LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 13 (2019), pp. 63–82
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24799
Revised Version Received: 26 Nov 2018

Accessing, managing, and mobilizing an ELAN-based language documentation corpus: the Kwaras and Namuti tools

Gabriela Caballero
UCSD

Lucien Carroll
Cisco

Kevin Mach
UCSD, Adobe

This paper introduces *Kwaras* and *Namuti*, two new tools for building, managing, accessing, and mobilizing ELAN-based language documentation corpora. Kwaras integrates WAV files, ELAN annotations, and document metadata into a web-based corpus, allowing immediate access to annotations and recordings. Namuti builds from Kwaras and enables different uses of language documentation products for different audiences and provides links from linguistic analyses to language documentation corpora. The main goal of these new tools is threefold: (i) to facilitate the use of language documentation in linguistic analysis; (ii) to increase transparency of documentation-based analyses, providing interested users full access to the data on which generalizations are based and contextualization of the projects that generated the data; and (iii) to enable uses of language corpora that may serve the interests of multiple stakeholders, including academic researchers and community members interested in language maintenance and revitalization. We provide a basic overview of how Kwaras and Namuti work, lay out instructions on how to download and use Kwaras, and discuss what uses it currently supports. This article also issues a call for increased collaboration between linguists, community members, language activists, and software developers to further develop these and other similar resources.

1. Introduction[1] Research on the different subfields of linguistics from both theoretical and typological angles has had a long history of addressing phenomena in understudied languages with data obtained through field research. The development of technological innovations in the recording and assembly of corpora of primary

[1]We would like to thank two anonymous reviewers, Kate Lindsey, and Laura McPherson for helpful comments and suggestions. The work reported here was supported by the National Science Foundation/DEL through a grant awarded to Gabriela Caballero (award #1160672). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

E-ISSN 1934-5275

data allow us to move from traditional description to more comprehensive documentation, still largely missing for these languages, which addresses needs from both stakeholder communities and academics concerned with creating more representative records of natural languages in a variety of contexts of use (Himmelmann 2006; Woodbury 2010). This development goes hand in hand with increased concern in typological/theoretical research to pair quantitative analyses with the more traditional qualitative approach, and implement new methodologies (instrumental, experimental, computational, and corpus-based) that are still largely missing for lesser-known languages (Norcliffe et al. 2015; Whalen & McDonough 2015). This brings up important questions and challenges for those interested in bringing data from underdocumented languages to bear on typological and theoretical questions, including: what constitutes valid empirical evidence in research in phonology, morphology, syntax, etc.? How do we develop theoretically and typologically sound analyses that respect the patterns of variation inherent in any speech community? What role do we confer to variation patterns that may stem from language obsolescence processes (often found in understudied languages)?

These questions are also crucial in the development of comprehensive grammatical descriptions of understudied languages. Reference grammars continue to provide the empirical backbone of developing linguistic theories, research in linguistic typology, and the creation of pedagogical materials geared towards language maintenance and revitalization. But, as described in Evans & Dench's (2006) metaphor of grammar writing as "catching" language, reference grammars only capture static pictures of complex linguistic systems. Documentary corpora, on the other hand, provide a more representative window into language as a dynamic system with significant variation and change in progress. However, the link between documentary corpora and specific products, such as linguistic analyses, needs to be clearly articulated in each individual case. As pointed out in Mosel (2014), reference grammars based on documentation-based corpora do not generally provide many details of the content and structure of the corpus on which they are based. Ideally, users of linguistic analyses and their source documentation corpora would be able to go back and forth between these two products (see Thieberger (2009) for a proposal of how grammars embedded in data could be conceived and developed).

There are multiple tools available to develop linguistically annotated language corpora, but ELAN, developed by the Max Planck Institute for Psycholinguistics in Nijmegen (Sloetjes & Wittenburg 2008), is perhaps the most used by language documentarists.[2] ELAN is a flexible, open source software that enables time-aligned, XML-encoded annotations, following best practice in language documentation for long-time preservation and archiving (Bird & Simons 2003). Several tools have been developed to make ELAN files viewable and usable outside of the ELAN platform, in order to improve mobilization of corpora among multiple audiences, to facilitate access to linguistically annotated data, e.g., by viewing or searching data contained in documents of different formats in a single repository (for a review, see Dobrin & Ross 2017), as well as improving accountability of linguistic analyses (see also

[2]http://tla.mpi.nl/tools/tla-tools/elan/.

Thieberger 2009; Berez-Kroeker et al. 2018; Kaufman & Finkel 2018). This paper introduces two new open-source tools that further contribute to the management and mobilization of ELAN-based language documentation corpora, *Kwaras* (https://github.com/ucsd-field-lab/kwaras) and *Namuti* (https://github.com/ucsd-field-lab/namuti-webapp-template/).[3]

Kwaras is an interface that integrates WAV files, ELAN annotations, and document metadata into a web-based corpus, allowing immediate access of annotations and recordings and searches through different fields, as well as generating time-stamped, unique identifiers for individual annotations that can be used as citation forms in linguistic analysis. Namuti, a web-based user interface that inherits the structure of Kwaras, enables different uses of language documentation products for different audiences, and provides links from linguistic analyses to language documentation corpora. The main goal of these new tools is three-fold: (i) to facilitate the use of language documentation in linguistic analysis; (ii) to increase transparency of documentation-based analyses, providing interested users full access to the data on which generalizations are based and contextualization of the projects that generated the data; and (iii) to enable uses of language corpora that may serve the interests of multiple stakeholders, including community members interested in language maintenance and revitalization and academic researchers. Both Kwaras and Namuti were developed in the context of the Choguita Rarámuri (Tarahumara) language documentation project (Caballero 2017), and we exemplify their use through products arising from this project. We provide instructions on how to access and use Kwaras, for which we have developed a user-friendly package and instructions that require no prior programming experience. Crucially, this article also issues a call for increased collaboration between linguists, community members, language activists, and software developers to further develop these and other similar resources.

This paper is structured as follows. In §2, we provide an overview of the basic functions of Kwaras and Namuti, what uses they currently support, and their contribution in the context of a growing set of tools and resources available to language documentarists. In §3, we describe where to access Kwaras and provide instructions on how to use it, as we have developed a user-friendly version that does not require users to have any programming knowledge. We conclude in §4, highlighting aspects that require improvement for both tools and lay out possibilities for future developments.

2. Kwaras and Namuti: an overview

2.1 Kwaras Kwaras is an ELAN corpus management tool created by Russell Horton (Linguistics MA 2012, UCSD) and further developed by Lucien Carroll (Linguistics PhD 2015, UCSD). Its main function[4] is converting ELAN data into an html table

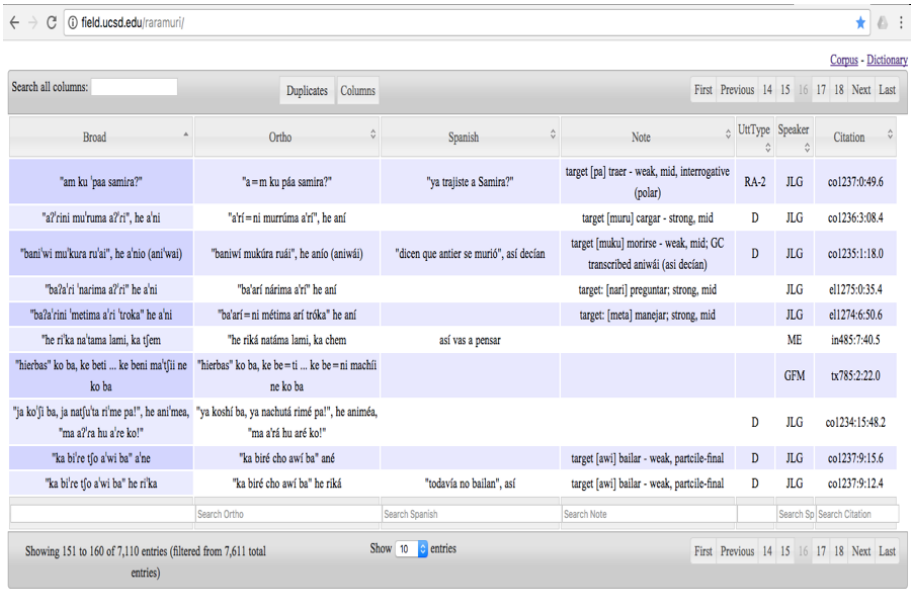
[3]Software is released under the MIT license (https://github.com/ucsd-field-lab/kwaras/blob/master/MIT-LICENSE).

[4]Kwaras is described here as an application, but it may also be used as a Python library to perform a couple of related functions: (1) Allow bulk edits in a corpus of ELAN files. This provides the ability

linked to sound clips. The result is a searchable interface that allows immediate access to transcriptions, annotations, and corresponding audio, with audio (WAV) clips accessible by clicking on the annotation tier.[5] The interface enables regular expression (regex) searches on individual fields (tiers from the ELAN annotation files) and cross-field searches where terms are matched as substrings in the data, regardless of order.[6] Audio files are also accessible for download and a unique identifier is generated by the software. Contributor information may also be added via the metadata spreadsheet that users may include when running Kwaras with their documentation files (more details about metadata information in Kwaras are provided in §3.3 below).

An example of a Kwaras-generated corpus is shown in Figure 1, with data from the Choguita Rarámuri corpus (http://field.ucsd.edu/raramuri).

Figure 1. Choguita Rarámuri corpus in Kwaras



As shown in Figure 1, the Choguita Rarámuri corpus provides broad phonetic and orthographic transcriptions, Spanish translations, linguistic notes, and codes (*Note* and *UttType*, respectively), source metadata information (including contributing speaker), and a citation system for referencing and finding individual utterances and lexical forms (e.g., 'co1237:0:49.6'), which are derived from the source file name (e.g., 'co1237') plus the time stamp of the annotation referenced (e.g., '0:49.6'). Other types

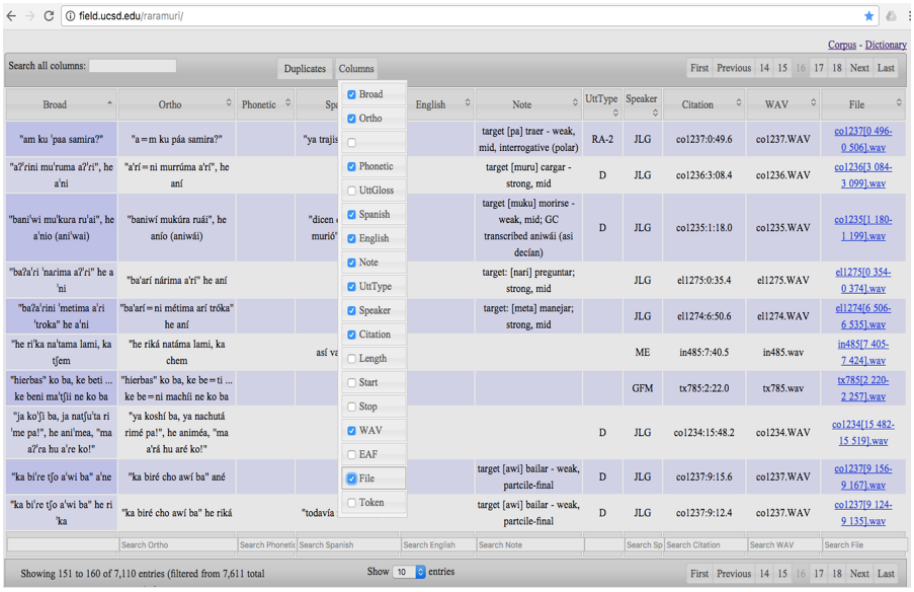
to enforce template conventions defined for the language and corpus, e.g., ensuring that translation tiers are dependent on transcription tiers, and that orthography tiers only use the standard characters. (2) Transform LIFT lexicon files (e.g., as exported from FieldWorks Language Explorer) to EAFL lexicon files used by the morphological analyzer in ELAN-Corpa.

[5]Clicking anywhere in the row generated by Kwaras will display a sub-table where a link to the corresponding audio file is available.

[6]The search patterns in Kwaras are the standard ones found in similar architectures, e.g., '$' matches the end of a string, '^' matches the beginning of a string, '.' matches any single character, etc. These are case-insensitive Javascript (mostly Perl-compatible) regexes.

of information available for the Choguita Rarámuri corpus include morpheme-by-morpheme glosses (available for a subset of the files), as well as other annotations (grammatical or other) available in the original ELAN annotations. This is possible due to the flexibility afforded by Kwaras, given it allows the analyst to choose which annotation tiers to import from ELAN for each individual project. Thus, the structure of Kwaras-generated corpora results from a selection of ELAN tiers that is required upon installation (more details about this function are given in §3). This flexibility is not only available during set-up, but also can be manipulated by individual users through the selection of columns to display, an option that can be selected in the webpage. This feature is exemplified in Figure 2.

Figure 2. Selection of columns available for display in the Choguita Rarámuri corpus

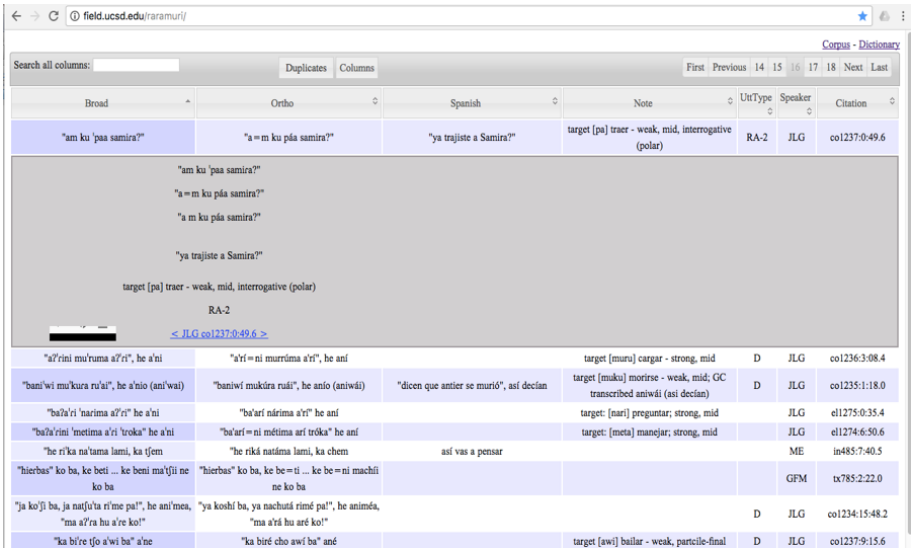


It should be noted that tier types, language associations, and tier hierarchy from ELAN annotation files are not preserved in Kwaras. The first tier name provided in the configuration window (described in §3.4) is used as the "main annotation tier" (any time-aligned tier in ELAN is suitable), and Kwaras groups annotations by time stamp, keeping all aligned tiers but ignoring the tier hierarchy. Clips are only reliably provided for annotations on time-aligned tiers, in addition to annotations on dependent-tiers that are aligned one-to-one with time-aligned tiers. Any other annotations will be exported but not linked correctly to the sound file.

In addition to the specific fields that are defined by the structure of ELAN files in each individual project, Kwaras generates additional information for all projects, including the citation form referenced above and a link to *File*, which provides a hyperlink to the individual annotation referenced. This link allows the user to either download the sound recording directly (via the browser's *Save Target* contextual menu option) or to load the audio file in the browser, enabling the user to listen to the sound recording or to download it.

As mentioned above, the audio recording is also accessible through the main corpus page by clicking on an annotation line. The audio file will play automatically in Chrome and Edge browsers when the annotation is clicked open, while in other browsers the user needs to explicitly open the provided link to hear the audio. An example of the display after clicking on an annotation line in the corpus is exemplified in Figure 3.

Figure 3. Clicking on an annotation line opens an interlinear display and allows access to individual audio files in the Choguita Rarámuri corpus.

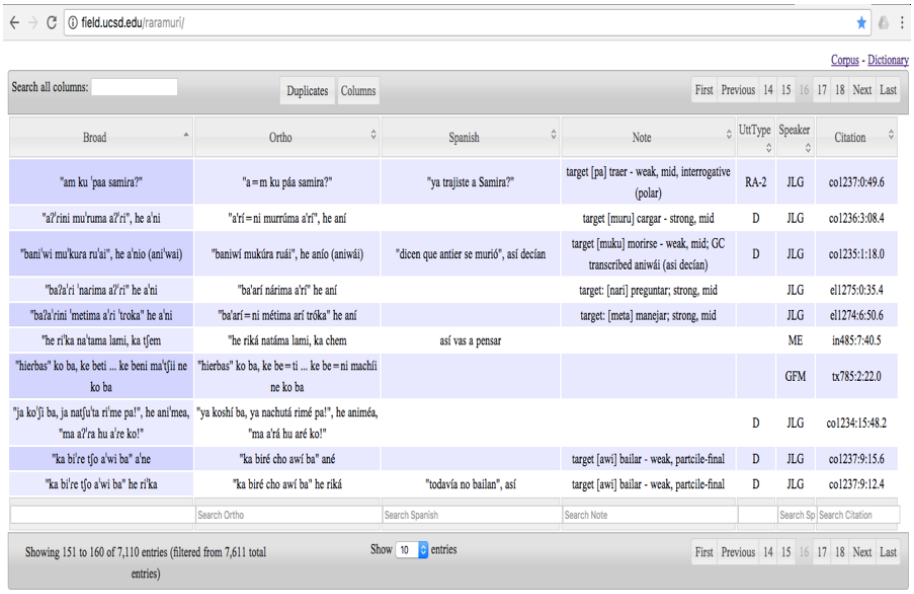


As shown in Figure 3, accessing the audio also enables a display where the annotations are provided in interlinear fashion, with the speaker information and hyperlink of citation information provided at the bottom.

A global search function is placed at the top of the webpage, with field-specific search functions at the bottom of each field (highlighted in Figure 4 with arrows). As shown in Figure 4, the field-specific search boxes allow looking up transcribed data, translations, and any other information contained in the annotation fields, as well as data contributed by individual speakers. There is also the possibility of carrying out searches of data contained in individual annotation documents (the *Citation* field in the rightmost column).

Kwaras users have thus a variety of resources available when accessing language corpora, including fast and easy access to annotations and corresponding sound files (including the ability to download sound files for further analysis or inspection) and an automatic way of generating unique identifiers for individual annotations which can be used in linguistic publications, materials in archives for long-term preservation, or other references on the language. These two aspects of the software increase the potential for data transparency by allowing users to access original recordings and annotations and the ability to have access to the larger contexts from which individual examples of linguistic analysis come from. This allows one to ask, for instance,

Figure 4. Individual search fields of the Choguita Rarámuri corpus (highlighted with arrows)



whether any data point arises from elicited data, a text, a conversation, or another kind of document, what kind of methodology or protocol was used in obtaining the data, and other information that may reveal the cultural and/or linguistic context of the data examined.[7] Furthermore, it enables the user to find neighboring utterances within a conversation or elicitation session to reveal the discourse context of particular examples. Finally, Kwaras is a powerful tool in grammar writing and other forms of linguistic analysis, since it allows finding examples of specific constructions from a variety of speech genres and from several speakers through its search engines.

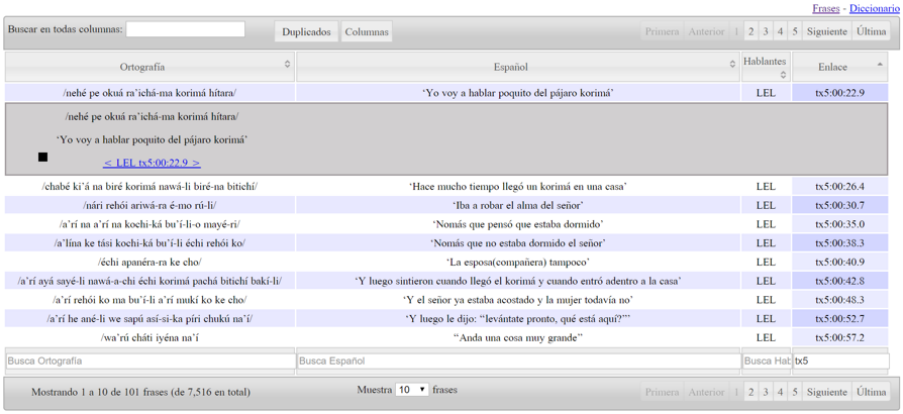
The Kwaras interface can be accessed over the internet, but it can also be loaded directly from a file directory for off-line access. Combined with the choice of shown fields, this enables versions suitable for use in remote communities where internet access is limited. Figure 5 shows a version of the Choguita Rarámuri corpus in use in Choguita. The default displayed columns in this version are limited to those of practical use to native speakers primarily interested in reviewing narratives or lexical items. Users can view a single narrative by searching for the narrative file name in the citation field. In this version, the default sort is on the citation (*Enlace*) column so that narratives are displayed in the proper order.

Though Kwaras can be useful for community members engaged in language teaching, the interface is still not optimal for display of full texts. Kwaras is thus best suited for developing and verifying linguistic analyses. It is primarily a web-based user interface containing non-curated materials that is maximally useful for the creators of

[7]Kwaras does not replace producing metadata for language documentation products, but it does provide the possibility of a much richer contextualization of the documentation.

language corpora, but not for other users who are not familiar with the structure of the ELAN files used for generating the annotations. The next section describes how a user-interface that delivers language documentation corpora to multiple audiences builds from Kwaras.

Figure 5. A version of the Kwaras corpus of Choguita Rarámuri adapted for community use (with a local orthography as main entry, a Spanish translation, a speaker code, and citation form)



2.2 Namuti The previous section highlights the advantages of a tool like Kwaras in the development of linguistic analyses given the ability to increase transparency of the analyses proposed. Increased transparency is also necessary for audiences that have a different relationship to language documentation corpora, such as community members who are interested in maintaining/revitalizing their ancestral language or creating materials for language learners within the main stakeholder community. A tool that delivers documentation corpora to stakeholder speech communities can also allow interested community members to enrich and further develop the existing language documentation materials as they deem necessary for their own language planning purposes.

Given these desiderata, a second tool has been developed within the context of the Choguita Rarámuri documentation project: Namuti. Namuti originated with the goal of creating a curated, open access user interface containing materials that individual contributors authorize as open access, with the option of visualizing the materials in a variety of configurations that may be audience-specific (i.e., community members or academic users). The features of Namuti are exemplified here through the Choguita Rarámuri Language Project webpage (http://raramuri.ucsd.edu), which is still under development.

Namuti has the same general capabilities as Kwaras, providing links to audio files and associated annotations (including transcriptions, glosses, and translations), as well as speaker and citation information. In addition to this, documents (in this case texts) can be visualized in a variety of ways. One way for displaying texts is in a

Story View, which may be appealing for community members; in Figure 6, texts are displayed with a Rarámuri practical orthography transcription side-by-side a Spanish translation, the language of heritage speakers for whom language shift has taken place. In *Gloss* views, which may be of greater interest to academic users, texts are displayed with linguistically annotated data, as well as Spanish and English translations. This is shown in Figure 7. All views support access to audio, either for entire documents, or for individual annotated utterances, with clips provided for each individual utterance.

Figure 6. *Story* view display of texts in the Namuti corpus of Choguita Rarámuri

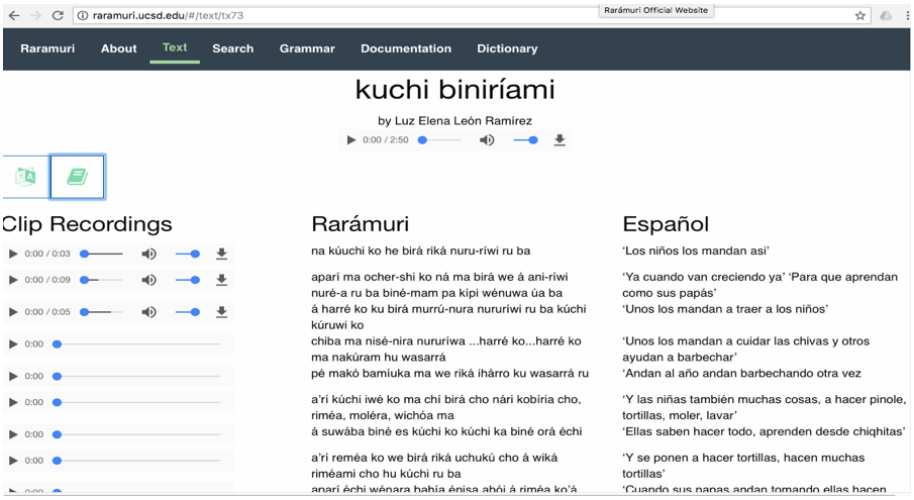
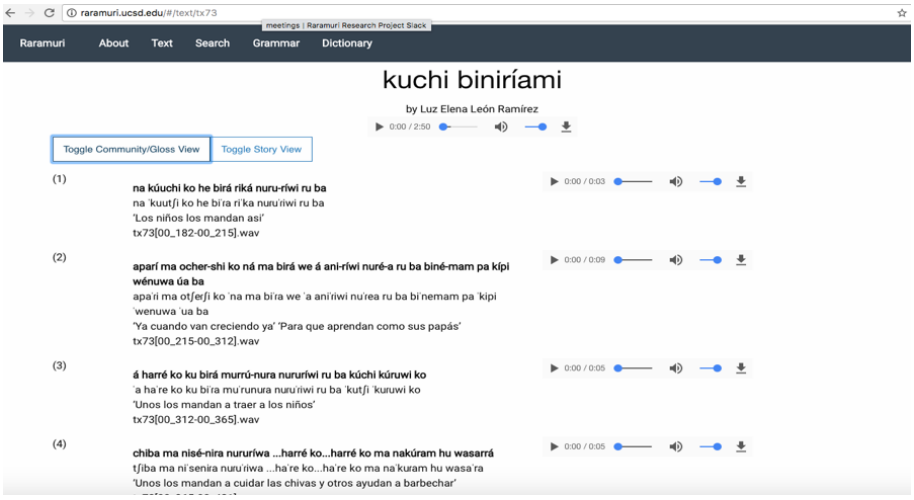


Figure 7. *Gloss* view display of texts in the Namuti corpus of Choguita Rarámuri



Texts are listed by title (Figure 8) or by contributing authors (Figure 9).[8] Providing texts listed by contributor may be of interest to community members who want to access legacy materials of particular families. Accessing materials by contributing speaker may also be relevant for users interested in inter- and intra-speaker variation patterns in this language.

Figure 8. Views of texts listed by title in the Namuti corpus of Choguita Rarámuri

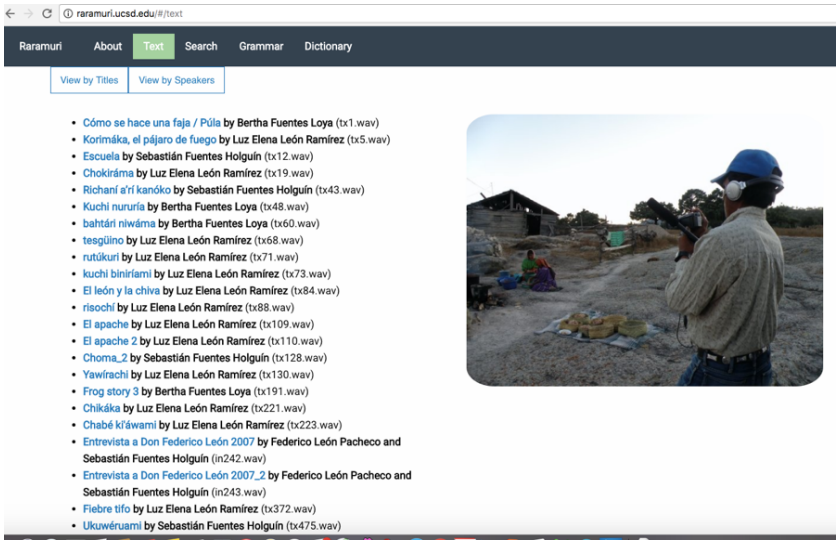
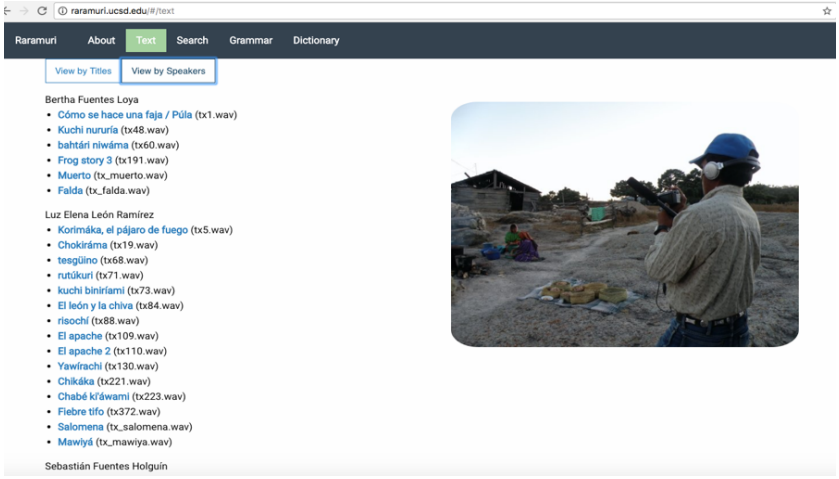


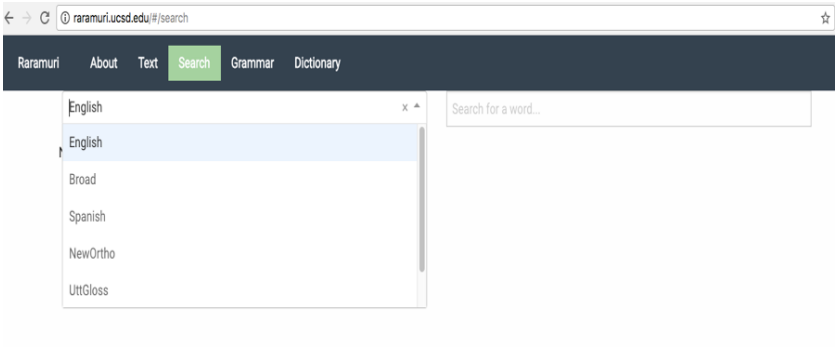
Figure 9. Texts listed by contributing author in the Namuti corpus of Choguita Rarámuri



[8]As shown in Figure 8, titles are highlighted providing links to each individual document. Each entry provides the contributing author's name, as well as the unique identifier for each text.

As mentioned above, Namuti inherits from Kwaras the ability to access audio files of individual annotations, as well as its search capabilities (both global and field-specific). This is illustrated in Figure 10.

Figure 10. Search engine in the Namuti corpus of Choguita Rarámuri



Namuti, like Kwaras, also has the crucial function of providing links between products of linguistic analysis and the documentary corpus. Individual utterances of texts are provided with their unique identifiers, and each text is also provided with its unique identifier, which references the deposited collection of Choguita Rarámuri in ELAR (available at http://elar.soas.ac.uk/deposit/0056). While this may be redundant, providing unique identifiers for both texts and individual annotations serves the purpose of linking the most recent annotations to their corresponding archived versions that are maintained for long-term preservation. In the case of the Choguita Rarámuri corpus, data is cited with the citation information generated by Kwaras (also available in Namuti) and provided with a link that directs the user to the source text from which the individual example was extracted. This is illustrated in Figure 11.

Figure 11. Result of accessing unique identifier hyperlink of an individual example within its source text in the Namuti corpus of Choguita Rarámuri

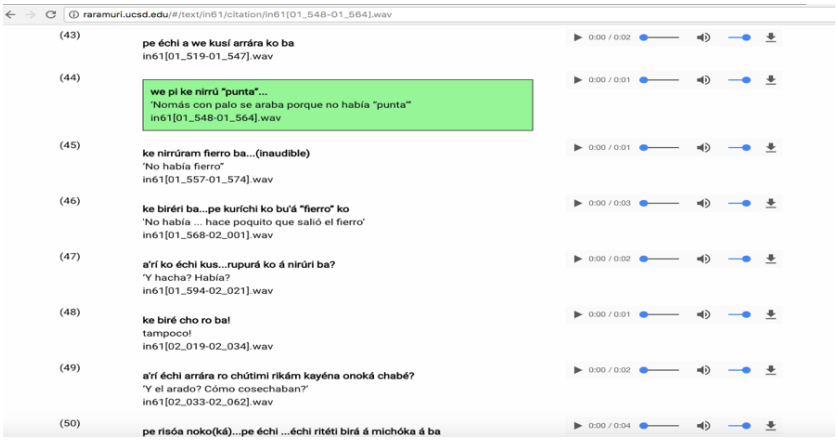


Figure 11 exemplifies the display after accessing the hyperlink of a particular unique identifier (in61[01_548-01_564].wav). Unique identifiers and corresponding hyperlinks (e.g., http://raramuri.ucsd.edu/#/text/in61/citation/in61 [01_548-01_564].wav) can be provided as the citation form in published linguistic analyses. A user interested in exploring the relationship between that particular piece of data and the documentary corpus from which it has been extracted can use the hyperlink to situate any given example within the source document. As shown in Figure 11, Namuti will highlight the individual example referenced (in this case, utterance (44) of document 'in61'). In the case of the Choguita Rarámuri Language project, a reference grammar of the language currently in progress will provide a substantial number of examples with links to this corpus, which will be open access upon its completion.

2.3 Kwaras and Namuti: new contributions In this section we address the contributions that Kwaras and Namuti bring *vis a vis* other existing tools designed for managing and accessing language documentary corpora. Comprehensive reviews of recently released tools are provided in Dobrin & Ross (2017) and Kaufman & Finkel (2018), who introduce the IATH ELAN Text-Sync Tool (ETST) and Kratylos, respectively. ETST and Kratylos are two web applications developed for the purpose of enhancing access to texts and corresponding audio (ETST) and interlinearized glossed texts and lexical data (Kratylos). The development of these tools, like Kwaras and Namuti, emerges in the context of increased need of enhancing data "reproducibility", verification, and accountability in linguistic research and the humanities at large (Thieberger 2009). As defined in Berez-Kroeker et al. (2018), reproducibility involves providing access to the original data for independent analysis.[9] A full review of the features of ETST, Kratylos, and other tools and web-based user interfaces is left out of the scope of this paper. Here we highlight some of the features that Kwaras and Namuti offer within the current available options for managing and accessing language documentation corpora.

As mentioned above, Kwaras and Namuti inherit the structures and functionality afforded by ELAN in a number of respects (overviews and reviews of ELAN can be found in Berez (2007) and Sloetjes & Seibert (2016)). Some features are familiar to ELAN users, including the possibility of carrying out regular expression searches. The search functionality afforded by ELAN is indeed more powerful than the one available in Kwaras and Namuti, as it gives the user control over case sensitivity and durations, and the user can restrict the search to an arbitrary set of EAF files. This is a feature that can be improved in future versions of Kwaras and Namuti.

The main strengths of Kwaras (some of which are also found with other tools) are: (i) providing a user-friendly, web-based version of a language documentation corpus; (ii) enabling off-line access to the data for local playback;[10] (iii) generating

[9]See Berez-Kroeker et al. (2018) for a discussion of the difference between reproducible vs. replicable data.

[10]Dobrin & Ross (2017) review the Ethnographic E-Research Online Presentation System for Interlinear Text (EOPAS; Schroeter & Thieberger 2006), another tool that, like Kwaras and ETST, is a web-based user interface to visualize ELAN annotations and associated multimedia files. EOPAS, however, does not allow off-line use, which is a disadvantage for several language projects where internet access is still not available. ETST, like Kwaras, enables off-line access to the user interface.

reliable unique identifiers of annotated data (via file names and time stamps); and (iv) allowing users the flexibility to structure the content of the interface both during configuration and on-line or off-line display. Indeed, the flexibility of its structure may be the most powerful feature considering the multiple needs that a single documentary collection may be designed to satisfy.

Namuti also enables flexibility in terms of how to access language materials, as well as facilitating making reference to, searching, and contextualizing linguistically annotated data. Some of the features offered in Namuti are available for other user interfaces of other language documentation projects. A few examples and corresponding URLs include the following:

(1) Language documentation projects with available web-based user interfaces

   a. Yurok Language Project: http://corpus.linguistics.berkeley.edu/~yurok/index.php
   b. Northern Paiute Language Project: http://paiute.ucsc.edu/
   c. Moro Story Corpus: http://linguistics.berkeley.edu/moro/#/
   d. Dictionary and text corpus of the Karuk language: http://linguistics.berkeley.edu/~karuk/resources.php

Like Namuti, the goal of the user interfaces associated with these language projects is to allow users to access language documentation and other language legacy materials in a user-friendly way. They share in common with Namuti the ability to access materials: (i) in a variety of formats (with views for language learners and views for academic or other audiences, e.g., with glosses and other linguistic annotations); and (ii) enabling different kinds of data searches. In addition to these features, and like Namuti, some interfaces also provide links to audio recordings of materials and citations for individual utterances of texts. Namuti departs from these interfaces in providing users the ability to download/access full recordings of individual texts as well of individual utterances and of providing unique identifiers as citation forms for each annotated unit. These citation forms, as mentioned above, may reference materials archived for long-term preservation.

There are multiple ways in which both Kwaras and Namuti can be improved. Before addressing these, we describe a user-friendly package for accessing and configuring Kwaras.[11]

3. Accessing and installing Kwaras The code for Kwaras is publicly available at https://github.com/ucsd-field-lab/kwaras. The website https://sites.google.com/view/g-caballero/kwaras contains built installation packages of the program for both Windows and Mac which do not require programming knowledge.

The off-line scripts are written for Python 2.7, and the dependencies are specified within the package. The web process depends on jQuery DataTables, including the

[11]We have not yet developed a user-friendly installation package for Namuti.

plugins TableTools and ColVis. The tested versions of these modules are included in the Kwaras directory available for download. This feature was designed for documentary collections with smaller set of annotations for the purpose of facilitating use of the interface off-line. However, and as one anonymous reviewer points out, this feature would not scale well for larger documentary collections.[12]

We provide next the instructions on how to install and configure Kwaras (these instructions are also available at https://sites.google.com/view/gcaballero/kwaras).

3.1 Step 1: Installing Kwaras The instructions for installing Kwaras on a Mac are listed step by step in (2):

(2) Kwaras installation for Mac

a. Download the kwaras-mac-2.2.1 directory and unzip it.
b. Move the whole folder to your working directory.
c. Double click the file 'install-macos.COMMAND' to install the Python library.

Kwaras can be installed on Windows platforms following the instructions in (3):

(3) Kwaras installation for Windows

a. Download the kwaras-win-2.2.1 directory and unzip it.
b. Move the whole folder to your working directory.

3.2 Step 2: Setting up data directories After installation is complete, the next step is to export data from ELAN. The export process depends on setting up four main directories of data, which are listed in (4). Users must set up these four data directories in order to configure Kwaras.

(4) Kwaras data directories

a. Transcriptions: a directory of ELAN (.eaf) files

b. Recordings: a directory of WAV format sound files, minimally containing a WAV file for each of the ELAN files in the transcription directory

[12]DataTables supports dynamically loading data via Ajax, but so far the corpora we have worked with have been manageable as single objects.

c. Web: contains a 'css' folder, an 'js' folder and 'index_wrapper.html' (these files are found in the 'web' folder inside the 'kwaras' folder)

d. Corpus: a directory for temporary output files

The Web data directory described in (4c) is initialized as a copy of the 'web' directory in the Kwaras package and is also where the files generated by Kwaras after export (an index.html file and clips directory) will go. After these files are created in this directory, this whole directory can be uploaded to a web server.

Kwaras uses the EAF media file reference to find the corresponding WAV filename, but if no media file reference is found, it will assume they share the same base name. In either case, the WAV files are required to be in a single directory because the absolute path in the media file reference is not consistently interpretable if the corpus is moved from one machine to another or if a corpus kept on a remote file system is mounted on two different computers.

3.3 Step 3: Preparing metadata information Kwaras enables users to optionally incorporate metadata information for their annotation files. Kwaras pulls speaker codes either from the ELAN tier names or from a metadata file that users provide. If tier names follow the conventional pattern of "word@SPEAKER", Kwaras will use the first element as the column name and the second element as a speaker code. For files with tier names that do not have "@SPEAKER" annotations, Kwaras will expect a metadata file in either a utf8-encoded CSV or an XLSX format, with the minimal following columns listed in (5):

(5) Minimal columns to include in a metadata file for use in Kwaras

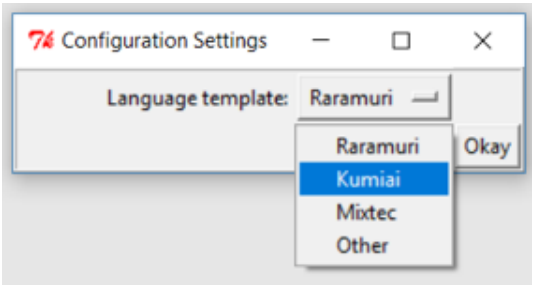
a. "File" for the basename (e.g. "tx143"),
b. "Contributor" for speaker codes (e.g. "BFL")
c. "Format" for the file extension (e.g. "wav")

Paired EAF and WAV files should either have the same basename or the WAV file should be linked media in the EAF file.

3.4 Step 4: Running the export-corpus function The fourth step for using Kwaras involves running the export-corpus function. Double-click the "export-corpus.COMMAND" file in Mac OS, or in Windows double-click "Kwaras.exe".

After this, the next step is to configure the export process. There are currently four language templates available (Figure 12).

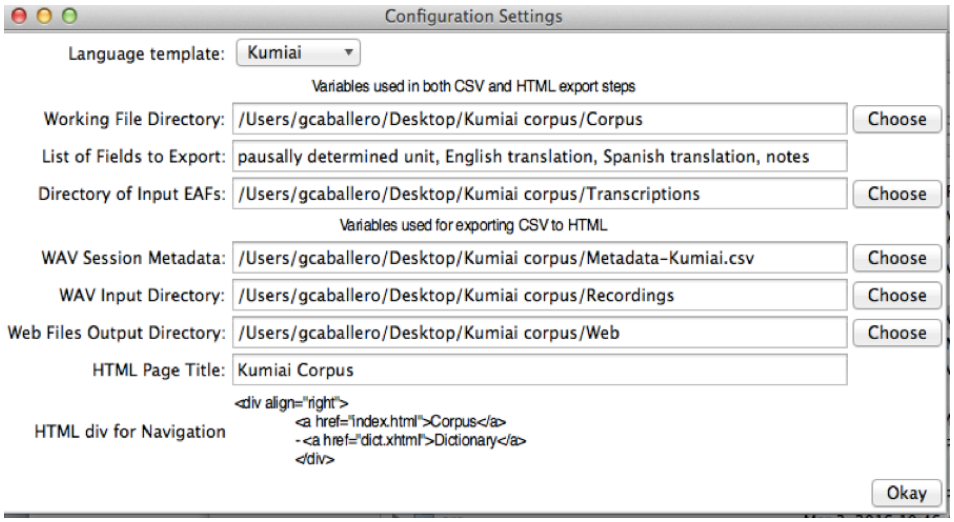
Figure 12. Select a language template in the Configuration Settings



The default template "Other" directly uses the EAF data, while the other templates have been designed for use in particular language documentation projects to enforce transcription conventions or populate empty fields (e.g., phonemic field from orthographic field or utterance glosses from word glosses). Users with new projects should select the "Other" template.[13]

After a template has been chosen, a configuration window is shown like the one in Figure 13.

Figure 13. Configuration settings for 'Kumiai' language template



The steps to follow once the configurations settings have been displayed are listed in (6):

[13]Currently, Kwaras does not allow users to create, load, or save their own templates, but interested users may develop this feature in the future. The current setup in Kwaras enables flexibility in the access of the documentary corpora by allowing users to select which ELAN annotation tiers may be displayed in the web version of the interface.

(6) Steps for setting up configuration in Kwaras

a. Working File Directory: select the Corpus directory.
b. List of Fields to Export: write down the tier names that should be extracted from the EAF files, separated by a comma and space. The tiers should be listed in the order that the user desires to display the data (Important: these tier names should be spelled *exactly* the way they are spelled in the ELAN files users are about to export).[14]
c. Directory of Input EAFs: select the 'Transcriptions' directory of EAF files.
d. WAV Session Metadata: select the metadata spreadsheet file (must be an XLSX or a utf8-encoded CSV).
e. WAV Input Directory: select the 'Recordings' directory with WAV files.
f. Web Files Output Directory: select the 'Web' directory.
g. HTML Page Title: Title showing in the header of the index.html file.
h. HTML div for Navigation: HTML code for a navigation bar (optional).
i. Press okay.

A Terminal window will open and display the process. Once completed, open the 'Web' folder and click on the 'index.html' file to display the corpus in your web browser.

4. Conclusion and future developments The goals of linguistic description and language documentation increasingly address the goals and needs of multiple stakeholders, which include enabling and improving access to original language data for a variety of purposes. This requires developing new resources and tools that are both powerful and flexible. This is the context in which Kwaras and Namuti were developed. These resources seek to address multiple audiences: (i) community members interested in language preservation; (ii) academics interested in language and culture research that pays attention to language ecology; and (iii) native speakers and language learners interested in language preservation and revitalization.

In terms of access to original data, this paper has provided illustration on how Kwaras and Namuti can be used as a tool in linguistic analysis through their search capabilities. Generating unique identifiers and having ready access to audio files of annotated data also allows closer inspection of the data referenced in linguistic analysis,

[14]As mentioned by one anonymous reviewer, this is another feature of Kwaras that makes it more suitable for smaller documentary collections with relatively little complexity vs. a project where documents include multiple speakers and a large number of annotation tiers. As mentioned in §3.3, the current version assumes participant codes in tier names are suffixed with '@' as the delimiter (e.g., 'phonetic@BFL'). Supporting other conventions will require adding other configuration options.

providing a unique opportunity to correct, expand, or replace existing analyses and allowing the possibility of identifying new phenomena previously overlooked. Community members (both native speakers and second language learners) can also have access to legacy materials that in the future could be developed through their continued annotation and development. This, for instance, can be made possible through incorporation of software features currently used in song lyric annotation websites (e.g., https://genius.com/), which could allow adding relevant linguistic, cultural, and historical context to language corpora from the perspective of the main stakeholders of these resources.

Both Kwaras and Namuti could be developed further to fit the needs of other documentary projects. Some areas of further development are identified in (7).

(7) Features to develop for Kwaras and Namuti[15]

a. Enable integration of video files with linguistic annotations.
b. Offer a compressed audio option (mp3 or ogg) for download.
c. Extend support to read annotations from Praat and other programs used by language documentarists.
d. Include support for integrating annotation of written resources in order to include digitized field notes, historical manuscripts, and other kind of materials that are often part of documentary collections.
e. Develop support for multiple audio tracks in ELAN transcripts.
f. Provide a mechanism to ensure stability of the data citations.
g. Enable higher compatibility with files exported from Flextext.
h. Develop the search capabilities of both Kwaras and Namuti further.

This list is, of course, non-exhaustive, and it only enumerates a few ways in which Kwaras can be improved to meet the needs of a wider array of users. In the case of the Choguita Rarámuri language, its rapidly changing sociolinguistic situation includes increasing displacement of native speakers into diaspora communities across Northern Mexico and, as a consequence, accelerated language attrition and obsolescence. On the other hand, Choguita Rarámuri speakers have increased access to new technology, which are more readily available in larger towns. Community members thus have changing needs and increased ability to access online language materials. Our hope is that Kwaras, Namuti, and other resources relating to language documentation and conservation will be of use to these and other communities and academics interested in bringing data from these languages to bear on theoretical and typological developments in linguistics and other academic fields.

[15]We would like to thank two anonymous reviewers for several suggestions included in this list.

References

Berez, Andrea. 2007. A review of EUDICO Linguistic Annotator (ELAN) from Max Planck Institute for Psycholinguistics. *Language Documentation & Conservation* 1(2). 283–289. http://hdl.handle.net/10125/1718.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice & Anthony Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18. doi:10.1515/ling-2017-0032.

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3). 557–582. doi:10.1353/lan.2003.0149.

Caballero, Gabriela. 2017. Choguita Rarámuri (Tarahumara) language description and documentation: a guide to the deposited collection and associated materials. *Language Documentation & Conservation* 11. 224–255. http://hdl.handle.net/10125/24734.

Dobrin, Lise M. & Douglass Ross. 2017. The IATH ELAN Text-Sync Tool: A Simple System for Mobilizing ELAN Transcripts On- or Off-Line. *Language Documentation & Conservation* 11. 94–102. http://hdl.handle.net/10125/24726.

Evans, Nicholas & Alan Dench. 2006. Introduction: Catching language. In Ameka, Felix, Alan Dench & Nicholas Evans (eds.). *Catching language. The standing challenge of grammar writing*, 1–39. Berlin, New York: Mouton de Gruyter. https://openresearch-repository.anu.edu.au/handle/1885/30297.

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

Kaufman, Daniel & Raphael Finkel. 2018. Kratylos: A tool for sharing interlinearized and lexical data in diverse formats. *Language Documentation & Conservation* 12. 124–146. http://hdl.handle.net/10125/24765.

Mosel, Ulrike. 2014. Corpus linguistics and documentary approaches in writing a grammar of a previously undescribed language. In Nakayama, Toshihide & Keren Rice (eds.), *Language Documentation & Conservation* SP08: *The Art and Practice of Grammar Writing*. 135–157. http://hdl.handle.net/10125/4589.

Norcliffe, Elisabeth, Alice C. Harris & T. Florian Jaeger. 2015. Cross-linguistic psycholinguistics and its critical role in theory development: early beginnings and recent advances. *Language, Cognition & Neuroscience* 30(9). 1009–1032. doi: 10.1080/23273798.2015.1080373.

Schroeter, Ronald & Nicholas Thieberger. 2006. EOPAS, the EthnoER online representation of interlinear text. In Barwick, Linda & Nicholas Thieberger (eds.), *Sustainable data from digital fieldwork*, 99–124. Sydney: Sydney University Press. http://hdl.handle.net/11343/34630.

Sloetjes, Han & Peter Wittenburg. 2008. Annotation by category – ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Sloetjes, Han & Olaf Seibert. 2016. New facets of the multimedia annotation tool ELAN. Poster presented at *Digital Humanities 2016 - Digital Identities: the Past and the Future*, Kraków, Poland. http://hdl.handle.net/11858/00-001M-0000-002B-98D4-7.

Thieberger, Nicholas. 2009. Steps toward a grammar embedded in data. In Epps, Patience & Alexandre Arkhipov (eds.), *New Challenges in Typology: Transcending the Borders and Refining the Distinctions*, 389–408. Berlin, New York: De Gruyter Mouton. http://hdl.handle.net/11343/26089.

Whalen, Doug & Joyce M. McDonough. 2015. Adaptable models: Taking the laboratory into the field. *Annual Review of Linguistics* 1(1). 14.1–14.21. doi:10.-1146/annurev-linguist-030514-124915.

Woodbury, Anthony C. 2010. Language documentation. In Austin, Peter K. & Julia Sallabank (eds.), *The Handbook of Endangered Languages*, 159–186. Cambridge: Cambridge University Press.

Gabriela Caballero
gcaballero@ucsd.edu

Lucien Carroll
lucien@discurs.us

Kevin Mach
kevin.mach88@gmail.com