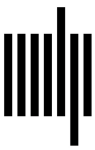
This is a section of doi:10.7551/mitpress/12200.001.0001

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

45 Managing Data for Theoretical Syntactic Study of Underdocumented Languages

Philip T. Duncan, Harold Torrence, Travis Major, and Jason Kandybowicz

1 Introduction: Project backdrop and methods of data collection

To shed light on helpful data management principles for theoretical syntax, this chapter draws from a recent and ongoing collaborative project to document two Indigenous Ghanaian languages, Ikpana (ISO 639-3: lgq) and Avatime (ISO 639-3: avn). Ikpana and Avatime are both underdocumented Ghana-Togo Mountain languages, spoken in the mountainous Volta region, in an area northwest of the regional capital Ho and east of Lake Volta. Data for this project were collected over a six-week period from July to August 2018, working with speakers in Logba Alakpeti, Amedzofe, Dzokpe, and Ho. Our eight-person research team divided into two groups, each dedicated to either Avatime (including authors Torrence and Major with Blake Lehmann and Kerri Devlin) or Ikpana (including authors Kandybowicz and Duncan with Bertille Baron Obi and Hironori Katsuda).[1] Given that extant work on Ikpana and Avatime remains limited, documenting general properties of the grammar of each language was a major focus of the project overall. To collect thematically unified data that would facilitate comparison between the two languages and give the project a more concrete direction, each group dedicated a majority of time to the documentation of interrogatives, which is what we highlight herein. Individuals also collected data for several interface topics (syntax-phonology, syntax-semantics, syntax-discourse, and so on), meaning that the data collection and management needed to be relevant and curated for robust linguistic inquiry across multiple subdomains.

Our methods of data collection included structured, direct elicitations, grammaticality judgments, and text collection across multiple genres (including short autobiographies, regional histories, and traditional *Ananse* [spider/trickster] stories, which are well-known in Ghana and other areas of West Africa, and still passed down among families there). For purposes of exemplification in this chapter, we primarily orient herein to principles underlying management of elicited material, though these do extend to our treatment of texts. The nature of our specific project—that is, non-Indigenous "outsider" (Ameka 2018) scholars working with Indigenous peoples and with Indigenous languages—informs aspects of data management that we feel are important to note but, given the purview of our chapter, are unable to elaborate on in detail. In particular, this includes taking appropriate steps to ensure that data use is transparent, aligned with language users' (and, when necessary, broader community) desires and expectations, and that data management and use are done in ways that promote Indigenous sovereignty, self-determination, and self-governance (see Holton, Leonard, & Pulsifer, chapter 4, this volume, and citations therein for discussion of data management issues specific to working with Indigenous peoples/nations/communities/families/individuals).

Our chapter is organized as follows. In section 2, we discuss aspects of our project pertaining to curating data, touching on themes such as data and file preparation practices that facilitate ease of interaction with the data along a typical syntax research pipeline, which involves cyclic integration of data collection, transcription, exploration, analysis, dissemination, and archiving. This highlights some of the more mechanical aspects of file preparation and management for the purposes of syntactic inquiry. Section 3 turns to conceptual and methodological issues. We reinforce what we see as a need for constant negotiation of descriptive and theoretical needs throughout data collection, which also has implications for data management (e.g., file curation, annotating transcriptions, and performing calls on data) in order to facilitate exploration of grammatical properties. Section 4 concludes.

Before proceeding, we would like to offer clarification on our use of *theoretical* in relation to syntax and fieldwork in this chapter. We recognize that *theoretical syntax* encompasses a plurality of diverse approaches, perspectives, and theoretical orientations. We intend the issues we discuss to be germane to theoretical syntax in the broad sense. This includes, but is by no means limited to, the generative framework in which we happen to operate (which itself is not singular). However, given that we exemplify various issues that arise in theoretical syntactic research with our own work, at times our discussion may be colored by peculiarities specific to one theoretical framework, that is, generativism (as among the many applications of theory that exist for syntax). By using the phrase *theoretical* in contrast to *non-theoretical*, then, we do not intend the misleading dichotomy of *generative* in contrast to *other approaches*.

2 Data management and project design

Underlying our data management was a file-oriented "database"[2] design informed by four interrelated principles: simplicity, ease of implementation, ease of deployment, and versatility. *Simplicity* here means that the interface, structure, and ability to interact with/access the data introduce as little complexity as possible and ensure that there is not-too-steep a learning curve. *Ease of implementation* for us means that the database could be constructed by leveraging existing technologies, architectures, and functionalities that are already present by default on personal computers. Relatedly, *ease of deployment* means that the database could be deployed seamlessly across multiple platforms (e.g., Windows, Mac, and Linux operating systems). Finally, *versatility* for our purposes was primarily directed toward output: we aimed to ensure that the initial formatting and data transformation would best prepare for dissemination (presentations and publications), as well as archiving (see Andreassen, chapter 7, this volume). Orienting to these principles is important in our project because it positions us to meet practical, ethical, and professional needs in ways that intersect with current best practices. We recognize from the outset that the approach we outline here is a bit ad-hoc and that ad-hoc solutions can provide challenges in linguistic data management. However, the reality of the current landscape is such that there is no single application or software that meets project needs universally. Even with a project-specific database, though, we stress the importance of forward- and multipurpose-thinking in design, so as to facilitate the production of accessible, replicable, and durative materials.

All files in the database followed consistent naming conventions with persistent file formats (see Mattern, chapter 5, this volume, for issues related to data sustainability and file naming). This included, for example:

- Audio recordings from elicitation sessions and text production/performances as uncompressed and lossless WAV files
- Video recordings of text production/performances as MPEG-4 files
- Scans of handwritten transcriptions and notes, either as PDF/A (which is specialized for archiving) or compressed and lossless PNG files
- Typed transcriptions (completed following elicitation sessions based on original handwritten notes and review of audio) as plain text files with UTF-8 encoding

Our file-naming conventions were based on "semantic file naming" (Thieberger & Berez 2011:103), incorporating the following information:

- Target language's ISO 639-3 code
- Date, according to international format (YYYYMMDD)
- Ordered letter, if more than one file for a single day
- Initials of language consultants, listed alphabetically
- Genre of the data (e.g., elicitation, text)
- Initials of linguists participating in the session, with "ALL" appended for group sessions with all researchers present
- File extension

For example, the file LGQ_20180723b_KA_NH_RD_elic_ALL.wav is an audio file from a group elicitation session with three Ikpana speakers on July 23, 2018. The aforementioned elements are also embedded in the metadata added to transcription files, along with additional information such as location, provenance, and explanations/definitions of percent sign (%) tags used.

Though these types of details about file formatting and naming conventions are nowadays fairly typical for language documentation projects, we feel that they are important to note here because they begin to form the basis for the file-oriented database we implemented in our project. To preserve the relationships between files, all associated files from a single data collection event would be placed

in a parent folder with the same semantic file name as the audio file (e.g., the folder LGQ_20180723b_KA_NH_RD_elic_ALL would minimally contain the .wav audio file, one or more .txt transcription files, and .pdf files of linguists' handwritten transcriptions and notes). Folders pertaining to data collection events are in turn contained in parent folders for their associated language, and these two language folders are one level under the folder for the entire project. This provides a by-language chronological organization that is easily navigable when individual files need to be reviewed, and the directory structure allows for querying the data, which is a simple but indispensable feature.

With respect to data entry within transcription files, we used the three-tier glossing system, but without numbering or additional formatting (for reasons we discuss shortly).[3] In lieu of an added numbering system, project members used a text editor with built-in line numbering, such as Notepad++ for Windows (https://notepad-plus-plus.org/), BBEdit (https://www.barebones.com/products/bbedit/) for Mac OS, or Atom for either environment (https://atom.io/). Morphological glosses follow the Leipzig Glossing Rules (https://www.eva.mpg.de/lingua/resources/glossing-rules.php) to the extent possible. To annotate examples for grammatical information, we used % tags, which were entered in the line immediately following the English gloss. These tier tags are initially derived from descriptive grammatical properties, and they can be further added to in order to annotate examples based on theory-specific points or domains of grammar outside of syntax proper (e.g., a pragmatic feature or intonational pattern). Thus, a typical numbered example would be split across four lines as seen in figure 45.1, which shows, from Ikpana, four lines associated with each of two constructions: a transitive sentence in lines 30–33 and an object *wh*-question in lines 35–38.

The built-in numbering is helpful in cases where group members need to quickly identify and discuss or review a particular example within a .txt file. The primary reason that we avoided additional document-internal numbering relates to the distinction "between the *form* of the presentation of the data and its *content*" (emphasis in original). In addition to the need to accommodate multiple platforms for doing data entry, we also wanted to ensure that the content would "derive many presentation forms" (Thieberger & Berez 2011:94). Among members of our project, this included creating documents (papers/articles, handouts) and presentations, with products in the Microsoft Suite, Google Suite, and LaTeX (the choice being either personal preference or a requirement, say, of a particular journal). Organizing the workflow in this way means that further formatting and editing is required, but this can be minimized through scripting. For example, we partnered with a colleague to create a simple Python script that could take plain text files as input and generate LaTeX-ready numbered examples (see Han, chapter 6, this volume, for detailed discussion of converting data into different forms). Running the script on an example like the first of the two seen in figure 45.1 can generate the following gb4e-style numbered example:

```
\beging{exe}
\ex
\gll Fafá o-kplò ìdzɔ́ afàn udântʃì(ɛ).\\
Fafa 3\textsc{sg}-cook yam home morning\\
‘Fafa fried yams at home this morning.’
\end{exe}
```

This reduces time taken to format and typeset examples (gb4e uses single spacing to left-align elements in the gloss), and it also provides a different way to call and interact with data. For example, an entire text file can

```
29
30  Fafá o-kplò i-dzɔ́ afàn udânʃì(ɛ).
31  Fafa 3SG-cook CL-yam home morning
32  `Fafa fried yams at home this morning.'
33  %transitive %declarative %mass noun %temporal adverbial %bare prepositional phrase
34
35  Mé Fafá o-kplò afàn udântʃì(ɛ)?
36  what Fafa 3SG-cook home morning
37  `What did Fafa cook at home this morning?'
38  %transitive %temporal adverbial %bare prepositional phrase %content question %object wh-question
39
```

Figure 45.1
Entering linguistic examples in a plain text editor (Atom shown) using a three-tier format and % tags.

be generated. Alternatively, the script can search across files and subfolders within a directory based on features such as % tags. A LaTeX engine can then easily create a searchable PDF for team or community members to use, because PDF readers are readily available, and such files can be opened in browsers if a PDF reader is not available. At this point in our project, though, we do not have as easy of a way to prepare linguistic examples for numbering in proprietary formats such as Microsoft Word, which for us still require manual formatting/typesetting.

The system we adopted for data management also can be integrated into workflows that require use of more specialized linguistic software and tools, including ones that are quite standard among documentary linguistics, such as ELAN (https://tla.mpi.nl/tools/tla-tools/elan/).[4] For example, some members of our project were interested in exploring issues related to the syntax-phonology interface. While the % tags in our system can be used to indicate features relevant to any domain of grammar associated with a particular entry (e.g., "%F0 rise on right edge" to mark specific intonation on a prosodic unit), ELAN is more suited to exploring phonetic and prosodic properties of sentences and utterances because it can be used to synchronize annotation tiers with A/V. For linguists who want to use ELAN in addition to adopting our system for data management, constructing a template within ELAN that exports into the format that we utilized is rather straightforward. This allows one to separate the large ELAN files from the text files, which is desirable when, for example, one is trying to work with a database, but does not need to access the audio.

Although as Dimitriadis and Musgrave (2009) note, the type of file-oriented database we've described has weaknesses, such as being less suited to highly "data-intensive" projects, we find that it has many advantages, such as rapid implementation, a drastically reduced learning curve, and ease of shareability of work product. Because the database leverages existing structures, this reduces computing power needed, and it also means that users interact with data in a way that is already familiar, which is helpful for researchers, as well as for language users who may wish to access the data. The database is also not operating system–dependent, which means that it can be easily deployed. These two components also allow for a means of interacting with data that is crucial in theoretical syntax, in particular: searching for, reviewing, and compiling examples based on grammatical features. In addition to building search features into scripting, a more basic way to achieve this is through the use of regular expressions (regex). In our project and database design, we can use regex to return files based on properties of a file name. For example, the regex (.txt and "elic") will return all elicitation files in the directory. Regex can also be performed on elements within a text file, by searching for, for example: a string in the target language, a particular morpheme in the gloss tier, or a particular grammatical feature listed in the % tags. The learning curve for using regex is not steep, and it does not require additional software (e.g., these can be entered into the Spotlight tool or Finder window in Mac OS, or a Windows folder). We also use additional % tiers to tag examples to simplify future regex queries that return examples based on more theory-oriented properties. For instance, if one is investigating *wh*-questions in a particular language or set of languages, there are a variety of properties that should be tracked throughout the data-entry process. Some examples include *wh*-in-situ, partial movement, full movement, arguments versus adjuncts. Elicitation sessions are often organized to investigate these particular issues, but this organization is lost when the entire database is considered. Tagging example sentences is a way of maintaining the content of the examples at a global level across the database (for instance, %FM+Q = full movement with question particle, %PM-Q = partial movement without a question particle, %IS+Q = in situ with question particle, and so on).

3 Data collection

3.1 Planning for data collection: Preparing for what is known and what is unknown

The basic challenge confronting the researcher in data collection for theoretical syntax is a familiar one that is common to scientific and/or scholarly investigation more broadly (Logan 2009). To borrow an infamously lampooned but surprisingly insightful comment from former US Secretary of Defense Donald Rumsfeld, "There are known knowns. . . . There are known unknowns. . . . But there are also unknown unknowns." That is, there are things we know, things that we don't know but know more or less about their existence, and things that we don't know, which we don't really know exist. Theoretical syntactic research—especially when working with an underdocumented Indigenous language—is by necessity

characterized by continuous navigation of these three components. This is partly because at the outset one often begins without necessarily knowing what will be theoretically interesting from the perspective of the language(s) investigated. This should not be taken to mean that one falls blindly into research. Instead, we advocate a cyclic, reflective approach, one that is based on using known knowns and known unknowns derived from theoretical and descriptive work to illuminate known unknowns and unknown unknowns. The idea is that over time the unknowns both become reduced and serve as the impetus for continued investigation. It is therefore simultaneously crucial to stay abreast of theoretical developments and their application to the language(s) of study and to understand the language(s) in theoretically neutral terms. The theory-neutral emphasis ensures that one's research is (1) accessible to non-linguists as well as linguists, (2) relevant to researchers of other frameworks, and (3) sound in description to be sound in theory (whatever theory).[5] In short, data collection for theoretical purposes begins with language description, just as data collection does for non-theoretical purposes.

The data collection process for this project began by consultation of prior work on Avatime and Ikpana. In this regard, the existence of excellent descriptive and analytical work (Ford 1971; Schuh 1995a, 1995b; Dorvlo 2008; van Putten 2014; Defina 2016), although they did not focus on *wh*-question formation, provided critical baseline data that allowed us to make much more progress than would have been possible otherwise.

We chose to look at *wh*-questions for reasons both theoretical and practical. On the theoretical side, some of us (Kandybowicz and Torrence) had already investigated *wh*-questions in the Akan group (Asante Twi, Bono, Wasa) and in Krachi, a North Guang language (Kandybowicz & Torrence 2013, 2015). These studies revealed some of the complexities of the construction in languages of the area. It seemed likely that Avatime and Ikpana would also share such complexity, which would be of theoretical interest. On the practical side, first, all languages have *wh*-questions, so we could be assured that we would find them in the languages. In addition, given the large theoretical and descriptive literature on *wh*-question formation, we could readily compare the Avatime and Ikpana data to what has been found in other languages. Finally, there was a limited time window in which the fieldwork could be carried out because of scheduling conflicts among team members. Typically, we would not go into a fieldwork situation on a "new" language with a theoretical agenda. This is because of the simple fact that it cannot be predicted which areas of the grammar will be of the most theoretical interest and one cannot know in advance where the evidence for theoretical analysis will be found. Related to this point, we basically agree with Dixon (2007:13), who states that a poor reason for doing fieldwork is to test or prove a theoretical point. If one knows nothing about the language, then we concur. However, if one knows a great deal about a language then testing a theoretical point is an excellent reason for conducting fieldwork!

3.2 Types of data and their use in theoretical syntax

3.2.1 Elicited material One way in which fieldwork for theoretical linguistics differs from classical descriptive fieldwork is that elicitation can play a central role in data collection for theoretically minded analysts. As a whole, our impression is that fieldworkers are divided on the role of elicitation, both within and across theoretical orientations. However, we take it as a given that the methods employed depend on the questions that one is trying to answer. If a theory makes a prediction about what stress pattern should be found or what word order should be ungrammatical, elicitation provides a straightforward way of testing the theory (although there are other, less immediately practical means). As Rice (2001:244) says, "given a particular theoretical claim, one should expect to find certain things are grammatical and other things are ungrammatical." Similarly, Aissen succinctly summarizes the importance of direct, controlled elicitation in theoretically oriented (in this case, generative) fieldwork by saying, "While nothing in generative linguistics excludes text collection, direct elicitation is unavoidable. The view of a generative grammar as a hypothesis about the internalized knowledge of the native speaker . . . makes central the classification of tokens as well-formed or ill-formed since the predictions of these hypotheses concern well- and ill-formedness. While well-formedness can be supported (though not established) by the existence of attested examples, ill-formedness cannot be supported by their absence" (1992:9).

Regarding "grammatical elicitation," Dixon enjoins fieldworkers that "such elicitation should play no role whatsoever in linguistic fieldwork." However, he goes on to say that "What I do is make up Jarawara sentences (that

are generated by the grammatical rules I am positing) and ask if these are bona fide utterances. . . . Or else I will quote some sentence that I know is alright (because I have heard it in a text or conversation) and ask about variants of it, perhaps changing the verb . . . or adding or subtracting an affix or a word. Consultants get the idea of what I am trying to do and either confirm that my made-up sentence is correct, or else offer an appropriate correction" (2007:23). Because this is almost exactly what we do in eliciting data for theoretical purposes, we take it that the differences in the role of elicitation among theoretical fieldwork approaches, and even in non-theoretical-oriented fieldwork, is one of degree, not quality. The only difference that we can readily discern in methodology is that, unlike Dixon, we see no problem in asking for translations from the contact language. If we imagine asking for the translation of a sentence from the contact language into the target language, Dixon correctly observes that the construction in the contact language may or may not literally correspond to a similar construction in the target language. However, this is a general property of translation, independent of fieldwork. The same issue arises, for example, if one were to ask for the Spanish translation of "I like puppets" and then assume that the subject "I" in English must be expressed as a subject in Spanish. With theoretically oriented fieldwork, analytical problems of this kind can probably be avoided entirely if one simply does not assume that *translations* must correspond to isomorphic structures across languages.

One advantage of working in a community of users of the target language is that the linguist may be able to work with users of different dialects, ages, and genders, for example. For theoretically oriented fieldwork, this is important because it may yield critical information about the range and sociolinguistic determinants of variation in the phenomenon under investigation. Expressing a similar view (specifically for syntax, but of more general applicability), den Dikken et al. (2007:339) note that "for the generative syntactician, the more informants [*sic*] you have, the more data from individual grammars you have, which gives you the potential to find micro-variants you might otherwise not have found (this is not unlike the general desire to study as many 'languages' as possible)." Certainly, in generative syntax, there is a long tradition of theoretical studies that makes crucial use of dialectal variation to argue for particular analyses (Bayer 1984; Henry 1995; Munaro 1997; Zanuttini 1997; Aboh 2004; Barbiers 2009, among others). The main point here though, is that it is that, by being embedded in the community, we were able to work with language users from a variety of social backgrounds. Thus, while we consider data collection from a single language user to be entirely legitimate (and we do this ourselves), while working in the Ikpana and Avatime communities we took it as imperative that we should seek out speakers from a variety of backgrounds to the extent we could. For descriptive/documentary purposes, this would yield a snapshot (although imperfect) of language variation in the community. Additionally, for our theoretical orientation, we hoped that this would result in a clearer idea of the factors that correlate with differences between individual grammars ("I-language[s]," Chomsky 1965, 1986).

3.2.1 Texts For this project, we also collected various genres of text, but the focus was on traditional stories. We also attempted, to the extent that we could, to find any previously written material that we could in either language. Given our focus on theoretical analysis, the role of text in theoretically oriented fieldwork is not the same as in other kinds of fieldwork. Unlike classical Boasian fieldwork (with its focus on text collection and the construction of a grammar and dictionary), the collection of negative data is necessary in theoretically oriented fieldwork, and ungrammatical data (with exceptions such as speech errors and stops and starts, and such) are not found in texts. This follows from the basic fact that native language users do not speak ungrammatically. This is a limit on the usefulness of texts. However, this does not mean that texts cannot be enormously useful in theoretically oriented fieldwork.

Concerning texts, Aissen (1992:9) observes that "while generative linguists may collect texts, publishing them has no place in the generative literature." Indeed, texts have not played a prominent role as data sources in the development of generative theoretical linguistics. However, fieldwork for theoretical purposes should include text collection, both audio and video if possible. For theoretically oriented fieldwork, texts are important because they are extremely rich sources of linguistic information. A text may reveal constructions that the linguist would never have thought to elicit explicitly. It is not surprising that there are limits to the usefulness of elicitation, just as with any other data source. In addition to "new" constructions, texts may also provide discourse contexts that license particular word orders, for example. Based

on the text, the theoretical linguist could then, based on the text, provide a felicitous discourse context to check sentence acceptability through direct elicitation, as suggested by Dixon. In our (i.e., authors Kandybowicz and Torrence) work on Krachi, we were able to obtain a copy of the New Testament book of Mark (GILLBT 2011). From this, we found a number of complex verbs that would have been practically impossible to find through direct elicitation. We then used these verbs in direct elicitation with the Krachi consultants. Probing the syntactic properties of these complex verbs proved critical in our analysis of predicate clefts in Krachi (Kandybowicz & Torrence 2015, 2016). For the present project, we were given a copy of the New Testament book of Mark. This text was useful because of the many examples it has involving question particles. These examples were then checked with the native speaker consultants, and we were able to get a clearer picture of the (very) complex distribution of these particles. The text also proved useful because the language is high register. The consultants would note that certain passages, while grammatical, sounded very formal and not at all like ordinary speech. This naturally led to elicitation about how one would express the Bible passage in "regular" speech, which revealed new constructions of theoretical interest. As part of the current project, we have collected audio and video texts of traditional folktales. It was intended that the great complexity of textual material could contribute to both the descriptive/documentary aspects of the project and the theoretical ambitions.

4 Conclusions

In this chapter, we discussed key principles of data management for doing theoretical syntax, including aspects related to database design, data collection and entry, and multipurpose data curation. We recommend that a database be simple, easy to implement, easy to deploy, and versatile. Attending to these interrelated principles enables high functionality and ensures that best practices can be followed—even with a highly project-specific implementation. In terms of practical and mechanical aspects, keeping in mind the distinction between data presentation form and content is essential. At the content level, tagging linguistic examples with grammatical features during data entry allows for performing calls on data and returning subsets bearing specific features. This, we feel, is an indispensable functionality, as it gives theoreticians the ability to interact with, probe, and explore data with theoretical issues and questions in mind. Methodologically, for data collection we advocate using an array of techniques, especially when a project requires more open-ended exploration of linguistic properties as a precursor to theoretical analysis. Finally, a major theme that we highlighted herein is that data collection for theoretical purposes (regardless of theoretical orientation) begins with language description, just as data collection does for non-theoretical purposes. Adopting a descriptive stance while doing theoretically oriented work has implications across the cycle of data collection and management and provides a solid foundation for robust theoretical inquiry.

Notes

1. On behalf of all the members of our project, we would like to express heartfelt gratitude to the speakers we have had the privilege of working with: Mary Akum, Kwame Amedzro, Vivian Anka, Edward Antwi, Raymond Dzakpo, Nelson Howusu, Ogordor, Gifty Amu, Peace Awunyama, Akos Mawulorm, Vincent Azafokpe, Wisdom Ekissi, Kwame Jones, Philomena Kumatse, Paul Kwawu, and Agbenya Wisdom. We also thank two anonymous reviewers and the volume editors for many helpful insights. This research was supported by a grant from the National Science Foundation (BCS EAGER DEL—1748590), which we gratefully acknowledge.

2. We recognize that *database* is a general term that encompasses a wide spectrum of structured data collections. For the purposes of this chapter, we use *database* primarily to refer to a structured file system residing on a local computer or cloud-based web application.

3. Additionally, we used the IPA as the basis for our transcription, and not, for example, an Akan-based romanization that is regionally prevalent and typically used for rendering Ghanaian languages orthographically.

4. From the Max Planck Institute for Psycholinguistics, the Language Archive, Nijmegen, the Netherlands (see Brugman & Russel 2004, among others).

5. Ultimately, then, the goal is to be sound in description *and* sound in theory. Here, we intend to highlight the importance of a good descriptive foundation for theoretical inquiry and analyses.

References

Aboh, Enoch. 2004. *The Morphosyntax of Complement–Head Sequences: Clause Structure and Word Order Patterns in Kwa*. New York: Oxford University Press.

Aissen, Judith L. 1992. Fieldwork and linguistic theory. In *International Encyclopedia of Linguistics*, vol. 2, ed. William Bright, 9–11. New York: Oxford University Press.

Ameka, Felix K. 2018. From comparative descriptive linguistic fieldwork to documentary linguistic fieldwork in Ghana. In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. B. McDonnell, A. L. Berez-Kroeker, and G. Holton, 224–239. Honolulu: University of Hawai'i Press.

Barbiers, Sjef. 2009. Locus and limits of syntactic variation. *Lingua* 119 (11): 1607–1623.

Bayer, Josef. 1984. COMP in Bavarian syntax. *Linguistic Review* 3 (3): 209–274.

Brugman, Hennie, and Albert Russel. 2004. Annotating multimedia/multi-modal resources with ELAN. In *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*, ed. M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, and R. Silva, 2065–2068. Paris: European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Defina, Rebecca. 2016. Events in language and thought: The case of serial verb constructions in Avatime. PhD dissertation, Radboud University.

den Dikken, Marcel, Judy B. Bernstein, Christina Tortora, and Raffaella Zanuttini. 2007. Data and grammar: Means and individuals. *Theoretical Linguistics* 33 (3): 335–352.

Dimitriadis, Alexis, and Simon Musgrave. 2009. Designing linguistic databases: A primer for linguists. In *The Use of Databases in Cross-Linguistic Studies*, ed. M. Everaert, S. Musgrave, and A. Dimitriadis, 13–75. Berlin: Mouton de Gruyter.

Dixon, R. M. W. 2007. Field linguistics: A minor manual. *Sprachtypologie und Universalienforschung* 60 (1): 12–31.

Dorvlo, Kofi. 2008. A grammar of Logba (Ikpana). PhD dissertation, Leiden University, Landelijke Onderzoekschool Taalwetenschap Dissertation Series. Utrecht, the Netherlands.

Ford, Kevin. 1971. Aspects of Avatime syntax. PhD dissertation, University of Ghana.

Ghana Institute of Linguistics, Literacy, and Bible Translation (GILLBT). 2011. *Maakı ɛ Kyıkyɛı*. Tamale, Ghana: GILLBT.

Henry, Alison. 1995. *Belfast English and Standard English: Dialect Variation and Parameter Setting*. New York: Oxford University Press.

Kandybowicz, Jason, and Harold Torrence. 2013. Comparative Tano interrogative syntax: The view from Krachi and Bono. In *Selected Proceedings of the 43rd Annual Conference on African Linguistics*, ed. Olanike Ola Orie and Karen W. Sanders, 222–234. Somerville, MA: Cascadilla Press.

Kandybowicz, Jason, and Harold Torrence. 2015. Wh-question formation in Krachi. *Journal of African Languages and Linguistics* 36 (2): 253–286.

Kandybowicz, Jason, and Harold Torrence. 2016. Predicate focus in Krachi: 2 probes, 1 goal, 3 PFs. In *Proceedings of the 33rd West Coast Conference on Formal Linguistics*, ed. Kyeong-min Kim, Pocholo Umbal, Trevor Block, Queenie Chan, Tanie Cheng, Kelli Finney, Mara Katz, Sophie Nickel-Thomspon, and Lisa Shortern, 227–236. Somerville, MA: Cascadilla Press.

Logan, David C. 2009. Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry. *Journal of Experimental Botany* 60 (3): 712–714.

Munaro, Nicola. 1997. *Proprietà strutturali e distribuzionali dei sintagmi interrogativi in alcuni dialetti italiani settentrionali*. PhD dissertation, University of Padua.

Rice, Keren. 2001. Learning as one goes. In *Linguistic Fieldwork*, ed. Paul Newman and Martha Ratliff, 230–249. New York: Cambridge University Press.

Schuh, Russel G. 1995a. Aspects of Avatime phonology. *Studies in African Linguistics* 24 (1): 31–67.

Schuh, Russel G. 1995b. Avatime noun classes and concord. *Studies in African Linguistics* 24 (2): 123–149.

Thieberger, Nicholas, and Andrea L. Berez. 2011. Linguistic data management. In *The Oxford Handbook of Linguistic Fieldwork*, ed. N. Thieberger, 90–118. Oxford: Oxford University Press.

van Putten, Saskia. 2014. Information structure in Avatime. PhD dissertation, Max Planck Institute for Psycholinguistics.

Zanuttini, Raffaella. 1997. *Negation and Clause Structure: A Comparative Study of Romance Languages*. New York: Oxford University Press.

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