiwa*. Berlin: Mouton de Gruyter.

Kratochvíl, Frantisek. 2007. *A Grammar of Abui*. Utrecht: Netherlands Graduate School of Linguistics. Doctoral dissertation.

Lidz, Liberty A. 2010. *A descriptive grammar of Yongning Na (Mosuo)*. Austin: University of Texas. Doctoral dissertation.

Mayer, Mercer. 1969. *Frog, where are you?* New York: Pied Piper.

Maxwell, Mike. 2012. Electronic grammars and reproducible research. In Nordhoff, Sebastian (ed.), *Electronic grammaticography* (Language Documentation & Conservation Special Publication No. 4), 207–234. Honolulu: University of Hawai'i Press.

McCracken, Chelsea. 2012. *A grammar of Belep*. Houston: Rice University. Doctoral dissertation.

Mihas, Elena. 2010. *Essentials of Ashéninka Perené grammar*. Milwaukee: University of Wisconsin. Doctoral dissertation.

Mosel, Ulrike. 2006. Grammaticography: The art and craft of writing grammars. In Ameka, Felix K., Alan Charles Dench & Nicholas Evans. *Catching language: the standing challenge of grammar writing*, 41–68. Berlin: Mouton de Gruyter.

Mosel, Ulrike. 2014. Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language. In Nakayama, Toshihide & Keren Rice (eds.), *The art and practice of grammar writing*, 135–157. Honolulu: University of Hawai'i Press.

Mushin, Ilana. 2012. *A grammar of (Western) Garrwa*. Berlin: Mouton de Gruyter.

Nakayama, Toshihide & Keren Rice (eds.). 2014. *The art and practice of grammar writing* (Language Documentation & Conservation Special Publication No. 8). Honolulu: University of Hawai'i Press.

Nichols, Peter John. 2011. *A morpho-semantic analysis of the persistive, alterative and inceptive aspects in siSwati*. London: University of London. Doctoral dissertation.

Pawley, Andrew. 2014. Grammar writing from a dissertation advisor's perspective. In Nakayama, Toshihide & Keren Rice (eds.), *The art and practice of grammar writing*, 7–23. Honolulu: University of Hawai'i Press.

Payne, Thomas E., & David J. Weber (eds.). 2007. *Perspectives on grammar writing*. Philadelphia: John Benjamins.

Pröll, Stefan & Andreas Rauber. 2013. Scalable data citation in dynamic, large databases: Model and reference implementation. In *2013 IEEE International Conference on Big Data*, 307–312.

R Core Team. 2013. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.

Renear, Allen H., Molly Dolan, Kevin Trainor & Melissa H. Cragin. 2009. Towards a cross-disciplinary notion of data level in data curation. *Proceedings of the American Society for Information Science and Technology* 46(1). 1–8.

Salffner, Sophie. 2010. *Tone in the phonology, lexicon and grammar of Ikaan*. London: SOAS, University of London. Doctoral dissertation.

Salffner, Sophie. 2015. A guide to the Ikaan language and culture documentation. *Language Documentation & Conservation* 9. 237–267. http://hdl.handle.net/10125/24639.

Sava, Graziano. 2005. *A grammar of Ts'amakko*. Cologne: Köppe.

Schnell, Stefan. 2010. *Animacy and referentiality in Vera'a*. Kiel: Kiel University. Doctoral dissertation.

SIL International. 2013. Toolbox (computer software), Version 1.6.1. http://www-01.sil.org/computing/toolbox.

de Sousa, Hilario. 2006. *The Menggwa Dla language of New Guinea*. Sydney: University of Sydney. Doctoral dissertation.

Starr, Joan, Eleni Castro, Mercè Crosas, Michel Dumontier, Robert R. Downs, Ruth Duerr, Laurel L. Haak, Melissa Haendel, Ivan Herman, Simon Hodson, Joe Hourclé, John Ernest Kratz, Jennifer Lin, Lars Holm Nielsen, Amy Nurnberger, Stefan Proell, Andreas Rauber, Simone Sacchi, Arthur Smith, Mike Taylor & Tim Clark. 2015. Achieving human and machine accessibility of cited data in scholarly publications.*PeerJ Computer Science* 1:e1. https://doi.org/10.7717/peerj-cs.1.

Swadesh, Morris. 1971. *The origin and diversification of language*. Sherzer, Joel (ed.) Chicago: Aldine Press.

Thieberger, Nicholas. 2004. Documentation in practice: developing a linked media corpus of South Efate. In Austin, Peter K. (ed.), *Language documentation and description* (2). 169–178. London: Hans Rausing Endangered Languages Project, Dept. of Linguistics, School of Oriental and African Studies.

Thieberger, Nicholas. 2006. *A grammar of South Efate: An Oceanic language of Vanuatu*. Honolulu: University of Hawai'i Press.

Thieberger, Nicholas. 2009. Steps toward a grammar embedded in data. In Epps, Patricia & Alexandre Arkhipov (eds.), *New challenges in typology: Transcending the borders and refining the distinctions*, 389–408. Berlin: Mouton de Gruyter.

Thieberger, Nicholas (ed.). 2012. *The Oxford handbook of linguistic fieldwork*. Oxford: Oxford University Press.

Thieberger, Nicholas. 2014. Digital humanities and language documentation. In Gawne, Lauren & Jill Vaughan (eds.), *Proceedings of the 44th Australian Linguistic Society Conference*, 144–159. Melbourne: The University of Melbourne.

Thieberger, Nicholas & Andrea L. Berez. 2012. Linguistic data management. In Thieberger, Nicholas (ed.), *The Oxford handbook of linguistic fieldwork*, 90–118. Oxford: Oxford University Press.

Thieberger, Nick, Anna Margetts, Stephen Morey & Simon Musgrave. 2016. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36(1). 1–21.

Vallejos Yopan, Rosa. 2010. *A grammar of Kokama-Kokamilla*. Eugene: University of Oregon. Doctoral dissertation.

van der Voort, Hein. 2004. *A grammar of Kwaza*. Berlin: Mouton de Gruyter.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. *Proceedings of* *the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1556–1559.

Woodbury, Anthony C. 2003. Defining documentary linguistics. *Language Documentation and Description* 1. 35–51.

Woodbury, Anthony C. 2011. Language documentation. In Austin, Peter K. & Julia Sallabank (eds.), *Cambridge handbook of endangered languages*, 159–186. Cambridge: Cambridge University Press.

Lauren Gawne
l.gawne@latrobe.edu.au
orcid.org/0000-0003-4930-4673

Barbara F. Kelly
b.kelly@unimelb.edu.au
orcid.org/0000-0002-0085-4917

Andrea L. Berez-Kroeker
andrea.berez@hawaii.edu
orcid.org/0000-0001-8782-515X

Tyler Heston
theston@hawaii.edu
orcid.org/0000-0002-1006-3684

Appendix A. Published grammars

Aikhenvald, Alexandra Y. 2003. *A grammar of Tariana, from northwest Amazonia*. New York: Cambridge University Press.

Alves, Mark. 2006. *A grammar of Pacoh: a Mon-Khmer language of the central highlands of Vietnam*. Canberra: Pacific Linguistics.

Baird, Louise. 2008. *A grammar of Klon*. Canberra: Pacific Linguistics.

van den Berg, René & Peter Bachet. 2006. *Vitu grammar sketch*. Ukarumpa: Summer Institute of Linguistics.

Bowern, Claire. 2012. *A grammar of Bardi*. Berlin: Mouton de Gruyter.

Burenhult, Niclas. 2005. *A grammar of Jahai*. Canberra: Pacific.

Childs, G. Tucker. 2011. *A grammar of Mani*. Berlin: Mouton de Gruyter

Chung, Ying Shing Anthony. 2005. *Descriptive grammar of Merei*. Canberra: Pacific Linguistics.

Coupe, Alec. 2007. *A grammar of Mongsen Ao*. Berlin: Mouton de Gruyter.

Courtz, Henk. 2008. *A Carib grammar and dictionary*. Toronto: Magoria Books.

Davies, William D. 2010. *A grammar of Madurese*. Berlin: Mouton de Gruyter.

Dhakal, Dubi Nanda. 2012. *Darai grammar*. Muenchen: Lincom.

Doctor, Raimond. 2004. *A grammar of Gujarati*. Muenchen: Lincom.

Epps, Patience. 2008. *A grammar of Hup*. Berlin: Mouton de Gruyter.

Fenwick, Rohan S. H. 2011. *A grammar of Ubykh*. Muenchen: Lincom.

Frajzyngier, Zygmunt. 2012. *A grammar of Wandala*. Berlin: Mouton de Gruyter.

Genetti, Carol. 2007. *A grammar of Dolakha Newar*. Berlin: Mouton de Gruyter.

Georg, Stefan. 2007. *A descriptive grammar of Ket (Yenisei-Ostyak)*. Folkestone: Global Oriental.

Givón, Talmy. 2011. *Ute reference grammar*. Philadelphia: John Benjamins.

Guillaume, Antoine. 2008. *A grammar of Cavineña*. Berlin: Mouton de Gruyter.

Heath, Jeffrey. 2008. *A grammar of Jamsay*. Berlin: Mouton de Gruyter.

Hedinger, Robert. 2008. *A grammar of Akoose*. Dallas: SIL International Publications.

Hellwig, Birgit. 2011. *A grammar of Goemai*. Berlin: Mouton de Gruyter.

Huber, Juliette. 2008. *First steps towards a grammar of Makasae: A language of East Timor*. Muenchen: Lincom.

Janhunen, Juha. 2005. *Khamnigan Mongol*. Muenchen: Lincom.

Kari, Ethelbert Emmanuel. 2004. *Degema* Muenchen: Lincom.

Klamer, Marian. 2010. *A grammar of Teiwa*. Berlin: Mouton de Gruyter.

Kruspe, Nicole. 2004. *A grammar of Semelai*. New York: Cambridge University Press.

LaPolla, Randy J., & Chenglong Huang. 2003. *A grammar of Qiang with annotated texts and glossary*. Berlin: Mouton de Gruyter.

Lichtenberk, Frank. 2008. *A grammar of Toqabaqita*. Berlin: Mouton de Gruyter.

Lu, Qiao Tan. 2008. *A grammar of Manoan*. Boca Raton: Universal Publishers.

Maslova, Elena. 2003. *A grammar of Kolyma Yukaghir*. Berlin: Mouton de Gruyter.

McLaughlin, John E. 2012. *Shoshoni grammar*. Muenchen: Lincom.

Morley, Eric A. 2010. *A grammar of Ajagbe*. Muenchen: Lincom.

Mu'azu, Mohammed Aminu & Katwal Pemak Isah. 2009. *A grammar of the Miship language*. Muenchen: Lincom.

Mushin, Ilana. 2012. *A grammar of (Western) Garrwa*. Berlin: Mouton de Gruyter.

Nau, Nicole. 2011. *A short grammar of Latgalian*. Muenchen: Lincom.

Olawsky, Knut J. 2006. *A grammar of Urarina*. Berlin: Mouton de Gruyter.

Peterson, John. 2011. *A grammar of Kharia*. Leiden: Brill.

Sava, Graziano. 2005. *A grammar of Ts'amakko*. Köppe: Köln.

Schneider, Cynthia. 2010. *A grammar of Abma: A language of Pentecost Island, Vanuatu*. Canberra: Pacific Linguistics.

Slater, Keith W. 2003. *A grammar of Mangghuer: A Mongolic language of China's Qinghai-Gansu sprachbund*. London: RoutledgeCurzon.

Teng, Stacy Fang-Ching. 2008. *A reference grammar of Puyuma, an Austronesian language of Taiwan*. Canberra: Pacific Linguistics.

Terrill, Angela. 2003. *A grammar of Lavukaleve*. Berlin: Mouton de Gruyter.

Thieberger, Nick. 2006. *A grammar of South Efate: An Oceanic language of Vanuatu*. Honolulu: University of Hawai'i Press.

Tsunoda, Tasaku. 2011. *A grammar of Warrongo*. Berlin: Mouton de Gruyter.

Van Otterloo, Roger. 2011. *The Kifuliiru language: A descriptive grammar*. Dallas: SIL International.

van de Velde, Mark L. O. 2008. *A grammar of Eton*. Berlin: Mouton de Gruyter.

van der Voort, Hein. 2004. *A grammar of Kwaza*. Berlin: Mouton de Gruyter.

Zeitoun, Elizabeth. 2007. *A grammar of Mantauran (Rukai)*. Taiwan: Academia Sinica.

Appendix B. Dissertations

Ahland, Colleen. 2012. *A grammar of Northern and Southern Gumuz*. Eugene: University of Oregon. Doctoral dissertation.

Ahland, Michael. 2012. *A Grammar of Northern Mao (Màwés Aas'è)*. Eugene: University of Oregon. Doctoral dissertation.

de Araujo, Gabriel Antunes. 2004. *A grammar of Sabané, a Nambikwaran language*. Utrecht: Netherlands Graduate School of Linguistics. Doctoral dissertation.

Berghäll, Liisa. 2010. *Mauwake reference grammar*. Helsinki: University of Helsinki. Doctoral dissertation.

Berthiaume, Scott Charles. 2003. *A phonological grammar of Northern Pame*. Arlington: University of Texas at Arlington. Doctoral dissertation.

van Breugel, Jonkheer Egbert Joost Seino Clifford Kocq. 2008. *A grammar of Atong*. Melbourne: La Trobe University. Doctoral dissertation.

Brochie, Amanda. 2009. *Tirax grammar and narrative: An Oceanic language spoken on Malakula, North Central Vanuatu*. Melbourne: The University of Melbourne. Doctoral dissertation.

Buragohain, Dipima. 2010. *A descriptive grammar of Tai Ahom*. New Delhi: Jawaharlal Nehru University. Doctoral dissertation.

Chacon, Thiago Costa. 2012. *The phonology and morphology of Kubeo: The documentation, theory, and description of an Amazonian language*. Honolulu: University of Hawai'i. Doctoral dissertation.

Cutfield, Sarah Anne. 2012. *Demonstratives in Dalabon: A language of southwestern Arnhem Land*. Melbourne: Monash University. Doctoral dissertation.

Doornenbal, Marius. 2009. *A grammar of Bantawa*. Utrecht: Netherlands Graduate School of Linguistics. Doctoral dissertation.

Fedden, O. S. 2007. *A grammar of Mian, a Papuan language of New Guinea*. Melbourne: University of Melbourne. Doctoral dissertation.

Felix Armendariz, Rolando Gpe. 2006. *A grammar of River Warihio*. Houston: Rice University. Doctoral dissertation.

Filchenko, Andrey Yury. 2007. *A grammar of Eastern Khanty*. Houston: Rice University. Doctoral dissertation.

Fleck, David William. 2003. *A grammar of Matses*. Houston: Rice University. Doctoral dissertation.

Fried, Robert Wayne. 2010. *A grammar of Bao'an Tu, a Mongolic language of Northwest China*. Buffalo: State University of New York at Buffalo. Doctoral dissertation.

Gaby, Alice Rose. 2006. *A grammar of Kuuk Thaayorre*. Melbourne: University of Melbourne. Doctoral dissertation.

Guerin, Valerie. 2008. *A grammar of Mavea: An Oceanic language of Vanuatu*. Honolulu: University of Hawai'i. Doctoral dissertation.

Hill, Peter M. 2011. *Morphology and sentence construction in Kurrama: A language of the Pilbara region of Western Australia*. Perth: University of Western Australia. Doctoral dissertation.

Hyslop, Gwen. 2011. *A grammar of Kurtöp*. Eugene: University of Oregon. Doctoral dissertation.

Jansen, Joana Worth. 2010. *A grammar of Yakima Ichishkiin/Sahaptin*. Eugene: University of Oregon. Doctoral dissertation.

Khalilova, Zaira. 2009. *A grammar of Khwarshi*. Utrecht: Netherlands Graduate School of Linguistics. Doctoral dissertation.

Kratochvíl, Frantisek. 2007. *A grammar of Abui*. Utrecht: Netherlands Graduate School of Linguistics. Doctoral dissertation.

Lanz, Linda A. 2010. *A grammar of Inupiaq Morphosyntax*. Houston: Rice University. Doctoral dissertation.

Lidz, Liberty A. 2010. *A descriptive grammar of Yongning Na (Mosuo)*. Austin: The University of Texas. Doctoral dissertation.

Loughnane, Robin. 2009. *A grammar of Oksapmin*. Melbourne: University of Melbourne. Doctoral dissertation.

McCracken, Chelsea. 2012. *A grammar of Belep*. Houston: Rice University. Doctoral dissertation.

Michael, Lev David. 2008. *Nanti evidential practice: Language, knowledge, and social action in an Amazonian society*. Austin: The University of Texas. Doctoral dissertation.

Mihas, Elena. 2010. *Essentials of Ashéninka Perené grammar*. Milwaukee: University of Wisconsin-Milwaukee. Doctoral dissertation.

Miller, Mark Turner. 2007. *A grammar of West Coast Bajau*. Arlington: University of Texas at Arlington. Doctoral dissertation.

Montgomery-Anderson, Brad. 2008. *A reference grammar of Oklahoma Cherokee*. Lawrence: University of Kansas. Doctoral dissertation.

Nchare, Abdoulaye Laziz. 20120. *The grammar of Shupamem*. New York: New York University. Doctoral dissertation.

Nichols, Peter John. 2011. *A morpho-semantic analysis of the persistive, alterative and inceptive aspects in siSwati*. London: SOAS, University of London. Doctoral dissertation.

Peterson, Tyler Roy Gösta. 2010. *Epistemic modality and evidentiality in Gitksan at the semantics-pragmatics interface*. Victoria: University of British Columbia. Doctoral dissertation.

Petzell, Malin. 2008. *The Kagulu language of Tanzania: Grammar, texts and vocabulary*. Gothenburg: University of Gothenburg. Doctoral dissertation.

Post, Mark. 2007. *A grammar of Galo*. Melbourne: La Trobe University. Doctoral dissertation.

Quick, Phil. 2003. *Grammar of the Pendau language*. Canberra: Australian National University. Doctoral dissertation.

Rapacha, Lal B. 2005. *A descriptive grammar of Kiranti-Kõits*. New Delhi: Jawaharlal Nehru University. Doctoral dissertation.

Salffner, Sophie. 2010. *Tone in the phonology, lexicon and grammar of Ikaan*. London: SOAS, University of London. Doctoral dissertation.

Sangdong, David. 2012. *A grammar of the Kadu (Asak) language*. Melbourne: La Trobe University. Doctoral dissertation.

Schnell, Stefan. 2010. *Animacy and referentiality in Vera'a*. Kiel: Kiel University. Doctoral dissertation.

Seyoum, Mulugeta. 2008. *A grammar of Dime*. Utrecht: Netherlands Graduate School of Linguistics. Doctoral dissertation.

de Sousa, Hilário. 2006. *The Menggwa Dla language of New Guinea*. Sydney: The University of Sydney. Doctoral dissertation.

Vallejos Yopan, Rosa. 2010. *A grammar of Kokama-Kokamilla*. Eugene: University of Oregon. Doctoral dissertation.

Vokurková, Zuzana. 2008. *Epistemic modalities in spoken Standard Tibetan*. Paris: Universite Paris 8. Doctoral dissertation.

Waldie, Ryan James. 2012. *Evidentiality in Nuu-chah-nulth*. Victoria: University of British Columbia. Doctoral dissertation.

Whitehead, Carl R. 2004. *A reference grammar of Menya, an Angan language of Papua New Guinea*. Winnipeg: The University of Manitoba. Doctoral dissertation.

Wilde, Christopher P. 2008. *A sketch of the phonology and grammar of Rajbanshi*. Helsinki: University of Helsinki. Doctoral dissertation.

Willis, Christina M. 2007. *A descriptive grammar of Darma: An endangered Tibeto-Burman language*. Austin: University of Texas. Doctoral dissertation.

Yoshioka, Noboru. 2012. *A reference grammar of Eastern Burushaski*. Tokyo: Tokyo University of Foreign Studies. Doctoral dissertation.