Analyzing Code-mixing in Linguistic Corpora Using Kratylos

RAPHAEL FINKEL, Department of Computer Science, University of Kentucky
DANIEL KAUFMAN, Department of Linguistics, Queens College CUNY & ELA
AHMED SHAMIM, Department of Asian Studies, University of Texas at Austin

Code-switching, code-mixing, and, more generally, multilingualism pose technological challenges for language documentation, the sub-discipline of linguistics that deals with the annotation and basic analysis of field recordings and other primary data. We focus here on a case study involving code-mixing in the endangered Koda language, which poses special problems for morphosyntactic analysis. We offer a robust approach to multilingual annotations that involves a combination of the popular open source software FieldWorks Language Explorer (FLEx) with Kratylos, a web-based corpus tool for display and query. Kratylos exposes linguistic data from various formats to powerful regular-expression queries that can exploit tier structure and other aspects of interlinear glossed text. We show how Kratylos can target mixed structures in our FLEx database of Koda that cannot be easily identified within the original FLEx software itself.

CCS Concepts: • **Applied computing** → **Document management and text processing**; • **Information systems** → *Web applications;*

Additional Key Words and Phrases: Language archives, linguistics, interlinear glossed texts, lexicons

ACM Reference format:
Raphael Finkel, Daniel Kaufman, and Ahmed Shamim. 2022. Analyzing Code-mixing in Linguistic Corpora Using Kratylos. *J. Comput. Cult. Herit.* 15, 1, Article 3 (January 2022), 15 pages.
https://doi.org/10.1145/3480238

1 INTRODUCTION

The present article tackles a persistent technical challenge with analyzing text in the field of language documentation: the presence of multilingualism. We approach this challenge by using features of Kratylos, a web-based tool for display and query.

Language documentation is a sub-discipline of linguistics whose ambit includes the collection, annotation, archiving, and presentation of primary language data [32, 33, 52]. Language documentation has taken a key position in the field since the early 1990s due in large part to the recognition that linguistic diversity is under severe threat [14, 19, 26, 45, inter alia]. Hand in hand with the development of new methods and technologies to document languages, a number of digital archives have emerged as institutional repositories for the output of

The programming and field work reported here are supported by NSF DEL Grant #1500753 to Raphael Finkel and Daniel Kaufman. Ahmed Shamim's field work on Koda was also supported by the Endangered Language Initiative at the CUNY Graduate Center.

Authors' addresses: R. Finkel, Department of Computer Science, University of Kentucky, Lexington Kentucky 40506, USA; email: raphael@cs.uky.edu; D. Kaufman, Department of Linguistics, Queens College, Flushing, NY 11367, USA, and Endangered Language Alliance, 3 West 18th Street, 6th Fl, New York New York 10011, USA; email: dkaufman@qc.cuny.edu; A. Shamim, Department of Asian Studies, University of Texas at Austin, 120 Inner Campus Drive, Austin Texas 78712; email: ahmedshamim@utexas.edu.

1556-4673/2022/01-ART3 $15.00
https://doi.org/10.1145/3480238

language documentation, the most prominent of which are ELAR (elar.soas.ac.uk), DOBES (dobes.mpi.nl/), AILLA (ailla.utexas.org), Paradisec (paradisec.org.au/), and Pangloss (pangloss.cnrs.fr) (see References [8, 28, 43] for the state of the art). Linguist fieldworkers, who are in many cases describing languages for the first time, commonly employ computerized tools to build lexicons and interlinear glossed text (henceforth IGT) corpora for documentary purposes and to provide the evidentiary basis for descriptive and theoretical work. An example of one line of IGT is shown in (1) from Wakhi, an Iranic language spoken in the Pamir region. Each line of IGT is called a *tier*. The first tier in (1) is a phonological representation of the language segmented by morpheme, the smallest unit of meaning. Following the Leipzig glossing conventions, affixes, which form part of the word, are indicated with a dash, and clitics, which attach to the outer edge of words, are indicated by an equals sign. The second tier contains a morpheme-by-morpheme gloss, thus nə=ra-t in the segmented transcription corresponds to a negation-marking clitic, glossed as NEG=, followed by a root meaning ‘give’, which takes a past tense suffix glossed as -PST. Finally, the third tier is a free translation into English and a citation linking the example to the primary data,

(1)	hetʃ	kuj	hetʃ	ʧiz	izzat	nə=kər-t	kuj	j-a-r	ʂapik	nə=ra-t
	none	who	none	what	respect	NEG=do-PST	who	DEM-MED-DAT	bread	NEG=give-PST

‘Nobody respected him, no one gave him bread.’ (ELA kimpir dyor, WBL_2017_04_23b, 1.5)

The three pillars of language description, often referred to as the “Boasian trilogy” after the pioneering linguist-anthropologist Franz Boas, consist of a text collection, optimally formatted as in (0), a dictionary, and a descriptive grammar that clarifies the rules and patterns of the language’s sound system, word-building processes, and syntax. The classic descriptive enterprise, especially in the American structuralist tradition, centers around the linguist extracting the rules and patterns of a language based on an extensive text collection (typically transcribed from oral speech). Although this work was carried out manually in Boas’s time, nearly all linguists today employ software to build a lexicon and text corpus. Despite considerable technological advances since the early 2000s, the electronic infrastructure linguists employ for these purposes is still very much geared toward idealized monolingual speech events. Consequently, there is no easy way to harness the computational power of corpus software to examine code-switching, code-mixing, and other effects of contact. Multilingualism thus gets “flattened out” in the annotation process. It may be identifiable to speakers of the language and to the linguist, but it is not indexed in the digital record and is therefore invisible for computer-aided analysis. This problem is not marginal; language mixing is in fact the norm in modern language-documentation scenarios. The modern literature on language contact, beginning with Weinreich [50] and continuing onwards [1, 31, 39, 48], shows the enormous breadth of the phenomenon throughout the world. The intense multilingualism of certain regions, for instance, the Bainounk area of Senegal as described by Lüpke [38] to take but one example, casts doubt on whether certain documentary records can even be sensibly “flattened” to begin with.[1] Certain attempts have been made to accommodate the annotation of multilingual speech in popular programs such as ELAN [16], but none of these have been widely adopted, as seen in the persistence of manual language tagging using ad hoc methods (e.g., Reference [49]).

There is a good deal of work on code-mixing in the computational linguistics literature, but the needs of field linguists, in many cases, working on underdescribed and endangered languages, differ greatly from the goals of current work in computational linguistics. The latter field is concerned with topics such as language identification, named entity recognition, sentiment analysis, dependency parsing, and other tasks that often have clear industrial applications. Linguists, however, are interested in extracting meaningful generalizations about the structure of a language from a universalist perspective, something of primary interest to linguistic

[1] Traditional views of language description and purist language ideologies occasionally lead both linguists and language community members to adhere to a single code for the purposes of documentation. As multilingualism increasingly becomes a focus of study itself and as technology develops to accommodate its analysis, the pretense of monolingualism will likely fade in future work in language documentation.

science but of little current significance to industry. Thus, few of the tools developed around code-mixing in computational linguistics are relevant to either descriptive or theoretical linguists.[2]

In addition to technical problems, intense multilingualism and code-mixing also raise ontological problems. Cysouw and Good [17] and Good [23] have addressed the core problem of how to define the object of language documentation when a community makes regular use of a diverse communicative repertoire. Their suggestions open up new ways of conceptualizing "language community" and language documentation, for instance, by making the target of documentation individuals across multilingual contexts rather than linguistic events chosen to capture a particular language.

Here, we focus narrowly on just one of several technical obstacles in the path toward true multilingual documentation as envisioned by Good [23]: the annotation and querying of interlinear glossed text tagged by source language. We propose a method for handling multilingual language documentation using common open source software tools. The first is Fieldworks Language Explorer (FLEx), described further below, and the second is **Kratylos**, a web-based corpus tool designed by the first two authors of the present article.

The article is structured as follows. In Section 2, we discuss the current digital tools employed by field linguists for corpus building. Section 3 then presents some basic features of Kratylos. Section 4 describes Kratylos's scheme for storage of data and metadata. Section 5 focuses on a case study involving code-mixing in the endangered Munda language, Koda (ISO 639-3 [cdz], Glottocode: koda1236), which has been heavily influenced by Bangla. Section 6 includes thoughts about multilingual corpora and further directions.

2 DIGITAL TOOLS AND FORMATS FOR LANGUAGE DOCUMENTATION

Language documentation data comprise audio and video recordings, PDF files containing texts and analyses, images, lexicons, and IGT in various formats. Common IGT formats include the file types produced by FLEx [10] (FlexText and LIFT XML format), ELAN [51] (eaf XML format), and Praat [12] (TextGrid format); earlier work made heavy use of Shoebox (software.sil.org/shoebox/) and its later reincarnation, Toolbox (`software.sil.org/toolbox/`), for lexicons and IGT corpora in addition to Transcriber [6] for time-aligned annotations.[3] There are also various custom XML formats, such as that designed for the Pangloss archive [40]. Predictably, a problem has arisen with accepting the increasing number of formats and displaying them within the context of a digital archive. Beyond displaying audio and video media, few language archives have facilities for viewing and querying data within a browser [35]. Kratylos presents various formats in a unified manner and provides a means for querying the diverse data types mentioned above from a single interface.[4]

There are several clear desiderata for corpus software to support linguists engaged in descriptive, typological, or theoretical work. Bouda and Helmbrecht [13] propose several design maxims for evaluating linguist-oriented corpus software:

i. Search results should be presented as interlinear text.
ii. The user should be able to find the source utterance in its context in the original file from the search result.
iii. The user should be able to search on all tiers.
iv. Relationships among search terms:
  (a) It should be possible to define relationships among search terms on one tier.

[2] It is indicative that a recent, hefty handbook on linguistic annotation for corpus and computational linguistics [34] contains no real discussion of annotating corpora containing code-switching or code-mixing.

[3] For building corpora and lexicons, the open source FLEx software is of particular importance to the field of language documentation due to its sophistication and popularity. The most recent stable version has been installed 4,404 times and appears to have at least 3,000 active users (Jason Naylor, p.c. 11/13/19). FLEx databases can be stored at `languagedepot.org`, a cloud-based repository that allows users to synchronize a project with collaborators. Among those, 576 projects have been active within the last two years (Christopher Hirt, p.c. 11/17/19).

[4] Although Kratylos presents diverse data types in a unified way, it does not aim to create a universal interchange format for different programs and frameworks. For such proposals, see SALT [53] and the Poio API [11], among others.

(b) It should be possible to define relationships among search terms on different tiers.

v. The user should be able to search within search results.
vi. Search should be possible in a set of files, not only in one file. The more file formats supported, the better.
vii. The user should be able to search for substrings in annotations and use regular expressions.
viii. The user should be confronted with few dialogs and windows during a search task.
ix. It should be possible to export searches and search results, to save and archive them for later reference.

Although Kratylos satisfies all of these desiderata (except, for the time being, v), our primary focus here is (iii) in its relation to multilingual corpora.

The audience for language documentation materials is diverse, and different users approach these materials with very different goals. Some may be concerned with structural aspects of a language; others might be interested in hearing elders from their community or learning their heritage language. The major challenge of any user interface to such material is to be flexible enough to satisfy various academic and cultural goals equally. We now introduce Kratylos, highlighting several research-oriented features as well as those functions geared toward making language documentation more accessible to the community at large.

3 BASIC FEATURES OF KRATYLOS

Kratylos [35], available at `www.kratylos.org`, is a web-facing application that standardizes lexicons and IGT formats through a suite of importer modules and provides an interface so researchers can search the imported data and perform complex queries in the browser. Kratylos is designed to help linguists uncover patterns in richly annotated texts and to easily export such examples to manuscripts as well as to disseminate their work to the public. The program allows queries that are not easily achievable within currently available fieldwork-based corpus software.

Although not designed as an archiving tool, Kratylos can be a valuable adjunct to archives. We have built bulk importers that acquire data from several of the archives mentioned above and import the data into Kratylos. As discussed below, Kratylos adds value to the archives by making the data directly searchable for the first time without ancillary software as well as providing display functions.

3.1 Levels of Access

Digital language archives must contend with complex issues surrounding who can view what, because many archival deposits of endangered languages contain sensitive material from societies with distinct notions of privacy and access [44]. Anyone can make full use of Kratylos's features without registering, but registering allows users to upload their own projects and thus become a "project maintainer." Project maintainers can grant access to particular registered users to view and comment on their projects. There are thus three levels of access that a project maintainer can bestow on a dataset. The data can be (i) completely private, visible only to the project manager, (ii) shared with selected registered users, or (iii) completely public to all registered and unregistered users. Some datasets within a project may be public at the same time that others are more restricted.[5]

3.2 Commenting and Editing

Crowdsourcing input on language documentation projects remains an elusive goal but is part of a strong movement toward "participatory archiving" on behalf of all types of cultural institutions. In the context of a linguistic corpus, creating a public space for feedback on particular entries within a text or lexicon provides a way to involve the language community more deeply while enriching the materials with additional information (Reference [8], pp. 351–353). In practice, crowdsourcing may not be feasible on a grand scale. Kratylos does make

[5] The responsibilities of an intermediary repository when hosting material from an archive are complex. Our approach is to only host archival deposits that are unrestricted and to refer users to the archival deposit's URL for full terms of its use.

Table 1. Stated Location of Primary Data in Descriptive Grammars and Dissertations [22, p.174]

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

a step in this direction by allowing two types of feedback. Registered users can send a message to the project maintainer with an automatic reference to a particular entry. This message is sent as an email with the user's comment together with an attached image of the entry. Collaborators and maintainers can also annotate entries with text, images, and sound files. Once an annotation is saved, it is linked to the entry and can be viewed by all users. Collaborators and maintainers can also edit entries directly. However, Kratylos was explicitly designed for display, query, and export, rather than archiving or manipulation of original annotation files. Thus, any changes made to entries are only made to Kratylos's own database and cannot in general be exported back into original corpus or annotation software.

3.3 Citations and Export

There has been a continued focus in the field of language documentation on improving the transparency between description and analysis, on one hand, and the primary data, on the other hand [4, 7, 21, 22]. Gawne et al.'s (2017) study of 50 dissertations and 50 published descriptive grammars shows just how rarely authors make their full field data available either online or in print, as seen in Table 1.

Similarly, only a very small number of published grammars and dissertations cite linguistic data in such a way that it can be traced easily back to the source documentation (i.e., an annotated recording). The Austin Principles of data citation in linguistics [7] posit eight points to improve current, sub-optimal norms.[6] Kratylos addresses the fourth and fifth principles most directly:

4. **Unique Identification:** A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.
5. **Access:** Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

Kratylos converts utterances and annotations from various formats into entries with a unique URL. This URL is automatically included in exports from Kratylos into a manuscript or webpage, facilitating the linking recommended above. The original media, metadata, and ancillary information can all be made available by the user in Kratylos.

3.4 Customized Keyboards

If the author of a project follows the convention of capitalizing functional morphology [37], then Kratylos automatically extracts the glosses of all functional morphemes in the project and creates a custom pop-up keyboard containing them to facilitate morphological searches. This keyboard incidentally gives a useful snapshot of a language's morphological inventory. In one glance, a user can see whether aspect, tense, agreement, and so on, are

[6] See `https://site.uit.no/linguisticsdatacitation/austinprinciples/`.

marked morphologically, whether the language has marked case, and what cases are present. Similarly, Kratylos automatically constructs a keyboard containing all non-standard orthographic characters used by a project so that users need not copy and paste IPA symbols or other Unicode characters.

3.5 Subtitles

In typical archival deposits of modern language documentation projects, a bundle representing a recording session contains at least a media file (audio or video) and, optimally, a time-aligned annotation. Linguists who are familiar with formats and software can easily download and open these files in open source tools such as ELAN. However, for most non-linguist users, the annotation file is opaque and of little value. Kratylos transforms the annotation files into subtitles that are visible when users view videos in the browser; the subtitles typically include transcriptions and translations. This function lets the members of a language community more fully employ the materials in their language, bringing together familiar media with difficult-to-access annotations.

4 METADATA, DATA, AND TEMPLATES

Each dataset uploaded into Kratylos is identified by a language and a title. It is associated with metadata such as datatype (such as FlexText, ELAN eaf, or Pangloss), provenance (genre, topic, participants, recording date, location, researcher), maintainer, and timestamps, which Kratylos stores in a mySql database keyed by the language and title. The bulk archive importers derive provenance metadata directly from the archives. Manual import by researchers explicitly elicits provenance information, and project maintainers may update that information later.

Kratylos stores the linguistic data in Unix directories named by language and beneath that, by title. Within that dataset directory, Kratylos stores the original data files (there could be several in a dataset) and an XML file resulting from converting the uploaded data files if necessary.[7] The resulting XML format is good for data transmission and storage [9] but less appropriate for searching and browsing. It typically contains much information that is irrelevant for practical purposes and that impedes querying and display. We therefore convert the XML into the lightweight internal format used by the Qddb database software [29, 30]; it is this format that Kratylos searches and displays.

The data comprise individual units of interest, which Kratylos calls *entries*. For a lexicon, an entry is a lexeme and its definitions. For text, an entry is an utterance demarcated by terminal punctuation (in FlexText) or a span identified by a time-aligned media indicator, along with any translation and glossing that the researcher has provided.

Kratylos uses a datatype-specific *template* file to coordinate three parallel representations of the data: XML, Qddb, and CSS. Kratylos has standard templates for some datatypes, such as FlexText, but datatypes without a standardized structure, such as Toolbox, require a specialized, per-dataset template. Figure 1 shows the template used for datasets of the XIGT XML datatype [24]. The + symbol at the end of a line indicates that it continues on the next line. Each line in the template has three components separated by the # symbol. Indentation is significant.

The first component is an XML XPath expression. Here, for instance, the expression `xigt-corpus/igt/tier[@state="normalized" or @type='odin-clean']` splits the XML file into individual entries. Within each entry, the indented XPath expressions identify tiers.

The second component of each line names the tier identified by the first component. The reserved word `TUPLE` indicates an entry boundary; the other tiers in this example are `Source`, `Text`, `Gloss`, `Translation`, and `Annotation`.

The third component of each line provides optional CSS information for HTML display. It also includes reserved words such as `GLOSS`, which introduces small caps where appropriate in the output, and `VERTICAL`, which

[7] Some datatypes, like FlexText and Pangloss, require no conversion; others, like ELAN eaf files, require rearrangement of the XML data, and others, like Toolbox, are not XML at all and require parsing to build an XML file.

```
xigt-corpus/igt/tier[@state="normalized" or @Type='odin-clean'] # TUPLE
    ../../metadata/meta/*[local-name()='source'] # Source
    item[substring(@tag,1,1)='L'] # Text # font-family:monospace;font-style:italic;+
        white-space:pre-wrap;
    item[substring(@tag,1,1)='G'] # Gloss # GLOSS;font-family:monospace;+
        white-space:pre-wrap;
    item[substring(@tag,1,1)='T'] # Translation # font-family:serif;
    annotation # Annotation # color:#006633; VERTICAL
```

Fig. 1. Template for XIGT.

causes multiple instances of a given tier (such as citation forms employing multiple writing systems, discussed below) to appear vertically aligned instead of the default horizontal alignment.

We now turn our attention to a case study that addresses the problem we outlined earlier: the annotation of multilingual field recordings that employ code-switching, code-mixing, and other contact effects.

5 CASE STUDY: KODA LANGUAGE CONTACT

Koda[8] is an endangered Munda language spoken in the Rajshahi district in Bangladesh and across its western border in India. The Munda languages comprise a subgroup of the larger Austroasiatic family, which span from eastern India to southeast Asia. In Bangladesh and India, Munda-speaking ethnic groups, including the Koda, are classified as *adivasi*, or indigenous. Within Munda, Koda is part of the Kherwarian subgroup, together with Santali, Mundari, Ho, and several neighboring languages with fewer speakers. An overview of the entire Munda subgroup is found in Anderson [2]; Kim et al. [36] provide a sociolinguistic sketch of the Santali cluster, including Koda. Koda is one of the least described Munda varieties in the region, having merited brief mention in Grierson [25] but barely noticed elsewhere. Koda does, however, bear a strong resemblance to Ho, which has been described in some detail [3, 15, 46]. A description of the phonology, morphology and basic syntax of Koda is the topic of the third author's dissertation [47], based on original fieldwork.

Bangla has strongly influenced Koda as a result of long-term contact, more so, perhaps, than neighboring Kherwarian languages, which also show strong signs of contact. "Language contact" covers a range of phenomena that are often treated individually. Code-switching typically refers to switching languages across sentences; code-mixing refers to switching within units smaller than the sentence (i.e., the phrase and the word). Borrowing refers to the full adoption of a word (or phrase) from one language into another. In practice, it is often difficult to distinguish these phenomena from one another; Koda is no exception. In (2), the words in blue are of Bangla origin; the others are native Koda. The sentence alternates between words of Koda and Bangla origin. It is not at all clear from casual inspection what type of contact phenomena we are witnessing here,

(2) buluŋ c^{h}ara utu bɛʃ ka laga
salt without curry good NEG taste
'Without salt, curry doesn't taste good.' (Phonemes Distribution framed sentences read by Shohag, CDZ_2015_05_12b, 354)

Contact effects shed light on two distinct domains. First, they inform us about the intense nature of social relations between two groups, in this case, the Munda ancestors of the Koda and their Indo-Aryan neighbors. Second, each case of language contact further clarifies the proper limits of contact phenomena, which are crucial for developing formal models of multilingualism. A major strand of research in this area seeks to understand what areas of grammar are more and less liable to be altered through contact and to what extent the typology of the languages involved dictates their possible interactions. In the remainder of the article, we focus on the distribution of Bangla lexemes in Koda and how Kratylos can help us find patterns in a mixed language.

[8] ISO 639-3 identifier *cdz*.

5.1 The Bangla Element in Koda

There is general consensus that the elements most likely to cross languages include the categories of nouns and conjunctions (e.g., ‘but’ and ‘or’) [39]. On the other end of the spectrum, there is also general consensus that bound, inflectional morphology, such as subject-agreement affixes and case markers on noun phrases are among the rarest elements to be transferred across languages.

Koda is of special typological interest here. With regard to case marking (e.g., affixes that indicate the grammatical role of a noun phrase as a subject, object, etc.), Matras [39, p. 215] concludes that “Evidence of cross-linguistic diffusion of productive nominal inflection is extremely rare,” noting that few examples of case borrowing have come to light, since Heath [27] documented this phenomenon among the Aboriginal languages of Arnhem Land.

However, Koda provides robust evidence for the reality of case-marker transfer. In (3), we see a Koda sentence (a) and its Bangla equivalent (b). We see that the noun-phrase morphology and the verb stem itself are of Bangla origin.

(3) a. daʔa-ʈa hɔn-ʈa-kɛ kʰawa-iʔ=m
water-DEF kid-DEF-OBJ feed-3SG.ANM.OBJ-2SG.SBJ
‘Help the kid drink the water.’ (ELICITATION IX 204.1)

b. pani-ʈa chele-ʈa-kɛ kʰawa-o
water-DEF kid-CL-OBJ feed-2SG.PRES
‘Help the kid drink the water.’

The objective case marker -kɛ, as well as other Bangla-origin case markers, are employed so regularly that they can no longer be considered to result from code-mixing. They are part and parcel of the Koda language. In addition to such instances of borrowing, we find code-mixing within the word. This observation is also of special interest in Koda, because Indo-Aryan and Munda languages have radically different ways of categorizing roots and words. Indo-Aryan languages have a somewhat strict categorization into the familiar cardinal categories of verb, noun, adjective and adposition. Kherwarian languages such as Mundari, Santali, and Koda, however, have very weak noun-verb distinctions at the root level. Lexical roots appear to be largely acategorial, taking canonical noun morphology when functioning as subjects, objects or other types of arguments, but taking canonical verbal morphology when functioning as predicates (see Evans and Osada [20] for the controversy this fact has engendered).

This divergence raises the question of what happens when Bangla-categorial (verbal or nominal) roots are incorporated into a largely acategorial system. A similar question is whether Koda roots can be incorporated into Bangla morphological templates and would they show the same flexibility. To approach these questions, we first introduce our strategy for handling multilingualism in a single corpus.

5.2 Unmixing the Codes in FLEx and Kratylos

Kratylos has a complex template for ingesting FlexText XML, a portion of which is shown in Figure 2. This template introduces a `Word` tier, which itself contains sub-tiers `Text` (the original unsegmented text), `POS` (part of speech), and `Morpheme` (the corresponding allomorph, as listed in the lexicon). In turn, the `Morpheme` tier contains sub-tiers of its own. The XML data may contain multiple instances of each of these tiers and sub-tiers within a single entry.

Figure 3 shows how Kratylos displays a formatted entry from our Koda corpus. At the level of the full entry, `Source` and `Segnum` indicate the provenance (CDZ_2015_05_07b) and the segment number (2) of the example. At the word level, the `Text` tier contains an orthographic representation (ɟɔtɔ), and `POS` indicates the word’s part of speech (adj). At the morpheme level, the `Morph` tier contains the allomorph/surface form (ɟɔtɔ), and `Citation` contains the citation form(s) as found in the lexicon (ɟɔtɔ).

```
words/word # Word ( #
   item[@type="txt"] # Text
   item[@type="pos"] # POS
   morphemes/morph # Morpheme (
      item[@type="txt"] # Morphx (
         @lang # MLang #
         text() # Morph
      item[@type="cf"] # Citation
      item[@type="msa"] # MSA
      item[@type="gls"] # Gloss
```

Fig. 2. Template fragment for FlexText.

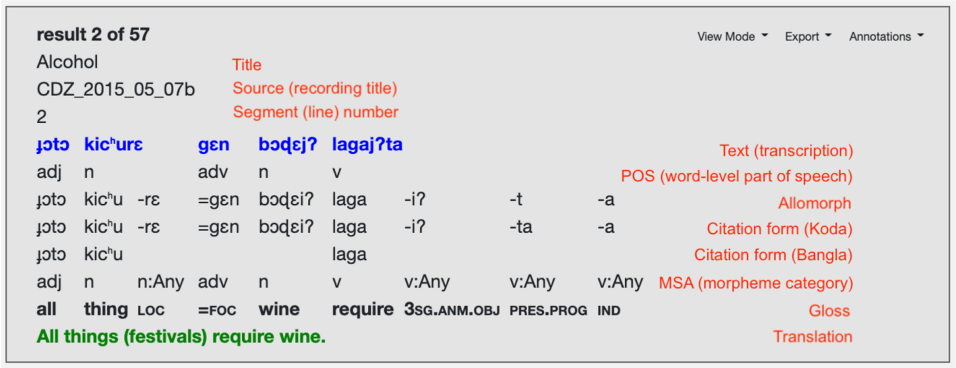
Crucially, FLEx allows users to employ multiple "writing systems" in the analysis of a single language. "Writing systems" need not employ different scripts; they can also be employed to represent a language in multiple orthographies using a single script. If a FLEx project contains multiple writing systems, then Kratylos creates vertically aligned citation tiers for each one. In this case, the linguist has opted to represent both the native Koda vocabulary as well as the Bangla vocabulary using an IPA-based orthography. Although FLEx provides ample fields for etymological and source-language information in the lexicon, it does not make this information visible to the parser or the user in its corpus component. FLEx is able, however, to display multiple writing systems in the corpus component; we have exploited this feature to allow us to query code-mixing in Kratylos. However, FLEx, in its current version, cannot parse words across writing systems in a single text. For this reason, we have encoded all morphemes uttered in a "Koda text" in the Koda writing system, regardless of etymology. However, we employ the Bangla (IPA) writing system solely for morphemes with a Bangla provenance. Thus, a native Koda word only has one representation, in the Koda writing system, whereas a Bangla word in a Koda text has a representation in both the Koda and Bangla writing systems. This strategy allows FLEx's parser to operate over the entire text while maintaining a precise representation of code-mixing that is visible to Kratylos's search function. This representation moves toward documenting the speech of individuals as opposed to idealizations of a "target language" [23]. Our "Koda lexicon" is in fact a lexicon of all the words uttered by our Koda collaborators in their recordings regardless of etymology. Nonetheless, we have not lost the ability to extract the native (Munda) vocabulary from this mixed database. It is trivial to exclude lexemes with an additional Bangla representation for this purpose.

Because loan-words from Bangla can follow Koda phonology, at least in the examples given here, the Bangla-origin morphemes are generally identical in the Koda and Bangla writing systems. Hence, ɟɔtɔ, kic^{h}u, and laga also appear on the Bangla `Citation` tier. The `MSA` tier indicates categorial and selectional information on the morpheme level. For instance, the morphemes -iʔ, -ta and -a are of the category v:Any, indicating they can attach to any verbal stem. The `Gloss` tier (**all**) shows the meaning or function of each morpheme. Finally, the `Translation` tier (**All things ...**) applies to the entire entry.

A standard IGT representation of the example as produced by Kratylos's LaTeX export function is given in (4). We have highlighted (in blue) Bangla elements, as represented in a second `Citation` tier (the fourth tier from the bottom in Figure 3). The Kratylos export function automatically creates a unique URL for the entry in Kratylos (in this case linked to the source's filename in parenthesis), which allows for verification against the original data.

(4)	ɟɔtɔ	kic^{h}u-rɛ=gɛn	bɔɖɛiʔ	laga-i-t-a
	all	thing-LOC=FOC	wine	require-3SG.ANM.OBJ-PRES.PROG-IND

'All things (festivals) require wine.' (Alcohol, CDZ_2015_05_07b, 2)



result 2 of 57

Alcohol — Title
CDZ_2015_05_07b — Source (recording title)
2 — Segment (line) number

ɟɔtɔ	kicʰurɛ		gɛn	bɔɖɛjʔ	lagajʔta				Text (transcription)
adj	n		adv	n	v				POS (word-level part of speech)
ɟɔtɔ	kicʰu	-rɛ	=gɛn	bɔɖɛiʔ	laga	-iʔ	-t	-a	Allomorph
ɟɔtɔ	kicʰu	-rɛ	=gɛn	bɔɖɛiʔ	laga	-iʔ	-ta	-a	Citation form (Koda)
ɟɔtɔ	kicʰu				laga				Citation form (Bangla)
adj	n	n:Any	adv	n	v	v:Any	v:Any	v:Any	MSA (morpheme category)
all	**thing**	LOC	=FOC	**wine**	**require**	3SG.ANM.OBJ	PRES.PROG	IND	Gloss

All things (festivals) require wine. — Translation

Fig. 3. Formatted Koda example.

```
<Word*
    <Morpheme
        <Citation \S>                               (Koda)
        <Citation \S>                             (Bangla)
        <MSA v$>>
    <Morpheme
        <Citation ->                                (Koda)
        <Citation $>                              (Bangla)
        <MSA v:Any>>>
```

Fig. 4. Query for Bangla stem with Koda suffix.

5.3 Querying in Kratylos

Kratylos allows query types of various complexity, including searches for a simple string, a regular expression, and a Boolean combination of regular expressions. Here, we are interested in multi-tier regular-expression search, which allows searching for attributes across different levels of analysis in the IGT. As shown below, this search method can help us distinguish between morphemes with or without a representation in a particular writing system. It is notably more powerful than FLEx's equivalent search function.

A multi-tier query is composed of nested units. A unit has the form `<tierName content>`. The content can be a Perl regular expression, a nested tier, or empty. The character `*` after the tier name means "any subsequent instance of this tier."

Figure 4 presents a multi-tier query that searches for any word in an entry (`<Word*>`) containing a non-whitespace character (`\S`) in both Koda and Bangla citation forms (i.e., a Bangla-origin morpheme), followed by a suffixal morpheme (a morpheme beginning with a dash) without a Bangla citation form. Empty cells in the IGT are matched by `$`, the regular expression for the end of a line.[9] The query further specifies that the category (MSA) of the first morpheme must be verbal (tagged in FLEx as `v`), and the second morpheme must be a verbal inflection that attaches to any type of verbal host (tagged in FLEx as `v:Any`).[10]

This query searches for all entries in which native verbal morphology attaches to a Bangla stem, as we saw in (4) with the Bangla verb root laga and the Koda suffixes that follow. Conversely, we can search for a Koda

[9] Because the Perl patterns are anchored to the beginning of the line (with the implicit regular-expression symbol `^`), the resulting query, `^$`, only matches empty cells.

[10] The regular expression `v$` is matched by the beginning of a line (implicit), followed directly by `v`, followed by the end of the field and thus matches the abbreviation for 'verb' in the category tier.

```
<Word*
    <Morpheme
        <Citation \S>                                  (Koda)
        <Citation $>>                                (Bangla)
    <Morpheme
        <Citation ->                                   (Koda)
        <Citation ->>>                               (Bangla)
```

Fig. 5. Query for Koda stem with Bangla suffix.

morpheme hosting a Bangla suffix, as shown in Figure 5. Here, the query targets a morpheme with a non-empty (\S) Koda citation form and an empty ($) Bangla citation form (i.e., a native Koda morpheme) followed by a suffix that has content in both writing systems (i.e., a morpheme of Bangla provenance).

This query in Figure 4 finds examples like (5), where the native root bɔɖɛiʔ takes a Bangla suffix -ʈa (highlighted in blue, as above),

(5) tac^{h}ara bɔɖɛiʔ-ʈa abu-a pɔd^{h}an hiʃabɛ cala-ɔʔ-ta-a
besides wine-DEF 1PL.INC-GEN.PRON main as move-INTR.A-PRES.PROG-IND
'Besides, the wine remains as our main (thing).' (Alcohol, CDZ_2015_05_07b, 6)

The query in Figure 4 yields over 300 results in the Koda corpus, showing that Bangla verb roots happily host Koda functional morphology, as seen with cala-ɔʔ-ta-a above. It is especially surprising that some Bangla auxiliaries are also incorporated into Koda morphological templates, as evidenced by hui-ak-a in (6). Functional items such as inflected auxiliaries are part of the grammatical glue that is most resistant to borrowing and change.

(6) hɔn-ʈa mɔrɛ-a hui-ak-a
child-DEF.SG five-CLF be-PRF.INTR-IND
'The child turned five.' [47]

The query in Figure 5, looking for Koda verb stems hosting Bangla functional morphology, finds no results at all. There are, in fact, very few *verbal* affixes of Bangla origin in the Koda corpus at all. On the other hand, there is a large number of Bangla-origin *nominal* suffixes: -kɛ OBJECTIVE, -(t)ɛ LOCATIVE, -ɛr GENITIVE, -ʈa CLASSIFIER, -ra PLURAL. Surprisingly, these attach both to Bangla-origin nouns as well as to native Koda nouns. The Bangla noun-phrase markers have been so thoroughly integrated into the language that they can no longer be considered the result of creative code-mixing; native speakers consider them obligatory in most contexts under elicitation (although they are occasionally omitted in discourse). The sentence in (7) is a typical example of an utterance comprised entirely of Koda-origin morphemes except for the case marker on the object.

(7) iŋ am-kɛ raʔa-tɛ-m-a=jŋ
1SG 2SG-OBJ call-AOR.TR-2SG.OBJ-IND=1SG.SBJ
'I called you.' (VALENCY I, CDZ_2013_02_22b, 29)

Searching for the longest strings of Bangla morphemes in the corpus we find that the verbal suffixes remain staunchly native, even when the entire preceding sentence is of Bangla origin, as in (8).

(8) tac^{h}ara puɟa-ʈa hɔ bɛʃ ʃundɔr hui-ak-a
besides worship-DEF EMPH very beautiful be-PRF.INTR-IND
'Besides, the puja was very beautiful.' (Shohag thanks Arun for the Puja, CDZ_2013_02_22j, 16)

There is a striking parallel here to the fascinating case of "language intertwining" described for Michif [5], a mixed language based on Cree and French and spoken by the Métis people of Canada. Michif displays a surprisingly strict separation according to lexical category; to a very large extent, nouns have a French origin and verbs (with all their associated morphology) come from Cree, as exemplified in (9), where French-origin morphemes are in blue.[11]

(9)	li	bɔn	mark	a	lıkɔl	mIšI-mIjœstam-Ih-Iko-w
	DAP	good.F	mark	at	school	big-be.glad-CAUS-INV-3

‘Good marks in school make him very happy.’ [5, p.112]

We also find “Lexicon-Grammar mixed languages,” in which the language’s functional glue (i.e., markers of case, voice, tense, aspect, valency, mood, etc.) come from one language, but the lexical material (nouns, verbs, adjectives and, to a certain extent, adpositions) come from another language. Such is the case of Media Lengua [41, 42], a language whose lexical material is largely Spanish but whose grammatical/functional morphology is almost entirely Kichwa, as can be seen in (10), where Spanish morphemes are in blue.

(10)	Mujer-ca	semilla-cuna-ta	tierra-wan	cubri-n.
	woman-TOP	seed-PL-ACC	earth-with	cover-3SG

‘The woman covers the seeds with the earth.’ [18, p.407]

With the help of Kratylos, we have uncovered a mixed-language pattern that appears to be a cross between the Michif noun-verb opposition and the Media Lengua lexical-grammatical opposition. In Koda, both verb stems and functional morphology can originate in Bangla. However, there is a strict separation between the functional morphology of the noun and verb: Verbal morphology is native to Koda, but nominal morphology is overwhelmingly Bangla. This separation is further exemplified in (11), where the noun raɟa and its two functional suffixes are Bangla while the following verb and all its suffixes are native Koda.

(11)	tɛ	tikin	dɔl	nimtɔʔ	raɟa-ʈa-kɛ	rɛŋgɛjʔ-t-iʔ-a
	SO	noon	towards	now	king-DEF-OBJ	hunger-AOR.TR-3SG.ANM.OBJ-IND

‘Then at noon the king was hungry.’ (Love and salt, CDZ_2013_02_22l, 24)

We leave a fuller discussion of these facts for another occasion. We hope that these examples demonstrate the utility of Kratylos in teasing out such patterns from a complex dataset.

6 CONCLUSION

Multilingual texts that involve mixing, switching, and borrowing are uniquely valuable to our understanding of language both in the mind and in society. Computer-aided analysis of such texts requires, at the very least, morphemic tagging by language, for which there are many options. However, a real challenge presents itself for linguists concurrently building a lexicon and concordance, because the linguist must distinguish among borrowing, mixing, and switching. Borrowed elements that have become obligatory to express some function or meaning, such as Bangla-origin -kɛ in Koda, are included as a bona fide part of the language. However, running discourse in the contact language or any other “non-target” language is generally left unanalyzed in an IGT corpus and excluded from the lexicon. But there is a large gray area in any multilingual setting between clear borrowings and active instances of mixing and switching. We advocate for a treatment that approaches Good’s [23] more holistic vision of documenting the speech of an individual or community rather than a single target code. The “target language” includes all linguistic material without regard to named language; identifiable “non-target languages” are separated out by an additional representation in their own tier or writing system. We have demonstrated an implementation of this idea in FLEx and Kratylos and exploited it to discover generalizations about the distribution of Bangla-origin morphemes in Koda.

A real examination of language contact in Koda must await further work [47], but we hope to have shown how Kratylos can exceed the query power provided by current tools available to field linguists with regard to

[11] Bakker’s abbreviations for glossing Michif are as follows: DAP, definite article plural; INV, inverse voice.

multilingual text. As a result, a linguist can conjecture and then verify patterns, such as those dealing with code-mixing and switching that would be otherwise hidden in the IGT data structures commonly employed by linguists. Furthermore, Kratylos allows such patterns to be verified and explored by other users through the creation of an online public corpus [35].

In an influential article, Bird and Simons [9] propose “seven dimensions of portability” in language documentation with recommendations for linguists in handling content, format, discovery, citation, rights, access and preservation. Nearly twenty years later, there have been great advances in the shared understanding of best practices but little consensus on how to implement them. The major archives have been slow to evolve because of the cost and institutional commitment involved in replacing infrastructure on a large scale. As a result, it is not much easier to query an archived IGT corpus today than it was a decade ago. Kratylos has the potential to significantly raise current standards for discovery, citation, presentation of content and query power. Issues of access and preservation will always fall within the responsibilities of the archive but, as a web application, Kratylos can be adopted just as easily by a digital archive as by an end user to provide a simple, low-cost solution to the other dimensions of portability.

We continue to develop Kratylos in response to requests from users and invite linguists to browse its public collections and to create their own collections with a free account.

REFERENCES

[1] Alexandra Y. Aikhenvald. 2002. *Language Contact in Amazonia.* Oxford University Press, Oxford.

[2] Gregory D. S. Anderson. 2008. *The Munda Languages.* Routledge, London.

[3] Gregory D. S. Anderson, Toshki Osada, and K. David Harrison. 2008. Ho and the other Kherwarian languages. In *The Munda Languages.* Routledge, Abingdon, 195–255.

[4] Peter K. Austin. 2006. Data and language documentation. In *Essentials of Language Documentation*, Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel (Eds.). De Gruyter, Berlin, 87–112.

[5] Peter Bakker. 1997. *A Language of Our Own : The Genesis of Michif, the Mixed Cree-French Language of the Canadian Metis.* Oxford University Press, New York.

[6] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. 2000. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Commun.* 33, 1–2 (Jan. 2000).

[7] Andrea L. Berez-Kroeker, Helene N. Andreassen, Lauren Gawne, Gary Holton, Susan Smythe Kung, Peter Pulsifer, Lauren B. Collister, the Data Citation and Attribution in Linguistics Group, and the Linguistics Data Interest Group. 2018. *The Austin Principles of Data Citation in Linguistics.*

[8] Andrea L. Berez-Kroeker and Ryan E. Henke. 2018. Language archiving. In *The Oxford Handbook of Endangered Languages*, Kenneth L. Rehg and Lyle Campbell (Eds.). Oxford University Press, Oxford, 347–369.

[9] Steven Bird and Gary Simons. 2003. Seven dimensions of portability for langauge documentation and description. *Language* 79, 3 (2003), 557–582.

[10] Andrew H. Black and Gary F. Simons. 2008. The SIL FieldWorks language explorer approach to morphological parsing. In *Texas Linguistics Society 10: Computational Linguistics for Less-Studied Languages*, Nicholas Gaylord, Stephen Hilderbrand, Heeyoung Lyu, Alexis Palmer, and Elias Ponvert (Eds.). CSLI Publications, 37–55.

[11] J. Blumtritt, P. Bouda, and F. Rau. 2013. Poio API and GraF-XML: A radical stand-off approach in language documentation and language typology. In *Proceedings of Balisage: The Markup Conference 2013.*

[12] Paul Boersma and David Weenink. 2017. Praat: Doing Phonetics by Computer [Computer Program]. Version 6.0.35. Retrieved October 16, 2017 fromewblock http://www.praat.org/.

[13] Peter Bouda and Johannes Helmbrecht. 2012. From corpus to grammar: How DOBES corpora can be exploited for descriptive linguistics. In *Electronic Grammaticography*, Sebastian Nordhoff (Ed.). University of Hawai‘i Press, Honolulu, Hawai‘i, 129–159.

[14] Matthias Brenzinger. 2007. *Language Diversity Endangered.* Mouton de Gruyter, Amsterdam.

[15] Lionel Burrows. 1915. *Ho Grammar with Vocabulary.* Catholic Orphan Press, Calcutta.

[16] Onno Crasborn and Han Sloetjes. 2014. Improving the exploitation of linguistic annotations in ELAN. In *Proceedings of the9th International Conference on Language Resources and Evaluation (LREC’14)*, Nicoletta Calzolari, Khalid Choukri, Hrafn Declerck, Thierry Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA).

[17] Michael Cysouw and Jeff Good. 2013. Languoid, doculect, and glossonym: Formalizing the notion ‘language’. *Lang. Document. Conserv.* 7 (2013), 331–359.

[18] Isabel Deibel. 2019. Adpositions in Media Lengua: Quichua or Spanish? Evidence of a lexical-functional split. *J. Lang. Contact* 12 (2019), 404–439.

[19] Nicholas Evans. 2011. *Dying Words: Endangered Languages and What They Have to Tell Us*. John Wiley & Sons, New Jersey.

[20] Nicholas Evans and Toshki Osada. 2005. Mundari: The myth of a language without word classes. *Linguist. Typol.* 9, 3 (2005), 351–390.

[21] Lauren Gawne and Andrea L. Berez-Kroeker. 2018. Reflections on reproducible research. In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton (Eds.). Language Documentation & Conservation Special Publication, Vol. 15. UH Press, Manoa, Hawai'i, 22–32.

[22] Lauren Gawne, Barbara F. Kelly, Andrea L. Berez-Kroeker, and Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Lang. Document. Conserv.* 11 (2017), 157–189.

[23] Jeff Good. 2018. Reflections on the scope of language documentation. In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton (Eds.). Language Documentation & Conservation Special Publication, Vol. 15. UH Press, Manoa, Hawai'i, 13–21.

[24] Michael Wayne Goodman, Joshua Crowgey, Fei Xia, and Emily M. Bender. 2015. Xigt: Extensible interlinear glossed text for natural language processing. *Lang. Resourc. Eval.* 49, 2 (2015), 455–485.

[25] George A. Grierson. 1906. *Muṇḍā and Dravidian Languages*. Linguistic Survey of India, Vol. IV. Office of the Superintendent of Goverment Printing, Calcutta.

[26] Ken Hale, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne, and Nora C. England. 1992. Endangered languages. *Language* 68, 1 (1992), 1–42.

[27] Jeffrey Heath. 1978. *Linguistic Diffusion in Arnhem Land*. Australian Institute of Aboriginal Studies, Canberra.

[28] Ryan E. Henke and Andrea L. Berez-Kroeker. 2016. A brief history of archiving in language documentation, with an annotated bibliography. *Lang. Document. Conserv.* 10, 411-457 (2016).

[29] Erin H Herrin II and Raphael Finkel. 1991. An ASCII database for fast queries of relatively stable data. *Computing Systems* 4, 2 (1991), 127–155.

[30] Eric H. Herrin II and Raphael A. Finkel. 1995. Schema and tuple trees: An intuitive structure for representing relational data. *Computing Systems* 9 (1995), 93–118.

[31] Raymond Hickey. 2013. *The Handbook of Language Contact*. Wiley, West Sussex, United Kingdom. http://books.google.com/books?id=1fr5t1KLL6oC.

[32] Nikolaus P. Himmelmann. 2006. Language documentation: What is it and what is it good for. In *Essentials of Language Documentation*, Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel (Eds.). Mouton de Gruyter, Berlin and New York, 1–30.

[33] Nikolaus P. Himmelmann. 2008. Reproduction and preservation of linguistic knowledge: Linguistics' response to language endangerment. *Annual Review of Anthropology* 37 (2008), 337–350.

[34] Nancy Ide and James Pustejovsky. 2017. *Handbook of Linguistic Annotation*. Springer, Dordrecht.

[35] Daniel Kaufman and Raphael Finkel. 2018. Kratylos: A tool for sharing interlinearized and lexical data in diverse formats. *Language Documentation & Conservation* 12 (2018), 124–146.

[36] Seung Kim, Amy Kim, Sayed Ahmad, and Mridul Sangma. 2010. *The Santali Cluster in Bangladesh: A Sociolinguistic Survey*. SIL International. SIL Electronic Survey Reports 2010-006 http://www.sil.org/silesr/abstract.asp?ref=2010-006.

[37] Christian Lehmann. 2004. Interlinear morphemic glossing. In *Morphologie. Ein Internationales Handbuch zur Flexion und Wortbildung*, Geert Booij, Christian Lehmann, Joachim Mugdan, and Stavros Skopeteas (Eds.). Handbücher der Sprach- und Kommunikationswissenschaft, Vol. 17. W. de Gruyter, Berlin, 1834–1857.

[38] Friederike Lüpke. 2016. Pure fiction – the interplay of indexical and essentialist ideologies and heterogeneous practices: A view from Agnack. *Language Documentation and Conservation* Special Publication 10 (2016), 8–39.

[39] Yaron Matras. 2009. *Language Contact*. Cambridge University Press, Cambridge. http://books.google.com/books?id=GiTPVVxP7OoC.

[40] Boyd Michailovsky, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre Francois, and Evangelia Adamou. 2014. Documenting and researching endangered languages: The Pangloss Collection. *Language Documentation and Conservation* 8 (2014), 119–135.

[41] Pieter Muysken. 1981. Halfway between quechua and spanish: The case for relexification. In *Historicity and Variation in Creole Studies*, Arnold Highfield and Albert Valdman (Eds.). Karoma, Ann Arbor, 52–78.

[42] Pieter Muysken. 1997. Media lengua. In *Contact languages: A wider perspective*, Sarah G. Thomason (Ed.). John Benjamins, Amsterdam/Philadelphia, 365–426.

[43] David Nathan. 2011. Digital archiving. In *Cambrdige Handbook of Endangered Languages*, Peter K. Austin and Julia Sallabank (Eds.). Cambridge University Press, Cambridge, 187–208.

[44] David Nathan. 2014. Language documentation and description: Access and accessibility at ELAR, an archive for endangered languages documentation. In *Special Issue on Language Documentation and Archiving*, David Nathan and Peter K. Austin (Eds.). Vol. 12. SOAS, London, 187–208.

[45] Daniel Nettle and Suzanne Romaine. 2000. *Vanishing Voices: The Extinction of the World's Languages*. Oxford University Press, Oxford.

[46] Anna Pucilowski. 2013. *Topics in Ho Morphophonology and Morphosyntax*. Ph.D. Dissertation. University of Oregon.

[47] Ahmed Shamim. 2021. A description of the phonology and the morphology of Koda, an endangered language of Bangladesh. Ph.D. Dissertation. Graduate Center, City University of New York, New York.

[48] Sarah G. Thomason. 2001. *Language Contact: An Introduction*. Georgetown University Press, Georgetown.

[49] Ewald Van der westhuizen and Thomas Niesler. 2018. A first South African corpus of multilingual code-switched soap opera speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA), Miyazaki, Japan.

[50] Uriel Weinreich. 1953. *Languages in Contact: Findings and Problems*. Linguistic Circle of New York, New York.

[51] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*. LREC, 1556–1559.

[52] Anthony C. Woodbury. 2003. Defining documentary linguistics. In *Language Documentation and Description*, Peter K. Austin (Ed.). Vol. 1. Hans Rausing Endangered Languages Project, London, 35–51.

[53] Florian Zipser and Laurent Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, LREC 2010*. LREC, La Valette, Malta. https://hal.inria.fr/inria-00527799.

Received September 2020; revised April 2021; accepted July 2021