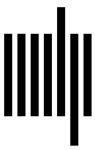
The Open Handbook of Linguistic Data Management

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

24 Managing Lexicography Data: A Practical, Principled Approach Using FLEx (FieldWorks Language Explorer)

Christine Beier and Lev Michael

> Wherever I turned my view,
> there was perplexity to be disentangled,
> and confusion to be regulated;
> choice was to be made out of boundless variety . . .
> —Samuel Johnson, 1755, preface to *A Dictionary of the English Language*

1 Introduction

In this chapter, we describe a methodology and workflow for developing lexical resources for underdocumented languages in the context of language documentation projects dedicated to one or both of the following goals: (1) to create and distribute a *dictionary* to a user community; and (2) to create a multipurpose *extensible lexical resource* that forms an integral part of a language documentation and is interdependent with other components of the project, including a text corpus and grammatical analyses. In particular, we describe a workflow that makes use of FieldWorks Language Explorer (FLEx),[1] a lexical and text corpus database application, together with an XML[2]-to-(Xe)LaTeX[3] Python[4] script, from which one can produce professional-quality typeset PDF files for paper or digital publications. All the software and applications we discuss are open source and/or free to obtain and use and have been stably supported for decades. In addition, we describe the methodology we have developed over more than twenty years of language documentation and description in Peruvian Amazonia that addresses concerns about both data *sharing* and data *validity* in the context of the lexicographical practicalities of documentation projects focused on underdocumented languages.

2 Lexicography and language documentation

Lexical resources occupy a prominent place in modern language documentation. A dictionary, for example, is often one of the outcomes of linguistic research that is most valued by the members of heritage communities of languages that are shifting. Being responsive to community priorities thus often means emphasizing the early development of a variety of lexical resources. Lexicographic work is also indispensable to the documentation and description of underdocumented languages, forming part of the classic Boasian trilogy, together with a grammar and a translated text corpus (Woodbury 2011). The methodology and workflow we describe in this chapter reflect our commitment to both long-standing motivations for lexicographic work and the influence of the principles and best practices of digital data management and preservation that have emerged in recent decades (see, e.g., Bird & Simons 2003; Bowern 2008).

As Nichols and Sprouse (2003:99) observe, "producing easily usable, professional-looking descriptive dictionaries on a shoestring budget in a short time span is a priority for documentation, but hard to achieve." The magnitude of the challenge can be appreciated by considering the technical and ethical desiderata that many language documentation projects face in carrying out lexicographic work. In the *short term*, a methodology and workflow are needed that will generate outputs that are useful and appropriate to host community collaborators on the one hand, and to linguistic scholarship on the other, in as short a time frame as possible, at as low a cost as possible. This methodology and workflow design, must, in the *medium term*, anticipate expansion, refinement, and adaptation of the project's structures and products as the project evolves over time. And finally, in the *long term,* the methodology and workflow need to be successful in producing data, analyses, infrastructures, and concrete products that are durable, secure, sharable, reusable, and archivable; while community-directed materials should be easy to access and easy to reproduce, both digitally and on paper, without the need for continual expert maintenance.

The methodology and workflow desiderata we've outlined must also be responsive to the practical realities that face many linguists' host community collaborators—especially their political and economic marginalization, often compounded by their geographical distance from urban political centers; lack of access to educational opportunities and corresponding low levels of literacy; lack of access to the Internet and other digital resources; and their desire for straightforward, easy-to-use language-learning materials. Here, these considerations are as important to project design and output as are the desiderata responsive to academic and scholarly practices.

These high-level desiderata also entail requirements for the digital tools we employ for lexicography in language documentation projects. First, the lexical database must have flexibility in organizing, categorizing, annotating, coding, sorting, and searching data. Second, our digital tools must have the autonomy, stability, and portability necessary for being able to work in situ and offline. Third, the system must be as simple and flexible as possible for the production of different types of print and digital outputs that themselves are portable and inexpensive. And finally, the digital tools must be as low cost or freely available as possible, so that neither access nor longevity constitute obstacles to sharing or preservation. The reader with a targeted interest in the technical aspects of our discussion may wish to skip ahead to sections 4.2 and 7.

3 Background

The lexicographic methodology we describe here, and the priorities that motivate it, have emerged dynamically since 1999 when we first began doing lexicographic work with various Amazonian communities, most of whom have experienced dramatic language shift toward Spanish over the last few generations.[5] In this context, our lexicographic objectives have been:

1. To gather as much *target language* lexical data as possible,
   1a. with glosses and definitions in the *local* (i.e., *contact*) *language* (typically Spanish, but also Quichua in two cases),
   1b. and with equivalent plus additional descriptive and analytical material in English;
2. then to organize and interpret that data in order to produce:
   2a. free, accessible, photocopiable print resources, generated from archivable, web-accessible PDFs, that we deliver rapidly to our local collaborators and other interested members of the heritage community;
   2b. and archivable results for scholars and unknown future users.

In practical terms, how we have operationalized these objectives has been shaped, in each project, by a core set of factors that we address briefly: project time frame (section 3.1) and scale (section 3.2); scope of dictionary content (section 3.3); and teamwork (section 3.4) and data sharing (section 3.5) as organizing principles.

3.1 Project time frames

In some cases, our projects have had specifically delimited time frames for fieldwork and production of the lexical resources, as in the case of our "rapid" projects with the Andoa, Muniche, and Záparo communities, where we were limited to a single in situ fieldwork visit (typically about two months) and an additional calendar year (sometimes less) for the creation of lexical (and other) resources. In other cases, the projects have been more open-ended, as in the case of our work with the Iquito[6] and Máíhùnà[7] communities, which remain ongoing. Even in these more open-ended projects, however, we have prioritized producing versions of lexical resources promptly and on a regular basis for community members and other project participants.

3.2 Project scale

The scale of each project that we have carried out has had a significant impact on the kinds of lexical resources we have sought to develop, as has the user community that we have sought to serve. Because most of our lexicographic work has been with "target" languages that no longer have active speech/user communities, but whose inheritors expected our fieldwork with them to result in materials that give them access to knowledge of and about the language and its history, we have chosen to make bilingual or trilingual dictionaries, with headwords in the target language and descriptive/definitional content in the contact language(s) that have replaced the target language in its heritage community. In some circumstances, we have opted for an alphabetically ordered dictionary only (e.g., the Iquito and Máíhɨ̃kì

dictionaries), while in other cases, we have included thematically organized content as well (e.g., the Muniche and Záparo dictionaries; see also Mosel 2011). In some cases, we have also produced pamphlets focused on particular semantic domains for pedagogical purposes, to complement the alphabetically ordered dictionary. As we discuss in detail in sections 7.2 and 8, the tools that are built into FLEx for organizing, categorizing, labeling, and sorting lexicographic data, as well as their extensibility, are indispensable to creating these varied types of language-specific outputs.

3.3 Scope of dictionary content

For the larger dictionaries we have developed, an important decision that we have made is to include as much "encyclopedic" information as we can—cultural, historical, and ecological—precisely because there is no other obvious place in which to systematically document, organize, circulate, and preserve this type of knowledge, which is of great interest to the communities with whom we work and, we hope, to future scholars. For the same reason, we have attempted to maintain high standards when it comes to defining ethnobiological terminology, aiming to provide descriptions and scientific names to the best of our abilities. We have thus avoided relatively uninformative definitions such as "bird species" or "type of tree" (unless that is literally all the local knowledge that remains of the lexeme). This encyclopedic goal, and the resulting data management requirements for such an internally complex data set, is a key reason that the preparation of certain of our dictionaries has been so time-consuming, but we have felt that a commitment to this scope is demanded by the circumstances of linguistic and cultural endangerment in which we tend to work. A more nuanced discussion of sources for lexical data is offered in section 5; relevant methodological strategies are offered in sections 8 and 10.

In sum, our typical output is a bilingual, bidirectional, quasi-encyclopedic PDF-based print dictionary that is designed to (1) describe the lexicon of the language as comprehensively as possible and (2) serve as an organized repository of diverse cultural knowledge associated with the language.

3.4 Teamwork as an organizing principle

Turning to methodological specificities, because all of our field projects have involved teamwork—minimally, the two of us, but in some projects, a team of six to nine collaborating researchers—we have been driven by necessity to develop methods and workflows that permit shared use and shared management of a lexical database and related resources. In this respect, we have found that the two most important elements of successful collaborative methodologies are, first, developing and documenting explicit detailed procedures, workflows, and standards that all team members use and, second, systematizing the means for adding metadata to the lexical database to ensure that there is a chronological record of the developments and changes made by identified participants.[8] For this approach to work, accommodating multiple collaborators and multiple project periods (e.g., fieldwork seasons), we have found that users of the lexical database must be committed to creating metadata and documenting work history at every step along the way; we explore this matter in greater detail in section 11.

3.5 Open linguistic data and sharing as an organizing principle

In every project, an important organizing principle has been to return usable outputs of our fieldwork to our collaborators and host community at the end of every field season. In our experience, doing this is a demonstration of respect and commitment that has a long-lasting positive impact on everyone involved. In practice, what this has meant is creating lexicographic materials that are designed for being harvested and shared in this time frame. For longer-term projects, this has meant producing "draft" dictionaries, clearly labeled as such, that can be exported, printed, circulated, and corrected by local collaborators and community members and which serve as good-faith down payments on the eventual production of a non-draft version.

A parallel commitment to sharing data and outputs with other scholars has led us to make durable digital materials available, including drafts of dictionaries, as soon as possible, through our own web sites and/or through digital archives, including, at different points, the Endangered Languages Archive,[9] the Archive of the Indigenous Languages of Latin America,[10] and the California Language Archive[11] (see also Buszard-Welcher, chapter 10, this volume).

4 Starting out on a lexicography project

4.1 General considerations for lexicography work

Having sketched out the frame around our own work, we now broaden our view to some general considerations. In beginning any lexicography project, we are embarking on a tremendously intricate and time-consuming undertaking (see Frawley, Hill, & Munro 2002a for an excellent summary), and we must calibrate our objectives to the temporal, material, financial, and human resources available.

We start by addressing fundamental questions asked by all lexicographers, including: Who is our audience? What is our time frame? What is the intended scope of our project? What are the intended outputs? What are the ethical entailments of the project? What will we do to ensure the long-term storage, protection, and dissemination of data and outputs? What will be the internal composition of, and division of labor among, the lexicography team? We think carefully about orthography design and the challenges of standardization. We consult with local collaborators about orthography, as well as ethics and permissions (see also Holton, Leonard, & Pulsifer, chapter 4, this volume; Collister, chapter 9, this volume). We consider whether, and how, to handle multiple media within, or associated with, the lexical database and whether the project will have an online component. Finally, we contemplate how to determine when the lexical database is "good enough" to publish in dictionary form and when and how to archive (all or part of) the data set (see also Andreassen, chapter 7, this volume; Buzsard-Welcher, chapter 10, this volume). To explore these and related issues further, we refer readers to resources on basic lexicography theory and practice (e.g., Atkins & Rundell 2008; Landau 2001; Zgusta 1971, among many others, as well as the *International Journal of Lexicography*) and to resources that specifically address lexicographic work in the context of language documentation and preservation (e.g., Frawley, Hill, & Munro 2002b; Mosel 2011; Haviland 2006; Ogilvie 2011; Rice 2018).

4.2 Overview of our lexicographic methodology and workflow

Returning our focus now to the specific lexicographic methodology presented in this chapter, the key sequential (and in some cases iterative) steps that guide our workflow are laid out in table 24.1; subsequent sections explicate the content summarized in the table.

5 Sources of lexical data

Clearly, fundamental to any good dictionary are broad, diverse, and nuanced lexical data. In our projects, we use a complementary set of methods for data collection, five of which we briefly discuss here.

5.1 Lexical data from parsed texts

Audio- or video-recorded texts that are transcribed,[12] translated, and subsequently parsed using FLEx (see section 7.2) are an obvious and excellent foundation for a FLEx lexical database. Not only do parsed texts yield glossed roots and stems that serve as the kernel of entries in the short term, but also, in the long term, this corpus provides numerous examples of the use of more common roots and stems, helping you to refine the definition and senses of the lexemes in question. And just as texts are invaluable for the investigation of grammatical phenomena because they manifest phenomena that are either difficult to elicit or that may not occur to the investigator to elicit, texts are also a rich source of lexical data that are either difficult to elicit or are not obvious candidates for lexicalization. Despite the great value of texts, however, it is important to emphasize that they are far from a panacea for lexicographic work. First, many lexemes of a language simply do not appear in corpora of the size typically developed in language documentation projects focused on underdocumented languages. This reflects the fact that low-frequency lexemes surface only sporadically in such corpora. For example, fewer than 25% of the large number of Iquito ethnobiological terms in our database actually surface in our quite extensive Iquito corpus. Similarly, we have found that less common senses of even high-frequency lexemes are often not represented in texts. To address these weaknesses, thoughtful and careful elicitation is indispensable.

5.2 Lexical data from stimulus word lists

Without a doubt, stimulus word lists that provide contact-language lexical items to be translated into the target language can be useful tools for quickly acquiring a large number of lexemes. Great care must be taken with this approach, however. Word lists of this sort suffer from

Table 24.1
Guide to workflow for our lexicographic methodology

Sequential step (i = iterative)	Type of work • Done by whom	Core tasks	Main discussion section(s)
1. Create a new project in FLEx	FLEx work • Linguist	• Set up orthography • Select built-in fields to structure your database • Create custom fields as needed per project	8
2(i). Document lexicon	Fieldwork • Linguist(s) with language users	• Gather lexical data	3.3, 5
3(i). Create FLEx Entries	FLEx work • Linguist(s)	• Start Entry History • Write definitions and other text using existing fields	7, 9
4(i). Develop FLEx Entries	FLEx work • Linguist(s) and consultants	• Expand, refine • Check, correct • Exemplify	10
5(i). Categorize FLEx Entries	FLEx work • Linguist(s)	• Populate existing fields • Add custom fields based on your analytical needs and output types	8
6. Export XML version of database	FLEx work • Linguist	• Export LIFT* file of lexicon	7.2
7. Create/obtain script, adapt it to project and output	Python work • Linguist and programmer	• Script structure will depend on desired content and form of output(s).	7.3
8. Run script	Python work • Programmer	• Process LIFT file to generate TeX file	7.3
9. Prepare TEX file	LaTeX work • Linguist and programmer	• Create/obtain preamble • Create/obtain mappings from XML-based tags to LaTeX commands	7.4
10(i). Create and share PDF	LaTeX work • Linguist	• Typeset TeX file • Disseminate PDFs of successive versions and types of dictionary	2, 3.5

* LIFT (Lexicon Interchange FormaT) is an XML format for storing lexical information.

two major types of weakness. First, they are typically based on whichever concepts are lexicalized in the contact language, and these may exhibit quite different patterns of lexicalization than those of the target language. This mismatch can be frustrating for both linguist and consultant in elicitation contexts. Second, the cultural and physical contexts presupposed by the word list may be quite different from those relevant to contemporary users of the target language, and/or relevant historically for the society in which the target language was used vitally, in the case of highly endangered languages. In such cases, the word list may be replete with concepts not lexicalized in the target language, while simultaneously lacking numerous important concepts that *are* lexicalized in the target language. In some regions of the world, scholars have developed areal or regional lexical elicitation lists that mitigate these problems to some degree.

For these reasons, we generally do not recommend starting lexical data collection through elicitation using word list stimuli. However, such lists can be very useful when employed strategically to identify gaps in lexical data collection, once the researchers who are compiling the lexical data have developed a familiarity with the patterns of lexicalization in the language and with the cultural and physical environments.

5.3 Lexical data from stimulus activities

Activity-based elicitation sessions (audio-/video-recorded or not) are another excellent source of lexical data. Concretely, for example, the lexicographer can accompany someone when they carry out an activity, such as going fishing, doing laundry, cooking, harvesting from a garden, going shopping, fixing a motor, or caring for a child, and ask a range questions about all the objects and actions that emerge: "What is that?" "What are you (is he/she) doing?" "Why are you (is he/she) doing that?" Alternatively, it may be the case that it is wisest to simply record words for future focused consultant work, rather than interrupting the activity.

5.4 Systematic exploration of semantic domains

In semantically cohesive and well-defined domains such as kinship terms, color terms, body parts, ethnobiology (mammals, birds, plants, and so on), and actions of particular types, such as cutting and breaking, it is fruitful to build up lexical data by exploring the semantic domain in a systematic way. This type of systematic exploration benefits substantially from specific methodologies and stimuli (such as using diagrams, color chips, ethnobiological publications, and the like), but some early progress can be made with *free listing*, that is, asking consultants to recall as many terms in a given semantic domain as possible.

5.5 Participant observation of language in use

Developing your conversational abilities in the target language and then participating in regular conversations, either as an interlocutor or as an accepted overhearer, is another excellent way of identifying lexemes that have not yet been documented via other methods, as well as for developing insights into the definitions of, or new senses of, lexemes that you have already recorded. For example, a few days before departing the Iquito community last month, in the flow of a conversation with us about a loud party near her house, one of our consultants used a word for "loud," *ihɨrana*, that we had never previously encountered. This word had never appeared in a text, and when we had explored the concept "loud" in prior elicitation sessions, we had only ever obtained the word *amátanana*, which has a broader range of meanings that include "strong" and "fast." But sure enough, when we asked another consultant about *ihɨrana,* he instantly recognized it and confirmed it as meaning "loud," while confirming that *amátanana* is used more commonly.

6 Structuring the dictionary as a whole

A common basic structure for a dictionary consists of either three or four main parts. The first part, the *front matter*, contextualizes the dictionary; describes the project that produced it; and identifies and thanks all participants, supporters, and funders. It also describes how to use the dictionary and provides essential grammatical information.[13] The second part, the *main body*, presents the lexical content of the dictionary, ordered alphabetically by target language headwords. The optional third part, the *thematic content*, presents the same lexical content as the main body but organized by semantic fields or themes. The final part provides the *reversals* from contact language(s) to the target language. In our workflow, each of these second, third, and fourth parts is the output of a different subset of our lexical database content, which is ordered, formatted, and typeset using specifications that are made possible by using LaTeX (see sections 7 and 8). Regardless of the structure you choose, note that it is essential to provide, before the main body, a detailed description of *how to read an entry*, along with a description of how to interpret the set of categories, labels, and abbreviations that you have used in your different types of entry.

7 Choosing digital tools for a lexicography project

In this era, we would not undertake lexicographic work without making use of flexible digital tools for data management and the production of outputs such as dictionaries. Here we describe the digital tools that we currently employ in our workflow.

7.1 A bird's-eye view

We use FLEx to house and organize our lexical data and the combination of Python and LaTeX to produce attractive, well-formatted outputs. Our use of Python and LaTeX is a response to the limited output options for FLEx. After trying out several variants, we settled on a digital pipeline that takes the native XML output from FLEx and converts it to (Xe)LaTeX markup using a Python script. The first version of this pipeline was

created in 2012 by Máíh̃ĩkì Project team members Greg Finley and Stephanie Farmer, and it served to produce the 2012, 2013, and 2014 drafts of the *Diccionario Bilingüe Máíj̃ĩkì—Castellano y Castellano—Máíj̃ĩkì* (Michael et al. 2013). The results were so satisfying that we have used updated and customized versions of this pipeline for all our FLEx-based dictionary outputs since then, including the *Diccionario Trilingüe Záparo* (Beier et al. 2014) and the dictionaries published by the Iquito Language Documentation Project (Beier et al. 2019; Michael et al. 2019). For the Iquito Language Documentation Project dictionaries, the Python script has been significantly reworked by Ronald Sprouse to handle the more complex database and dictionary organization required.[14]

7.2 FLEx

FLEx is a free, open source application, developed and made available by SIL International. FLEx is designed to create linked text corpus and lexical databases, where the lexical database can serve to parse texts in the corpus, and parsed texts can serve as one of the sources of data for the lexical database. The pros and cons of FLEx are laid out in table 24.2.

7.3 Python

Python is a high-level programming language widely used in academia for scientific programming. The pros and cons of Python are laid out in table 24.3.

7.4 LaTeX and XeLaTeX

LaTeX is a free and widely used document preparation and typesetting system that produces PDF documents of high typographical quality; XeLaTeX is an extension to LaTeX that accommodates Unicode in its native TeX files, thereby accommodating a vast range of characters. The pros and cons of LaTeX are laid out in table 24.4.

8 Structuring FLEx lexical database records

Regardless of which lexical database tools you use, the most consequential data management task you will face is the structuring of the *Entry*, by which we mean the unique database record that corresponds to a single headword. Since there are good resources available that discuss the essentials and logic of dictionary entry creation in general (e.g., Atkins & Rundell 2008; Landau 2001), we will note here only the most important elements of structure that

Table 24.2
FLEx: Pros and cons

Pro	Con
Easy to obtain and install; project backup is also quick and easy.	Relatively poor documentation makes it difficult to learn, use, fully exploit, or troubleshoot.
Backed by significant institutional infrastructure and commitment from SIL International, therefore likely to remain a viable tool for the foreseeable future.[15]	Native to PC and Linux, but not to Mac OSX (we have had mixed success using PC and Linux emulators on Mac OSX devices) and requires substantial RAM to run smoothly.
Accommodates Unicode characters—a crucial feature for orthographies that incorporate non-ASCII characters and diacritics.	The user has relatively limited control over the types and design of outputs.
Widely used among linguists, therefore it is relatively easy to share projects with colleagues, as well as to obtain guidance and advice from other linguists who use FLEx for similar purposes.	Has a relatively small online user community for purposes of sharing knowledge and experiences, or for troubleshooting.
Different versions of the database can be synchronized online via the built-in Language Depot tool, or by synchronizing copies on storage devices such as USB drives, which facilitates collaboration across multiple computers.	SIL International is a Christian missionary organization and the ethical considerations of having one's professional activities crucially dependent on such an organization are significant (Dobrin 2009; Epps & Ladley 2009 concerning Amazonia in particular).
The database can be exported in a variety of formats, including XHTML and XML, which allows repurposing and manipulation of the data using other tools.	
Has the capacity to add custom fields, as well as to bulk edit and merge fields, allowing for flexibility as the database evolves with use.	

Table 24.3
Python: Pros and cons

Pro	Con
Free, flexible, expandable, adaptable, powerful, and well-supported.	Requires a relatively high level of programming expertise to develop the script and employ it for the FLEx-to-LaTeX pipeline described here.
It is relatively easy to find people with Python programming experience.	The script employed for the FLEx-to-LaTeX pipeline must be tailored to the specific FLEx fields one wishes to typeset in LaTeX.
Python scripts developed for handing FLEx exports (e.g., XML) can be adapted from one project to another.	

Table 24.4
LaTeX: Pros and cons

Pro	Con
Free and easily accessible.	Requires learning its markup language, which involves a learning curve.
Very stable—in two different senses. First, whereas the files of applications such as Word can become quite unstable when they are both large and incorporate complicated formatting, LaTeX files never do. Second, whereas commercially popular applications change their features and encoding with some regularity, resulting in the permanent loss of formatting in older files, LaTeX files made decades ago are as readable now as they were when they were made.	Many publishers don't accept LaTeX files (although many academic publishers do); in such cases, only the PDFs it produces could be submitted.
Very powerful, well-designed, and seemingly limitless tool for producing publishable documents on one's own time frame, using in-house labor. Basic functions, plus a vast set of extensions (packages) satisfy a plentitude of linguists' specific needs, such as interlinearization, flexible example numbering, and tree diagrams.	
LaTeX and its variants have a huge and active user community that helps tremendously in learning and troubleshooting.	

we have come to rely on in the type of work we do using FLEx.

8.1 Basic components of an Entry

The default FLEx lexicon record, called an Entry, includes a very large number of built-in fields, of two types: fields that pertain to an Entry as a whole, and fields that pertain to Senses within an Entry. To a large degree, the earliest stages of developing the structure of FLEx database records involves selecting which fields to employ and which ones to suppress. Many of these choices will be obvious, based on a general familiarity with dictionaries (e.g., creating citation forms, parts of speech, and definitions), but others are perhaps less so. We will discuss some important fields of the latter type, and considerations involved in their use—but first, we offer two overarching observations.

First, it has been our experience that as our knowledge of the grammar of a language has deepened, we have discovered new aspects of lexical or grammatical irregularity, or relationships between lexemes, that merit documentation. In the Iquito case, for example, we discovered a large number of irregular plurals, morphologically conditioned root allomorphy, and active/middle verb pairs related by ablaut and consonant mutation, for all of which we eventually created dedicated custom fields in FLEx. In short, it is not possible to know at the outset all the fields that you will want or need, and it is helpful to accept from the outset that you'll continue to develop the structure of your Entries as your lexical documentation advances.

Second, in this context, a powerful advantage of the FLEx-Python-LaTeX pipeline described here is that the fields that make up a FLEx Entry can, to a large degree, be reordered at will via the Python script, in so far as the final output document is concerned. We have even taken advantage of this capability to display, in different parts of the dictionary entries, subtypes of a single FLEx field (e.g., the Variants field, discussed in section 8.2), in effect splitting a single FLEx field into distinct dictionary entry fields. In short, there is much about the structure of your output dictionary entry that can be reworked via the pipeline described here.

The single aspect of FLEx record structure that cannot be so easily finessed in this way is the distinction between fields linked to the Entry as a whole (e.g., Citation Form), and fields linked to particular Senses (e.g.,

Gloss and Definition). This is arguably most significant for the various types of note fields (see section 7.3), some of which are Entry-level fields and others of which are Sense-level fields. In this light, it is important to contemplate whether the note field you want to use supplies information that is relevant at the level of the Entry as a whole or at the level of an individual Sense associated with the lexeme.

Now let us turn our attention to several specific field types that are extremely useful in developing the structure of a FLEx Entry.

8.2 Variant forms

In FLEx, it possible to establish a Variant relationship between distinct Entries, such that one Entry is treated as the Main Entry, and the other as a Minor Entry that points back to the Main Entry. These Variant relationships are extremely useful for addressing both sociolinguistic variation and dialectal diversity. Significantly, it is simple to create customized Variant types by which to categorize your Variants. For example, a particular variant can be identified as pertaining to a specific dialect or as a "playful variant" or an "archaic form" of the Main Entry. For our Iquito dictionary, we also use the Variant relationship for a number of irregular form types, such as irregular third-person possessives, irregular imperfective roots, and irregular plurals. Importantly, because each Variant type is directly associated with the FLEx Entry in the XML that we export from FLEx, it is possible to customize both where and how these Variants appear in the associated dictionary entry, resulting in considerable power and flexibility in organizing the structure of the dictionary entry in the final output.

8.3 Built-in Note fields

The types of note fields built into FLEx—including Grammar Note, Semantic Note, and Anthropology Note, among many others—are excellent tools for organizing and categorizing different types of descriptive information and data (see section 5) related to an Entry. Using these various note fields means that, at the point of generating a dictionary, you are able to choose which of them will appear in the output. Similarly, if it turns out that you have overdifferentiated, it is possible to merge the contents of different note fields together using FLEx's Bulk Edit function.

8.4 Semantic domains

If one of your objectives is to create a dictionary resource that is organized thematically, the Semantic Domain tool in FLEx is indispensable. A highly elaborate tree of Semantic Domains is built into FLEx, but we have preferred to set up our own custom Semantic Domain categories, based on the type and quantity of data we have for a given project. (For example, for Iquito, it serves us well to have as distinct domains the following: Plants, Plants: medicinal, Plants: parts, Body parts, and Body parts: human.)

8.5 Custom fields

In the lexical databases that we have designed to serve as a source for multiple types of output, we have come to rely on the tremendous usefulness of creating custom fields. Custom fields allow us to annotate, categorize, and label multiple facets, and organizing principles, of content that is unique to a particular project (see sections 3 and 5). Thus, we have created custom fields not only to house information that is destined to be output for a dictionary entry, but also for categorizing FLEx Entries as candidates for certain outputs, as well as for managing workflow. For example, in the Iquito database, we have a field for tagging the subset of Entries that will appear in the teaching dictionary and another to label every Entry whose Spanish content has been proofread by a native Spanish user.

The most important by far of our custom fields, however, is the Entry History field, which we use to communicate with one another, and our future selves, regarding the creation, development, and editing of the Entry. We discuss this field at greater length in section 11.

9 What counts as a headword in your dictionary?

A decision that will have significant ramifications for your lexicographic project is settling on a principle that divides words that you select as headwords from those that you do not. From the perspective, common among linguists, that considers the lexicon to be the locus of unpredictable form-meaning pairings that constitute the input into grammatical processes and structures, this is, in principle, an easy decision: headwords should be restricted to morphologically simplex words and morphologically complex words with non-compositional semantics; while morphologically complex words with compositional semantics should be excluded as headwords.

While this is an excellent guiding principle, there are multiple ways in which matters can be more complicated. First, it has been our experience that as our knowledge and understanding of a language's lexicon deepens, we have discovered areas of the lexicon that exhibit gradience between compositionality and non-compositionality. Second, while a phenomenon may be compositional, its permissible realizations may not be predictable. We illustrate this with the example of Iquito pluractional suffixes and classifiers. The semantics of verbs and adjectives, respectively, when they bear these morphemes *is* entirely predictable on the basis of the verbal and adjectival stems involved. However, what *is not* entirely predictable is which verbal or adjectival stems can bear these morphemes. Our solution was to create a Related Forms field that allowed us to record these forms in the relevant verb or adjective Entry, but not to promote them to headwords.

Finally, a different issue is raised by the fact that the contact language will typically lexicalize concepts that are not lexicalized in the target language, in which these concepts may be expressed via entirely productive processes such as verb serialization or derivation. Especially in contexts of significant language shift, however, users may search for words on the basis of concepts lexicalized in the contact language. Take the example of Spanish *ladrón* "thief," whose translational equivalent in Iquito is a wholly productive nominalization of a verb glossable as *robar* "steal." This kind of situation presents the following dilemma: if one does not include the nominalized form as a headword, then someone looking for the translational equivalent of *ladrón* will fail to find it, which may be frustrating to them (and incidentally, give them a dim view of the dictionary). On the other hand, if we organize the dictionary around concepts lexicalized in the contact language, the documentation of the target language can be significantly distorted. The solution we adopted was to include the form for *ladrón* in the Related Forms field in the Entry for the Iquito verb meaning "steal." This way, the word for *ladrón* is included and findable (especially via search functions on PDF versions of the dictionary), but it does not impose Spanish lexicalization onto the principles for what counts as a headword for Iquito.

10 Checking, verifying, and expanding dictionary content

Lexicography is partly a *science* (in the spirit discussed by Berez-Kroeker et al. 2018:6–7; see also Berez-Kroeker et al., chapter 1, this volume; Gawne & Styles, chapter 2, this volume) but it is also an *art* in a significant sense, and in the case of endangered language lexicography, it is a research endeavor so heavily dependent on the knowledge of a small number of specific individuals that the totality of the qualitative results cannot be considered reproducible. Nonetheless, the methodology presented here is principled, rigorous, reproducible, and valid. And to the degree that dictionary work is linked to and exemplified by textual materials, a reasonable degree of intersubjectivity and verifiability can be obtained, for which we discuss some methods and strategies next.

10.1 Establishing the validity of lexical data

The glosses, definition(s), and exemplification that a consultant provides on any single occasion for a particular lexeme have a complex relationship to the meaning(s) and function(s) of that lexeme that emerge after prolonged study. Though consultants often do provide highly insightful and nuanced definitions that cannot be significantly improved, this is not always the case. In particular, the information a consultant provides on a given occasion may suggest either a narrower, or a broader, meaning or function than the precise picture you'll be able to present after more extensive investigation. Often, this is because the specific situation or discourse context that is evoked on a single occasion strongly affects how the consultant construes the word. Simultaneously, there may be crucial mismatches between forms available in the contact language used to gloss or translate words in the target language, on the one hand, and the meaning of the words in the target language itself, on the other.

The principle strategy that we now rely on for both verifying our data and clarifying our definitions is to consider the development of a definition as a multistep process through which we build up a coherent view of the meaning of a headword from multiple vantage points. This multiperspectival view can be developed by examining multiple uses of the lexeme in texts; by asking multiple consultants to reflect on the meanings of the form; and by asking the same consultants to do this on different occasions, such as during different field seasons. It also includes asking consultants to reflect on translations provided by others, and, in some cases, to explicitly evaluate the hypotheses we develop about the meaning of the form. This outcome of this process is a stable, intersubjectively valid entry

for the form, with its steps and participants annotated as appropriate in the Entry History (see section 11).

10.2 The role of exemplification

Example sentences that illustrate the meaning and/or grammatical properties of the definition (or senses) of a headword serve multiple ends. Most importantly, they provide evidence for the generalizations presented in the entry, while the process itself of obtaining examples is a valuable method for checking, evaluating, testing, and further developing an entry.

An obvious and excellent source for example sentences are texts that have been recorded and transcribed (and ideally parsed in FLEx as well; see section 5.1) as part of the broader documentation project, and such texts should, in many cases, be prioritized as a source of example sentences. That said, our experience suggests that the corpus available for a language that is the focus of language documentation will typically not be sufficient as the sole source of example sentences. Not only may the relevant lexemes simply be missing from the corpus (see section 9.1), but also the sentences found in the corpus may not clearly exemplify important aspects of meaning or function, or they may be overly long or otherwise unwieldy. Elicitation of examples in one form or another thus becomes inescapable.

We have found the following three methods for example elicitation especially fruitful. First, ask consultants to make up sentences including the lexeme in question. If a consultant is experiencing difficulties with doing this from scratch, we find it useful to ask about common contexts in which the lexeme would be used; then, after fleshing out the context, we ask the consultant to produce an utterance (or utterances) appropriate in that context.

Second, in collaboration with a consultant, adapt a sentence already in the corpus so that it exemplifies the meaning or function of a lexeme more clearly or succinctly. And finally, in collaboration with a consultant, develop a target-language translation of a contact-language sentence, which is sometimes the most practical and efficient means of focusing on a very specific aspect of meaning or function. As with all the strategies discussed in this section, sentences obtained in this way should evaluated with the multiperspectival approach sketched in section 10.1.

10.3 Verification and validation through principled comparisons

A third important method for checking and expanding generalizations regarding meaning(s) and function(s) is by drawing out and identifying the differences between two or more similar lexemes. We approach this elicitation-based strategy in two primary ways: first, by comparing lexemes within a given semantic domain, and second, by comparing lexemes that bear similar or identical glosses or reversal-list meanings. For example, Iquito exhibits two verbs that consultants readily gloss as "break," further qualifying that they apply to slender, rigid objects such as sticks and bones: *tihaka* and *nasikata*. But only when asked to compare and contrast these two forms were our consultants able to articulate the difference: the former entails that the two parts are completely separated, while the latter entails that the two parts remain connected in a flexible manner (e.g., a broken stick whose two parts remain connected by a piece of green bark). In addition, this comparison-based strategy has also proved very successful in helping consultants to identify variants of various types, including dialectal variants, archaisms, and synonyms with particular affective flavors.

11 Tracking and managing the development of an entry

As the preceding discussion makes clear, the development of any single lexical Entry is typically a prolonged process; it typically involves contributions from multiple consultants, and it may also involve the contributions of multiple linguists, as has been the case for several of the dictionaries that we have developed. Both of these factors—a prolonged development period and multiple contributors—make it likely that, without efforts to address the issue, relevant information will be forgotten and complementary viewpoints—or discrepant ones—overlooked.

The solution we have developed for addressing this specific issue is to create an Entry History field in our FLEx databases, which serves as a running log of changes made to the Entry. This field is absolutely indispensable for knowing what has been done when, by whom, and with whom, as well as for annotating what remains to be done, by whom, and with whom. It is a place to note complementary or discrepant information from different sources, to spell out apparent problems in the

Table 24.5
Entry History examples from the Iquito Language Documentation Project FLEx lexical database

FLEx Entry	FLEx Entry History
headword • Part of speech glosses	• Participants are identified by three initials. • At first use, abbreviations are spelled out between square brackets. • CHECK or TODO means not yet checked or done, while CHCK or TD means already checked or done.
mɨ̈isaji • Noun 1. woman 2. female	LDM 23.09.2006 added lx [new lexeme Entry]; LDM 23.10.2006 JPI now says 'tiene pausa' so recheck hw [headword]; LDM 24.10.2006 mod [modified] hw mɨsáji > mɨ̈isáji, expanded def [definition] with JPI, now 2 senses; BGG Praat [segment lengths annotated]; CMB 12sep2015 RNLT [removed non-lexical tone mark]; 2015CHCK confirm tone is non-lexical; 2015TD WED [write English definitions]; LDM 25sep2015 conf [confirmed] non-lex tone, WED [wrote English definitions];
iitimɨra • Noun irregular plural of: *mɨ̈isaji*	LDM 23.09.2006 added lx; LDM 28.09.2006 mod hw itímɨra > itimɨra; LDM 23.10.2006 conf lx with JPI; 2015CHCK all; LDM 13oct2015 hw itimɨra > iitimɨra, RNLT, WED; 2015TD conf for humans only? yes: pɨsɨkɨ mɨ̈isajika; LDM 15jul2017 ELY has n/poss [possessed vs. non-possessed] alt, nu-iItim+ra; LDM 10jun2018 confirmed, added to IrregPoss [field]; 2019CHECKJPI GrammNote: confirm and add what to use in non-human cases
awasi • Noun digit, finger, or toe	LDM 05.09.2006 added lx; LDM 08.11.2006 lx def; LDM 06jun2016 JPI conf lx, irreg.pl.; LDM 18jun2016 JPI conf def, WED; LDM 28jul2016 ELY has toneless awAsi for nposs [non-possessed] form; 2016TD add this variant; 2016CHCKJPI to see if he has same? he does not; 2017CHCKELY again on her variants; 23jun2017 ELY confirms nposs form is toneless, poss form is Awasi; 2017TD discuss w/CMB what to do in such cases where JPI does not show a variation; LDM 24jun2017 hw áwasi > awasi, created hw for poss form áwasi; 2017CHCKELY plurals for poss, nposs forms; LDM 06jul2017 ELY has awasi no lextone for nposs, but nAwasi, nawAsika for poss; 2017TD deal with how to annotate JPI variants? added to GramNote [field]; LDM 15jul2017 ELY confirmed n/poss alt, added to IrregPoss [field]; CMB 13nov2019 added 3.poss.Var [populated Variant field to generate a Minor Entry]

entry, and to summarize how problems were resolved. In our work, the Entry History has allowed us to identify, among other things, subtle errors in representation (e.g., by noting discrepancies between different contributing linguists' representations of lexemes), and dialectal or sociolinguistic variation (e.g., by noting differences among consultants).

In practice, every annotation in the running Entry History begins with the initials of the annotating linguist and the date of the annotation, followed by terse but specific prose, and also including a consultant's initials when relevant. In addition, we have employed the Entry History field to tag or spell out tasks that remain to be carried out, such as exactly what to check with consultants on various aspects of the entry. Three illustrative examples, edited for readability, are given in table 24.5.

12 Conclusion

It is our hope that this chapter has presented one approach to managing lexicography data with sufficient motivation and methodological detail that interested readers will be able build on this methodology for their own lexicography projects. The work of lexicography is neither simple nor swift, but it can be an immensely rewarding contribution to a language and its users.

Notes

1. https://software.sil.org/fieldworks/.

2. https://www.w3.org/TR/REC-xml/.

3. https://www.latex-project.org/.

4. https://www.python.org/psf-landing/.

5. This has included work with users/rememberers of: Nanti (ISO 639-3: cox; 1995–2010), Iquito (iqu; 2001–present), Omagua (omg; 2003–present), Aʔɨwa (ash; 2008, 2010), Andoa (anb; 2009), Muniche (myr; 2009–2010), Máíhɨ̃kì (ore; 2009–present), Caquinte (cot; 2010), Matsigenka (mcb; 2010–present), and Záparo (zro; 2010–2011). Each of these projects has resulted in tangible, usable outputs for community members or for scholars or both.

6. A growing collection of materials of the Iquito Language Documentation Project is available at the California Language Archive at http://dx.doi.org/doi:10.7297/X2PC30JV.

7. A growing collection of materials of the Berkeley Máíhɨ̃kì Project is available at the California Language Archive at http://dx.doi.org/doi:10.7297/X2DR2SGD.

8. For managing work on Iquito, we have, in addition, used a separate Wiki to coordinate some multiparty procedures and workflow.

9. https://www.soas.ac.uk/elar/.

10. https://www.ailla.utexas.org/.

11. http://cla.berkeley.edu/.

12. Our digital tool for text transcription and translation work is ELAN Linguistic Annotator, available at https://tla.mpi.nl/tools/tla-tools/elan/.

13. Time permitting, including a *grammatical sketch* in the front matter is extremely useful to users, particularly to help them connect the roots and citation forms in your dictionary with the more complex forms that they encounter in interactions and texts.

14. If you are interested in acquiring a version of our script to adapt for a project of your own, please contact Lev Michael at levmichael@berkeley.edu.

15. Over the years, we have witnessed a number of efforts to develop alternatives or competitors to FLEx or its antecedents. Unfortunately, not one has yet been backed by the long-term institutional infrastructure necessary, in our view, for such tools to be a responsible and defensible choice for a researcher to make.

References

Atkins, B. T. S., and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

Beier, Christine, Brenda Bowser, Lev Michael, and Vivian Wauters. 2014. *Diccionario Záparo Trilingüe*. Quito, Ecuador: Ediciones Abya-Yala.

Beier, Christine, Lev Michael, Jaime Pacaya Inuma, Ema Llona Yareja, Hermenegildo Díaz Cuyasa, and Ligia Inuma Inuma. 2019. *Diccionario Escolar Ikíitu Kuwasíini—Tawɨ Kuwasíini (Iquito—Castellano)*. Iquitos, Peru: Cabeceras Aid Project. https://escholarship.org/uc/item/03m736sz.

Berez-Kroeker, Andrea, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56 (1): 1–18. https://doi.org/10.1515/ling-2017-0032.

Bird, Steven, and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79:557–582.

Bowern, Claire. 2008. *Linguistic Fieldwork: A Practical Guide*. New York: Palgrave MacMillan.

Dobrin, Lise. 2009. SIL International and the disciplinary culture of linguistics: Introduction. *Language* 85:618–619.

Epps, Patience, and Herb Ladley. 2009. Syntax, souls, or speakers? On SIL and community language development. *Language* 85:640–658.

Frawley, William, Kenneth C. Hill, and Pamela Munro. 2002a. Making a dictionary: Ten issues. In *Making Dictionaries: Preserving Indigenous Languages of the Americas*, ed. William Frawley, Kenneth C. Hill, and Pamela Munro, 1–22. Berkeley: University of California Press.

Frawley, William, Kenneth C. Hill, and Pamela Munro, eds. 2002b. *Making Dictionaries: Preserving Indigenous Languages of the Americas*. Berkeley: University of California Press.

Haviland, John. 2006. Documenting lexical knowledge. In *Essentials of Language Documentation*, ed. Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 129–161. Berlin: Mouton de Gruyter.

Landau, Sidney. 2001. *Dictionaries: The Art and Craft of Lexicography*. 2nd ed. Cambridge: Cambridge University Press.

Michael, Lev, Christine Beier, and Stephanie Farmer, compilers. 2013. *Diccionario Bilingüe Máíj̱ɨkì–Castellano y Castellano–Máíj̱ɨkì. Borrador Agosto 2013*. Iquitos, Peru: Cabeceras Aid Project. http://www.cabeceras.org/mai_ore_diccionario2013.pdf.

Michael, Lev, Christine Beier, Jaime Pacaya Inuma, Ema Llona Yareja, Hermenegildo Díaz Cuyasa, and Ligia Inuma Inuma. 2019. *Iquito—English Dictionary*. Quito, Ecuador: Ediciones Abya-Yala.

Mosel, Ulrike. 2011. Lexicography in endangered language communities. In *The Cambridge Handbook of Endangered Languages*, ed. Peter K. Austin and Julia Sallabank, 337–353. Cambridge: Cambridge University Press.

Nichols, Johanna, and Ronald L. Sprouse. 2003. Documenting lexicons: Chechen and Ingush. In *Language Documentation and Description*, vol. 1, ed. Peter K. Austin, 99–121. London: SOAS.

Ogilvie, Sarah. 2011. Linguistics, lexicography, and the revitalization of endangered languages. *International Journal of Lexicography* 24 (4): 389–404. http://dx.doi.org/doi:10.1093/ijl/ecr019.

Rice, Keren. 2018. Reflections on documenting the lexicon. In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton. Special issue, *Language Documentation and Conservation* 15: 180–190. http://hdl.handle.net/10125/24819.

Woodbury, Anthony C. 2011. Language documentation. In *The Cambridge Handbook of Endangered Languages*, ed. Peter K. Austin and Julia Sallabank, 159–186. Cambridge: Cambridge University Press.

Zgusta, Ladislav. 1971. *Manual of Lexicography*. The Hague: Mouton.

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