XML transformation and structure-based pattern matching for endangered languages

Raphael Finkel, Daniel Kaufman, Ahmed Shamim

University of Kentucky, Queens College CUNY & ELA, CUNY Graduate Center
Department of Computer Science, Department of Linguistics, Linguistics Program
raphael@cs.uky.edu, dkaufman@qc.cuny.edu, ahmedshamim@utexas.edu

Abstract

We describe the use of Kratylos, a web-based corpus query tool, to analyze the morphosyntax of multilingual texts. Among other formats, Kratylos accepts interlinear glossed texts in XML as produced through Fieldworks Explorer (FLEX), the corpus/lexicon building software most commonly employed in language documentation and description. Kratylos facilitates multiliter regular expression queries over such XML corpora, which can be used to investigate interactions across languages in a multilingual FLEX corpus, a use case which FLEX itself is not well suited for. We exemplify with a Koda language (cdz) corpus in development that includes a range of borrowing, code mixing and code switching with Bangla.

1. Introduction

There exists a considerable divide between the technology concerns of linguist fieldworkers and those of the NLP community. Fieldworkers, who are in many cases describing languages for the first time, are interested in tools to build lexicons and interlinear glossed text (IGT) corpora for documentary purposes and to allow harvesting examples of grammatical phenomena for descriptive and theoretical work. The NLP community, on the other hand, while interested in various aspects of language use, is not in the business of describing or analyzing linguistic phenomena with an eye towards universal patterns. As a result, interaction between documentary and computational linguists continues to lag despite some prominent efforts to unite them. This misalignment applies especially to the documentation of multilingualism, which has less industrial application at present but is crucial to understanding the actual language use of many speech communities.

The application described here, Kratylos (Kaufman and Finkel, 2018), employs computational methods for purposes of linguistic documentation, description and theory (Himmelmann, 2006; Mosel, 2012; Xia et al., 2016; Bender and Langendoen, 2010). Kratylos, available at kratylos.org, is middleware, designed to aid 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. A user can upload lexical and corpus datasets from the programs most commonly used by fieldworkers, such as FieldWorks (Black and Simons, 2008), Praat (Boersma and Weenink, 2017), ELAN (Wittenburg et al., 2006), and Toolbox, as well as formats used by online archives such as Pangloss (Michaelovsky et al., 2014). Uploaded data can then be browsed and queried. Kratylos allows complex queries that are not easily achievable within the original corpus-building programs.

2. Metadata and templates

The metadata associated with each dataset (provenance, maintainer, timestamps) are stored in an SQL database. The data comprise individual units of interest, which Kratylos calls entries. For a lexicon, an entry is a lexeme. For text, an entry is an utterance demarcated by terminal punctuation (in FLEX) or a span identified by the uppermost tier of a time-aligned annotation.

Kratylos digests annotation files produced by any of the aforementioned software into an internal XML representation suitable for query and display. Some input formats require no modification, such as FLEXTEXT, EBible, and Pangloss. Some are XML but need reorganization, such as ELAN EAF files. Others are not XML at all but are converted to XML, such as TextGrid, ToolBox, and CSV files.

The resulting XML format is good for data transmission and storage (Bird and Simons, 2003) but less appropriate for searching and display. It typically contains much information that is irrelevant for practical purposes and that impedes querying and display. We therefore convert the XML into a lightweight internal format used by Qddb database software (Herrin II and Finkel, 1991a; Herrin II and Finkel, 1991b), and it is this format that Kratylos searches and displays.

Kratylos uses a template file to coordinate three parallel representations of the data: XML, Qddb, and CSV. Kratylos has standard templates for some data formats, such as FLEXTEXT, but data formats without a standardized tier structure, such as ToolBox, require a specialized, per-project template. Figure 1 shows the template used for the XIGT XML (Goodman et al., 2015). 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, 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 tiers identified by the first component. The reserved word TUPLE indicates a entry boundary; the other tiers in this example include Source, Text, and Gloss.
The third component provides optional CSS (Cascading Style Sheet) information for HTML display. It includes reserved words such as GLOSS, which introduces small caps where appropriate in the output and VERTICAL, which 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.

At the morpheme level, the Morph tier contains the allomorph/surface form, and Citation contains the underlying form as found in the lexicon. If the lexicon contains multiple writing systems, Kratylos creates citation tiers for each writing system, discussed further below. The MSA tier indicates categorial and selectional information on the morpheme-level morpheme. The gloss tier shows semantic information. Finally, the Translation tier (in green) applies to the entire entry.

3. Multi-tier regex queries
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 focus on multi-tier regular-expression search for investigating code-switching and mixing in the Koda FLEX database, as developed by the third author. Koda (cdz) is an endangered Munda language spoken in the Rajshahi district in Bangladesh and across the border with India. Bangla has strongly influenced Koda as a result of long-term contact. Multilingual speech has posed a challenge to language documentation, because the methods and software employed by linguists in the field have been primarily designed for monolingual projects. For instance, despite ample fields for etymological and source-language information in the lexicon, FLEX does not make this information visible to the parser or the user in the corpus portion. It does, however, allow the user to employ multiple scripts or writing conventions (“writing systems”) for a single language, and it makes them fully visible as part of the IGT. Our strategy thus involves treating the two target languages, Koda and Bangla, as distinct (IPA-based) “writing systems” in FLEX and designing a provision for searching the two writing systems simultaneously in Kratylos. All morphemes are represented in the Koda writing system, but Bangla morphemes are additionally represented in the Bangla writing system. A standard IGT representation of the example as produced by Kratylos’s LaTeX export function is given in (1). Bangla elements, as represented in a second Citation tier (the third tier from the bottom in Figure 3) are highlighted here. The Kratylos export function automatically creates a URL link to the entry in Kratylos, which allows for verification against the original data.

Figure 1: Template for xigt

We focus here on a particular transformation of FLEX-TEXT XML, the output of the open source Fieldworks Explorer (FLEX) software, which 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). FLEXTEXT requires a complex template, a portion of which is shown in Figure 2. This template introduces a Word tier, which itself contains tiers Text, POS, and Morpheme. In turn, the Morpheme tier contains sub-tiers of its own. The XML data may contain multiple tiers of each of these tiers and sub-tiers within a single entry.

Figure 2: Template fragment for FLEXTEXT

Figure 3: Formatted Koda example

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

wordy/word/word = Word ( #
item[@type="txt"] = Text
item[@type="pos"] = POS
morphemes/morph = Morpheme (  @lang = MLang #
text() = Morph
item[@type="cf"] = Citation
item[@type="msa"] = MSA
item[@type="gls"] = Gloss
Multi-tier search in Kratylos allows for regular-expression queries that can distinguish between morphemes with or without a representation in a particular writing system. A multi-tier query is composed of nested units. A unit has the form <tierName content>.

In Figure 3, the regular expression searches for any word in the utterance (<Word>) that contains a non-whitespace character (\S) in both Koda and Bangla citation forms, followed by a suffixal morpheme (a dash followed by a non-whitespace character: -\S) without a Bangla citation form (a dash followed by a whitespace: -\s, because FLEX generates morpheme break symbols in all the writing systems for a given tier even when there is no morpheme to display). Furthermore, the query specifies the category (MSA) of the first morpheme as v for verb (surrounded by word breaks, \b) and the category of the second morpheme as v:Any, indicating that it attaches to any verbal stem.

The query in Figure 5 results in over 300 hits in the Koda corpus, but none when the query further specifies that the first morpheme must be a verb in the MSA tier. This result reveals an interesting generalization in the data: code mixing is asymmetric. Bangla verb roots happily host Koda morphology, but Koda verb roots never host Bangla verbal morphology. A speaker’s intuition about which language an utterance is in largely correlates with the source language of the functional morphology. Thus, utterances like (3) and (4), despite having only a relatively small number of Koda morphemes, are still deemed to “be in Koda” rather than in Bangla.

The same asymmetry can be seen in the noun phrase, where Koda determiners can take Bangla noun-phrase complements, but Bangla determiners do not appear to take Koda complements. For instance, we find examples in the corpus such as the following, where the Bangla demonstrative fei ‘that’ introduces a Bangla noun phrase boj in (5) and dajitto in (6).

This absence of native Koda noun phrases following fei can be verified with the query in Figure 6, which matches instances of fei followed by a word that contains no Bangla citation form.
Figure 6: Query for a Bangla demonstrative with Koda complement

The only result is shown in (7), where kɔɖa is part of the larger phrase kɔɖa bʰaʃa headed by Bangla bʰaʃa ‘language’.

(7)  ar alɛ kɔɖa bʰaʃa alɛ jagar
daru-kɛ-n-a=lɛ
 can-PERF.TR-INTR.B-IND=1PL.EX.SBJ
‘And we can speak that Koda language of ours.’
(Koda Language 3, Sri Alamgir Shordar, 3)

Although only presented as a brief demonstration, the pattern uncovered militates against theories that allow free switching between languages that have parallel structures for a particular phrase (Poplack, 1980), as Koda and Bangla both show [Dem [NP]], as well as theories in which code switching is forbidden between functional heads (e.g. determiners) and lexically headed complements (e.g. NPs) (Belazi et al., 1994).

4. Conclusion

We have shown how Kratylos, a web-facing application, can exceed the query power provided by FLEX with regard to multilingual text. As a result, the linguist can conjecture and then verify generalizations with respect to code mixing and switching that would be otherwise hidden in the IGT data structures commonly employed by linguists. Furthermore, Kratylos allows such generalizations to be verified and explored by other users through the creation of an online public corpus [Kauffman and Finkel, 2018].

5. Acknowledgments

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

6. Bibliographical References


