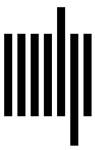
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022



The MIT Press

26 Managing Data for Descriptive and Historical Research

Don Daniels and Kelsey Daniels

1 Introduction

In this chapter we describe the data management workflow we have used in our fieldwork in Madang Province, Papua New Guinea. The bulk of the chapter is devoted to the data management practices we used during the first author's time as a postdoctoral researcher at the Centre of Excellence for the Dynamics of Language at the Australian National University. During this period, the authors conducted a pair of two-month field trips together in 2016, and the first author also conducted a solo trip in 2018. These were all to the Astrolabe Bay area of Madang. In D. Daniels's doctoral research, he conducted solo trips to a different part of the province, the Middle Ramu. We briefly describe the data management workflow used for this project later in the chapter.

In both cases, the goals of this research were primarily descriptive: We were documenting and describing a few essentially undescribed Papuan languages of the Rai Coast and Sogeram branches of the Madang branch of Trans New Guinea (Pawley & Hammarström 2018; Daniels 2015). But one of D. Daniels's primary research interests is historical-comparative linguistics (e.g., Daniels 2014, 2017, 2019), and this interest informed the research agenda in important ways. Most notably, it meant that we conducted shorter stints of research in multiple communities, rather than a longer spell in one community. For the Rai Coast project our research goal was to be able to write a decent grammatical sketch for each of three languages (Bongu, Soq, and Jilim) with the data we collected on these trips. In practical terms, this meant that, for each language, we hoped to collect word lists, conduct basic grammatical elicitation, and record and transcribe a sizeable corpus of naturalistic speech (we hoped for an hour per language per trip). We discuss how the historical orientation of the first author's research agenda influenced the research program toward the end of the chapter.

It will be helpful to describe the logistical situation in Astrolabe Bay before discussing the actual workflow. The villages that we worked in—Bongu, Kaliku, and Jilim—are all rural. The first two are usually serviced by a road during the dry season, but not the wet; during the wet season, the villages are accessible via small motorboats that travel along the coast. Jilim has no road access but is reached by a three-hour walk from the coastal road. None of the villages have electricity, but in each village, there are a few generators around—often in various states of disrepair—that individuals have acquired for one reason or another. A primary difficulty of fieldwork in this context is keeping your electronic equipment, especially laptops, functioning.

The bulk of this chapter is devoted to describing our data management process, which we do in the next section. Section 3 describes some of the differences between this process and the data management practices the first author used in previous research. Section 4 discusses the ways a focus on historical research informed our choices on the field, and we offer some concluding thoughts in section 5.

2 Fieldwork with portable solar panels

The process of managing data consists of three phases. The most involved workflow takes place in the villages where we conduct the fieldwork; this is where primary data collection and processing take place (section 2.1). Then when we return from the villages to the provincial capital of Madang we process some of the collected materials (section 2.2), and we complete the data processing once we return home (section 2.3).

2.1 Data management in the village

There is a lot to do in the village. We have broken our process down into three broad steps: making recordings

(section 2.1.1), transcribing recordings (section 2.1.2), and what we call "end-of-day processing," a daily routine of backing up files and processing metadata (section 2.1.3). An added task is equipment management, which is ongoing (section 2.1.4). A key point about this phase of data collection is that routine is paramount (Mattern, chapter 5, this volume). There are so many tasks to perform, and so much to keep track of, that without a solid routine to guide the researcher, the number of decisions to be made quickly becomes overwhelming.

2.1.1 Making recordings Three types of recordings are made on each trip: elicitation (both lexical and grammatical), natural-language recordings, and informed consent. Before arriving in the village, a field notebook is prepared with a word list based on Z'graggen (1980) and basic grammatical elicitation paradigms. After arriving in the village, the word list is usually the first recording made, if possible, with two language users. Elicitation of basic grammatical paradigms follows soon after. This includes verbal subject-agreement paradigms for several categories, including present, future, multiple pasts, habitual, imperative, and switch reference, and pronominal paradigms for transitive subject, object, and possessor. This initial skeletal session is later supplemented with additional elicitation. In each case, the session is recorded in audio but not video, and transcriptions are recorded in the prepared notebook.

Natural-language sessions are recorded in both audio and video format. We generally try to record a diverse corpus, with representation of men and women, multiple clans (Stanford 2009), and a variety of genres.

Informed consent is usually obtained after each recording so language users are fully aware of what stories or personal information they might be disclosing to others. A description of the purpose and use of the recordings is provided to each language user both in speech and in writing, but, due to low levels of literacy in the area, consent is recorded orally (see Collister, chapter 9, this volume, and Holton, Leonard, & Pulsifer, chapter 4, this volume, for more discussion around consent and rights to data).

Metadata about each recording and each language user is recorded in field notebooks. The first two pages of each notebook contain a table of contents for that notebook; recording metadata are listed next, usually on pages 3–6, and language user metadata around pages 7–10. Recording metadata include the original file names created by the recording devices, name(s) of the language user, and the title of the recording (often in two or three languages). If multiple audio and video recorders, or multiple secure digital (SD) cards, are being used, we also note which device and which memory card a recording was created on.[1] Language user metadata include name, date of birth, level of education, clan membership, childhood residence, languages used, marital status, and number of children. An example is given in figure 26.1.

2.1.2 Transcription Before natural-language recordings are transcribed, audio files are entered into a program called SayMore (Moeller 2014), where we segment the recording into intonation units. SayMore was chosen for a few reasons. It creates an ELAN file (.eaf) for each recording that makes the data easy to import into other programs. It has simple interfaces for audio segmentation and transcription; this helps engage consultants in the work of transcription, and the interfaces are easier to teach to advanced consultants than are the corresponding interfaces in a more powerful program, such as ELAN. Finally, SayMore automatically generates a file directory for each language project, which helps with keeping files organized.

After the audio file has been segmented, SayMore's transcription interface presents an ELAN tier for the target language transcription and another for the analysis language translation, each consisting of the same time-aligned audio segments. Clicking inside any segment automatically plays the corresponding audio on loop to facilitate the transcription or translation of that text.

We spend most of our days in each village transcribing texts. Transcription typically involves one of the authors sitting at a laptop with one language user, each of us with a pair of earbuds in our ears. (Our field laptops typically do not have good built-in speakers.) We play a chunk of the recording, which the consultant repeats slowly so we can transcribe it. Then the consultant translates the chunk into Tok Pisin, and we transcribe that as well.

As we proceed, we also take notes about the recording. SayMore works well for time-aligned line-by-line transcription and translation, but there is no place to record transcription notes or word-by-word translations. For this reason, we write word-by-word translations and other notes in our notebooks. To save time, we typically write only the first letter of each target language word with the word-by-word translation and notes below that. We also often conduct spontaneous elicitation inspired by the text we are transcribing, and this is also

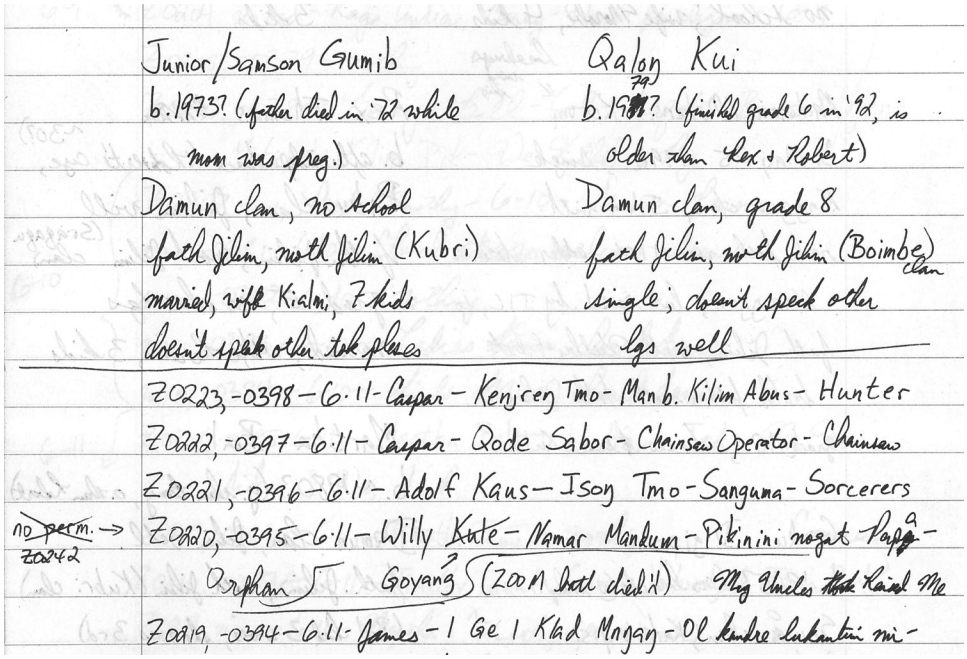


Figure 26.1
An excerpt from a field notebook, showing speaker metadata and recording metadata.

recorded in the notebooks. A sample is given in figure 26.2. These notebooks are then scanned when we arrive back in town.

2.1.3 End-of-day processing At the end of each day, all files are saved to an external hard drive, and all metadata from the field notebooks are entered in a master spreadsheet. Once all the recording devices have been collected, the contents of each SD card (photos, audio files, video files, and such) are copied to the external hard drive. Copies of all new audio files are also copied to the SayMore directory.

While these files are copying, metadata from all new recordings are copied from the field notebook to the master spreadsheet.[2] The master spreadsheet includes all the recording metadata contained in the field notebooks; it also includes information about corresponding SayMore names for each natural-language recording, the notebook locations of transcription notes, and the transcription status for each natural-language recording (i.e., whether it is untranscribed, only segmented, partially transcribed, or fully transcribed).

One of the most important steps at the end of each day is to standardize the files across the hard drive and both laptops. Copying an updated SayMore transcription file to another laptop requires overwriting the previous SayMore file on that laptop; therefore, it is extremely important for us to save a copy of the updated SayMore transcription file to the hard drive before overwriting any laptop files. This process of transferring and updating files on all the various devices easily gets confusing, so we created a column in the master spreadsheet as a "transfer log" to systematically document which files are most up-to-date and which files have been copied to which devices. The master spreadsheet is copied to the hard drive each day. At the end of this process, both laptops and the external hard drive have a copy of the most recent SayMore transcripts. All photos, videos, and audio recordings are stored both on the original SD cards and on the external hard drive. Depending on how much work is accomplished throughout the day, this process can take up to an hour every afternoon.

It is important to note that the challenges of version control mentioned have long been addressed by software engineers—for example, one reviewer suggested the use of an application such as Git, which can permanently store previous file versions to avoid the hazards of accidentally overwriting or deleting transcriptions. We did not use Git in the field, however, primarily because (1) we did not know about it, and (2) we did not have Internet access to sync laptop versions. Such version control applications seem promising for fieldwork situations like ours, but they require more a bit more investigation before implementation.

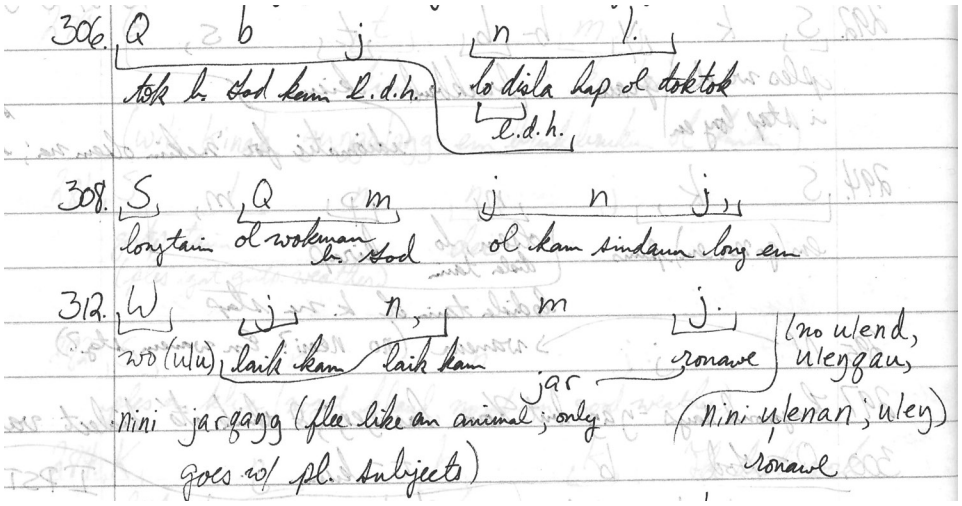


Figure 26.2
An excerpt from a field notebook, showing transcription notes.

2.1.4 Equipment management Working in rural villages with no electricity requires advanced planning and consistent routines to keep all the devices and batteries charged. To ease this burden somewhat, we have chosen, whenever possible, to use recording devices and transmitters that require AA batteries. And before leaving the United States, we buy the batteries we need, as the life spans of those bought in Papua New Guinea tend to be quite short.

The camera that we use requires a lithium ion rechargeable battery that we recharge using a small portable universal charger—bought in Papua New Guinea—that transfers power from AA batteries to the camera battery. We always bring a few spare lithium ion batteries, allowing us to charge one while we are using another.

Likewise, the two 24,000 mAh PowerGorilla power banks that charge the laptops must regularly be reconnected to solar panels. When we first arrive in the village, D. Daniels finds a trustworthy homeowner whose roof gets decent access to direct sunlight and asks them if we can install the canvas solar panels on their roof. After installing them, each morning D. Daniels goes to their house to check that the panels are functioning properly and to unplug the charged power bank and plug in the recently used power bank. Throughout the day, we alternate which laptop is connected to the power bank. Additionally, the laptops are always set to power-saving mode, to the lowest screen-brightness setting, and to airplane mode. These power banks and solar panels can usually power the laptops for approximately five to six hours a day for a bit more than one week. Data processing is particularly power-intensive, as the external hard drive we use draws power from the laptops, further draining their batteries.

When both laptops and the power banks are depleted, everything needs to be recharged using a generator. In each village we worked in in 2016, at least one person owned a generator. We purchased fuel for them, and they allowed us to recharge all our devices using their generator over the course of an evening. One danger in using locally owned generators is power surges. Depending on the quality of the generator, surges of power can abruptly run through the power cords setting fire to the laptop chargers and the power banks. On one occasion when this happened, we were fortunately able to buy a replacement charger from a computer supply store on our next trip to town. Because of this risk, it is important for someone to be near the generator at all times while it is running, to unplug electronics in case of a surge. Once all the laptops and the power banks are recharged, we can usually operate our devices using only the solar panels and power banks for another week. Another risk with using local generators is that the generator simply will not work; this happened in 2018 and meant that for the last few days in the village, the first author had to rely on solar power. Due to this, he is now planning on buying a high-quality generator for the village.

Another aspect of equipment management is preparing field notebooks. We use two-hundred-page composition notebooks and archival-quality pens, which ensure that ink does not run if it gets wet. Before arriving at the field site, we enter a notebook title and a table of contents on the first page, leave six or seven pages for recording metadata[3] after that, and then prepare any elicitation tasks we plan to complete (word lists, grammatical elicitation, and so on). We also number every page in the

notebook, which aids tremendously in tracking data after it has been entered into FieldWorks Language Explorer (FLEx) and other analysis programs.

We must also consider the safety of the equipment. Equipment is most vulnerable during travel, so we take special care when packing our bags to minimize the risk. All electronics are placed in dry-bags to prevent water damage. If we are transporting data (that is, if we are leaving the village), we try to ensure that duplicates are spread across different bags. For example, if SD cards with original recordings are in one bag, the hard drive with backups is put in another bag. In this way, most data will be safe even in the event of a lost or stolen bag.

2.2 Data management in town

When we return from the village, our data are primarily located in two places: field notebooks and an external hard drive. The field notebooks contain consultant metadata, basic metadata about recordings, transcripts of elicitation sessions, and notes from transcription sessions. The hard drive contains all our recordings and photos, the metadata spreadsheet, and transcripts of the naturalistic recordings. Redundant copies of these files are also located on SD cards and our field laptops.

Upon arriving in town, we have two primary tasks. The first is to scan the field notebooks. We use the offices of some friends to do this and copy the scans to our hard drive. The second task is then to copy everything from our first hard drive (the "village" hard drive) to our "town" hard drive. Then, if we return to the villages for more fieldwork, the town hard drive stays with friends in town. If it is time to leave the country, we once again distribute all data between our bags before boarding our flight home.

2.3 After returning from the field

Once we have arrived home, hopefully with all our hard drives intact, there is still a considerable amount of work to be done. In section 2.3.1, we describe the work involved in archiving the materials we collected, and in section 2.3.2, we describe preparing those materials for our own research.

2.3.1 Archiving One of the primary goals of this research is to produce an enduring and open record of the languages and communities under investigation (Gawne & Styles, chapter 2, this volume). As such, we archive the materials we collect in D. Daniels's (2018) Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) collection.[4] Preparing them for archiving involves preparing more detailed metadata, renaming the files, and delivering them to the archive.

D. Daniels maintains a personal metadata spreadsheet (Daniels 2010a), which contains significantly more detail than our field metadata spreadsheet, and filling this in is the first item of business after returning home. The field spreadsheet is organized by recording files. Each separate audio or video file is given its own line. This information is copied into the personal spreadsheet, but the personal spreadsheet also organizes recordings into events. This way, for example, an audio and a video recording of the same story are grouped together. The personal spreadsheet also includes a description of the content of each recording event, which is produced from memory at this stage in the workflow. Metadata about transcripts—both those in the field notebooks and ELAN transcripts produced with SayMore—is also entered into this spreadsheet. For the audio, video, and transcription files, the spreadsheet includes personal file names (which are meaningful) and the PARADISEC file names (which are not). It also includes detailed information about language users, which is copied from the field notebooks.

Once the personal metadata spreadsheet is filled out, copies of the files produced during fieldwork are created. They are renamed in accordance with PARADISEC's file-naming conventions and given to PARADISEC. This involves copying a subset of the metadata from the personal metadata spreadsheet onto PARADISEC's intake spreadsheet and transferring that, along with the renamed files, to PARADISEC. Personal copies of the files are retained with the first author's personal file-naming conventions.

In our experience, it is important to attend to data processing as soon as possible after returning from the field (see also Mattern, chapter 5, this volume). In our own process, responsibility for this stage of work has fallen to the first author, and he has found that there are invariably hiccups in the process that are much easier to get through when the whole experience is fresh in your memory. The longer data processing is delayed, the longer it takes, and the higher the likelihood that some information will be lost or corrupted (Han, chapter 6, this volume). D. Daniels generally expects the whole process to take two full weeks of eight-hour days.

2.3.2 Preparing materials for research Once everything is archived, it is time for D. Daniels to begin moving materials through the research workflow. This can

be conceived of as using primary data to create different kinds of more abstract data (Good, chapter 3, this volume; Han, chapter 6, this volume). At this stage he is primarily working with transcripts in FLEx. The first author also has experience using Toolbox, and he has found FLEx to be more powerful, more reliable, and far more useful. The ability to edit morphological analyses as one goes, in particular, is essential for an analyst working with an unfamiliar language. The search functionality is significantly better, as is the technical support from the developer. FLEx also interfaces much better with other programs, most importantly SayMore and ELAN. This allows the analyst to take advantage of ELAN's very powerful "Structured Search Multiple eaf" function.

How data are entered into FLEx depends on the type of data being entered. Transcripts created in SayMore can be automatically imported (Pennington 2014). First they are exported from SayMore by highlighting the relevant transcript and selecting "Export > FLEx interlinear text," which creates an XML file with a ".flextext" extension. This file is then imported into FLEx as an interlinear text, which imports each vernacular line with its Tok Pisin free translation and preserves the time stamps for future exporting. As each text is imported into FLEx, basic metadata about the text are copied to the "Info" tab.

D. Daniels also creates additional "texts" to store other kinds of multimorphemic data, such as morphological paradigms and syntactic elicitation. This is necessary so this data can be morphologically parsed. D. Daniels generally makes four of these dummy texts: one for verb paradigms, one for noun paradigms, one for all other multimorphemic utterances he has recorded, and one for all utterances he has recorded as being ungrammatical. Creating a special place for storing ungrammatical utterances ensures that (1) he will have access to that information when he is conducting research on a particular feature in FLEx, and (2) he will see immediately whether something is ungrammatical based on the text it is in.

Lexical data from our field notebooks needs to be copied manually into FLEx. Word lists can be entered in the Lexicon view or using the "Collect words" interface (this is more useful when dealing with a single semantic domain, such as words for trees, because all entries can be automatically tagged as belonging to the same semantic field). The notebook page where a word is recorded is entered into that lexeme's "Source" field, to facilitate double-checking the source material when necessary.

The final step is parsing. In each text in FLEx, whether it is a genuine transcript of a recording or a dummy text storing morphological paradigms, the user can parse words morphologically. In this process, the user splits the word into its constituent morphemes and tells the program which specific lexical entry each morpheme corresponds to. This turns the texts into a searchable database and allows the analyst to bring up all tokens of a particular morpheme at will. The first author enters lexical glosses in English in FLEx. This means that word-by-word and morpheme-by-morpheme translations are English, but line-by-line translations, which are imported directly from field transcripts, are in Tok Pisin. These choices are largely driven by convenience. The use of Tok Pisin in the field is dictated by the language abilities of our consultants. The use of English at the morphological level means that example sentences can be exported directly into papers and presentations; if Tok Pisin were used here, exporting would mean translating every morpheme afresh, each time.

D. Daniels prefers to proceed with morphological parsing by moving through field notebooks. He begins by parsing in FLEx whatever is recorded on the first pages of field notebook 1 and completes notebook 1 before continuing on to notebook 2. As he parses, he has SayMore open so that he can play back any line in the text that he is working on. He also has a skeleton outline of a grammar sketch open in Word. As he encounters interesting examples, or constructions that he does not understand, he copies them into the appropriate section of the sketch (for instance, an example involving a nominalized verb would go into a section on verb nominalization). He also keeps a list of follow-up questions to ask on his next field trip.

3 Fieldwork without portable solar panels

In the preceding sections we have primarily described the data management practices we used during two trips we went on together in 2016, and during a trip the first author conducted alone in 2018. But for the first author's PhD fieldwork, which he conducted between 2010 and 2014, he worked without field laptops or solar panels, and consequently had to develop different data management practices. We would not recommend this approach over the more technology-heavy approach described in section 2—we have found that the added hassle of solar panels and laptops is more than compensated for

by the added capabilities they provide. Nevertheless, we recognize that it may not always be possible to acquire all the gear needed for this kind of fieldwork in remote locations, so in this section we briefly describe the differences between the 2016–2018 fieldwork and the 2010–2014 fieldwork.

The main difference, as mentioned, was that for the 2010–2014 fieldwork D. Daniels did not bring a laptop to the village, which meant he did not need a solar panel. All the power he needed was provided by AA batteries, which could power his camera, his audio recorder, and his video recorder. This difference in equipment primarily affected the work in two ways. It changed the transcription process, and it limited the amount of metadata recording that was possible in the field.

Without a laptop, all transcription had to be done in field notebooks. Transcription is always a time-intensive task, but this limitation made the process even slower. Additionally, audio files couldn't be cut into intonation units before meeting with a speaker but had to be rewound manually and played back. For this reason, audio recorders were selected based, in part, on their built-in playback capabilities. D. Daniels opted for the Olympus LS-10 and LS-14, which enabled him to manually fast-forward and rewind through a recorded file and play back portions of it as needed. After returning from the village, the fact that all transcripts were handwritten in field notebooks also meant that they then needed to be typed into FLEx; the ease of exporting typed transcripts from SayMore to FLEx is a major improvement in this regard.

All metadata also had to be recorded in the notebooks and could not be converted to electronic form until leaving the village. This made the task of entering metadata more time-consuming because it had to be delayed until weeks after many recordings had been made.

4 Historical research

Readers will have noted that relatively little of the fieldwork methodology we have chosen seems to have anything to do with historical linguistics, even though the first author's work is primarily historical (e.g., Daniels, Barth, & Barth 2019; Daniels 2010b, 2020a,b) and this chapter is ostensibly about managing data for historical research. This impression is mostly correct: the research agenda in the field is primarily designed to produce data for synchronic description. We see historical linguistics as something that is done *after* synchronic analysis has been carried out and consequently aim to collect material with which to write grammatical descriptions.

Nevertheless, there are some ways that D. Daniels's historical research goals did inform our choices in the field. One is the collection of a standard word list. We collected a list of meanings based on Z'graggen's (1980) work for every language, which increased the odds of finding cognates for lexical reconstruction. Another is that we prioritized elicitation of formal linguistic features over less easily visible ones. By this we mean linguistic features that have a formal, phonological realization, as opposed to features that are simply stored in speakers' memory but have no overt realization. For example, a verb's morphologically irregular inflectional form is realized with overt phonological material, while that verb's subcategorization frame is not. We spent significantly more time collecting information of the former kind than the latter because it would be more likely to be useful for historical reconstruction.

D. Daniels's historical-linguistic goals also informed his choices during data processing and analysis. As he analyzed and described the data we had collected, he maintained a document with notes about constructions that might be cognate. This was particularly useful because he was not just interested in lexical cognates, but also syntactic ones, and prose notes could more easily express relations of cognacy between, for example, a particular stem shape in one language and a serial verb construction in another.

For keeping track of more traditional lexical and morphological cognates, the first author simply used spreadsheets. Separate worksheets were devoted to various morphological categories (one for verb suffixes, one for pronouns, one for lexemes, and so on), but each worksheet was organized similarly. Each language was given a column, and each row consisted of semantically similar forms. In cases of semantic innovation, where a form in one row was cognate with a form in another, D. Daniels used color coding to keep track of cognacy relations.

5 Conclusions

We hope these reflections will be useful to fieldworkers in the future and will help researchers design metadata workflows that are effective for their particular situations. We recognize that every project, every researcher,

and every language community is different, and we encourage readers to adapt from our workflow those things that they find useful and to ignore those things that seem ill-suited to their situation. Our primary recommendation is to set up a solid routine: create a set of daily expectations and stick to them. Like any chore, staying on top of data is manageable if it is done regularly, but it quickly becomes unmanageable if delayed for a few days. This is especially true if, as will hopefully be the case, fieldwork is going well, and you are recording lots of data.

Another recommendation we have is to pay attention to your relationships in the field. We hope this goes without saying, but fieldwork is much more than an exercise in collecting data. Although for this chapter we focused on the nuts and bolts of how to record and manage data, successful fieldwork will also involve making friends, building collaborative relationships, and learning how to navigate what will often be a very foreign cultural landscape. These aspects of fieldwork probably deserve even more attention than data, but that is a topic for another time.

Notes

We are grateful to the editors and two anonymous reviewers for comments on earlier versions. We are also immensely grateful to people in the various communities where we have conducted fieldwork for their tireless help and steady companionship. All remaining errors are our own. The authors' joint fieldwork was supported by the ARC Centre of Excellence for the Dynamics of Language.

1. We have found it helpful to distinguish SD cards by affixing a unique sticker to each one.

2. Although SayMore has specified fields for recording metadata, we have chosen to record this information in spreadsheets because the first author used spreadsheets when he first started his career in linguistics and because spreadsheets are easier for searching for and copying information.

3. As mentioned, SayMore can easily store metadata about texts and speakers; however, we choose to write this information in our field notebooks as they are more readily accessible when we are recording stories with speakers.

4. PARADISEC is an archive focused on preserving language materials, with a historical connection to the Pacific region.

References

Daniels, Don. 2010a. Metadata spreadsheet. doi:10.4225/72/56F00F2CCA5E9. http://catalog.paradisec.org.au/collections/DD1/items/068. Accessed December 12, 2018.

Daniels, Don. 2010b. A preliminary phonological history of the Sogeram languages of Papua New Guinea. *Oceanic Linguistics* 49 (1): 163–193.

Daniels, Don. 2014. Complex coordination in diachrony: Two Sogeram case studies. *Diachronica* 31 (3): 379–406.

Daniels, Don. 2015. A reconstruction of Proto-Sogeram: Phonology, lexicon, and morphosyntax. PhD dissertation, University of California, Santa Barbara.

Daniels, Don. 2017. A method for mitigating the problem of borrowing in syntactic reconstruction. *Studies in Language* 41 (3): 577–614.

Daniels, Don. 2018. Papuan Languages Collection. Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). http://catalog.paradisec.org.au/collections/DD1. (Archival collection, 960 items).

Daniels, Don. 2019. Using phonotactics to reconstruct degrammaticalization: The origin of the Sirva pronoun *be*. *Diachronica* 36 (1): 1–36.

Daniels, Don. 2020a. *Grammatical reconstruction: The Sogeram languages of New Guinea*. Berlin: Mouton.

Daniels, Don. 2020b. The history of tense and aspect in the Sogeram family. *Journal of Historical Linguistics* 10 (2): 167–208. doi:10.1075/jhl.18012.dan.

Daniels, Don, Danielle Barth, and Wolfgang Barth. 2019. Subgrouping the Sogeram languages: A critical appraisal of historical glottometry. *Journal of Historical Linguistics* 9 (1): 92–127. doi:10.1075/jhl.17011.dan.

Moeller, Sarah Ruth. 2014. SayMore, a tool for language documentation productivity: From SIL International. *Language Documentation and Conservation* 8:66–74.

Pawley, Andrew, and Harald Hammarström. 2018. The Trans New Guinea family. In *The Languages and Linguistics of the New Guinea Area: A Comprehensive Guide*, ed. Bill Palmer, 21–195. Berlin: De Gruyter Mouton.

Pennington, Ryan. 2014. Producing time-aligned interlinear texts: Towards a SayMore–FLEx–ELAN workflow. Unpublished manuscript, SIL International.

Stanford, James N. 2009. Clan as a sociolinguistic variable: Three approaches to Sui clans. In *Variation in Indigenous Minority Languages*, ed. James N. Stanford and Dennis R. Preston, 463–484. Amsterdam: John Benjamins.

Z'graggen, John A. 1980. *A Comparative Word List of the Rai Coast Languages, Madang Province, Papua New Guinea*. Pacific Linguistics D 30. Canberra, Australia: Pacific Linguistics.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363