LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 12 (2018), pp. 124–146
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24765
Revised Version Received: 2 Feb 2017

Kratylos: A tool for sharing interlinearized and lexical data in diverse formats

Daniel Kaufman
Queens College, CUNY & ELA

Raphael Finkel
University of Kentucky

In this paper we present Kratylos, at www.kratylos.org/, a web application that creates searchable multimedia corpora from data collections in diverse formats, including collections of interlinearized glossed text (IGT) and dictionaries. There exists a crucial lacuna in the electronic ecology that supports language documentation and linguistic research. Vast amounts of IGT are produced in stand-alone programs without an easy way to share them publicly as dynamic databases. Solving this problem will not only unlock an enormous amount of linguistic information that can be shared easily across the web, it will also improve accountability by allowing us to verify analyses across collections of primary data. We argue for a two-pronged approach to sharing language documentation, which involves a popular interface and a specialist interface. Finally, we briefly introduce the potential of regular expression queries for syntactic research.

1. Sharing IGT corpora and improving accountability[1] Good progress has been made over the last decade in creating browsable electronic dictionaries from common digital formats such as Toolbox and Fieldworks Language Explorer[2] (FLEx) (Black & Simons 2008). Among popular free software, we find Lexique Pro (SIL International (SIL International 2014), an early program for converting Toolbox data into web dictionaries, and the more recent Webonary (SIL International 2017), which takes a FLEx project as its input and is entirely web-based. In contrast to the progress being made for dictionaries, interlinear glossed text (henceforth IGT) has presented more serious challenges. IGT has played a central role in language documentation and description for over a century, yet there is no simple software for creating searchable,

[1]We thank audiences at the Tools and Methods Summit sponsored by the Arc Centre of Excellence for the Dynamics of Language (University of Melbourne) and ICLDC 5 (University of Hawai'i at Mānoa). Many thanks especially to an anonymous reviewer whose careful and insightful comments led to several improvements in the manuscript. The programming and field work reported here are supported by NSF DEL Grant #1500753 to Raphael Finkel and Daniel Kaufman. Jiho Noh serves as RA to Raphael Finkel at the University of Kentucky, and Husniya Khujamyorova and Daniel Barry serve as RAs on the fieldwork discussed below.

[2]https://software.sil.org/fieldworks/.

Licensed under Creative Commons
Attribution-NonCommercial 4.0 International

E-ISSN 1934-5275

online IGT corpora.[3] Online IGT corpora for endangered and under-resourced languages thus remain surprisingly few in number relative to their importance. Because corpus linguistics is now largely driven by the field of Natural Language Processing (NLP), it has favored representations designed for aiding machine translation, such as dependency trees and part-of-speech tagging. Language documentation and description, on the other hand, demand more detailed and typically language-specific annotations, which much work in NLP attempts to either derive automatically or circumvent entirely. For example, separate tiers for phonetic, phonological, orthographic, and lexical representations may all figure heavily in a language documentation project, but are rarely treated in tandem by NLP-oriented applications. The last ten years, however, have seen significant attention towards language documentation and field work on the part of computational linguists. In a discussion of how computational linguistics can support language documentation, Bender & Langendoen (2010) make several predictions that have now come to pass:

> "Questions that computational methods will soon be able to help answer include the following: Given a transcribed and translated narrative, which is not yet in IGT format, which words are likely to belong to the same lemma? Or given a collection of texts and a partial morphological analysis, which words are still unaccounted for?"
>
> (Bender & Langendoen 2010:3)

For many field linguists, these problems, to a large extent, have already been solved by the FLEx software, which does a commendable job of parsing and highlighting material that requires checking. What has been far slower is another goal identified by Bender & Langendoen (2010:20), namely, the "[…] need to establish a culture of expecting claims to be checked against web-available data." It appears that advances in technology have not yet made a strong impact on improving the reliability of linguistic data. Problems of reliability have been decried in many corners of linguistics; to take but one example, Abney (2011), a former generative syntactician, now computational linguist, remarks:

> "[…] neither the structuralists nor the generativists have understood the second half of the scientific method, namely, making and testing predictions. In generative linguistics, one does speak of the predictions of the

[3]Himmelmann (2006; 2012) draws a clear line between raw data, on one hand, and analyses at various levels of abstraction, on the other hand. In this view, the former is considered "documentation" while the latter is "description". While acknowledging the utility of this conceptual division, we do not make a strict separation between documentation and description, as, in practice, one linguist's documentation is another linguist's description. Note, for instance, that grants for "language documentation" such as the NSF's Documenting Endangered Languages program and the Hans Rausing Endangered Language Programme do not accept pure documentation without description as a product of their grants. Nonetheless, in Himmelmann's terms, Kratylos brings together part of the descriptive record, e.g. the various levels of analysis that comprise a FLEx text or ELAN annotation, together with the associated documentation, e.g. an audio or video recording. By "corpus", we simply refer to any collection of data on a single language, whether it is highly structured, as a FLEx project, relatively unstructured, as a plain time-aligned transcription, or some combination of the above.

> theory. But the predictions are almost never seriously tested, and untested predictions are nothing more than speculations."
>
> (Abney 2011:10–11)

At the same time, practitioners of language documentation, who are in most cases neither computational linguists nor "full-time" theoreticians, see their work as improving the accountability in linguistic analysis. Himmelmann (2006), for instance, states in the outset of his landmark paper on documentary linguistics:

> "Language documentations strengthen the empirical foundations of those branches of linguistics and related disciplines which heavily draw on data of little-known speech communities (e.g. linguistic typology, cognitive anthropology, etc.) in that they significantly improve accountability (verifiability) and economizing research resources."
>
> (Himmelmann 2006:1)

Powerful tools for morphological parsing as well as building and searching lexica and corpora are now available to field linguists, but they only bring us half-way to the above-stated goal of improved accountability. The second half involves being able to share data publicly so that anyone can verify a linguistic claim over the relevant corpus. Not only are tools lacking for sharing corpora generated in FLEx, data formats are rapidly changing (despite an apparent consensus since Bird & Simons 2003 that these formats be structured in XML). Long-term documentation projects thus accrue data in different electronic formats that cannot be viewed or searched as a unit. For instance, a language documentation project may have begun life in Shoebox or Toolbox, whereas later work on the same language may have been created natively in FLEx. Although FLEx has the ability to convert Toolbox files, a wealth of "legacy" documentation will in all likelihood never be migrated to newer programs. Furthermore, there are time-aligned transcriptions produced in ELAN (Wittenburg et al. 2006), Praat (Boersma & Weenink 2017) and other programs, which may either stand alone or not be fully integrated into an accompanying FLEx database. Finally, there are certain types of data, such as video, that cannot yet be integrated into FLEx. This summary only includes situations arising within a single project. There are also many cases of multiple projects on the same or closely related languages that a researcher would like to examine simultaneously.

Adding to the chaos, there is also the widely recognized problem of heterogeneous structures across IGT data itself (Bow et al. 2003). Schroeter & Thieberger (2006) conclude from their work in developing a standard IGT interchange format that:

> "[…] we must encourage standard data structures in whatever tool is used for transcription and annotation of linguistic data. While much of the data produced using current tools is reusable because it is in a non-proprietary format and has an explicit structure, it is not interoperable without considerable effort. The only method that seems to be capable

> of inducing linguists to use constrained data structures is the provision of templates as part of a clearly described workflow, resulting in data presentation that can be made available in a variety of views."
>
> (Schroeter & Thieberger 2006:119)

Easy-to-use templates may indeed the best hope for standardization across the field but here, too, the increasing adoption of FLEx has already gone a long way towards standardization (despite the current inability of FLEx to handle some common annotation types, e.g., phonetic transcription, exportable syntactic structure, or time-aligned annotations).

An effort to address current limitations in creating shareable IGT is especially timely given that computational linguists are now working to harvest legacy IGT data and transform it into XML, as in the ODIN project (Lewis 2006), and subsequent work (Palmer & Erk 2007; Xia et al. 2016; Gries & Berez 2017). The further transformation of published data into an XML format will not only allow us to check this data against naturalistic corpora, but also to uncover inconsistencies in grammaticality judgments in theoretically-oriented literature.

1.1 Existing tools for online IGT corpora TROVA[4] (Stehouwer & Auer 2011), a corpus search tool, and ANNEX[5] (Berck & Russel 2006), a web-based browsing version of ELAN, are two programs developed by the Language Archive at the Max Planck Institute for Psycholinguistics in Nijmegen. These tools do an admirable job in some respects but are problematic in others. Selection of a sub-corpus to search over, formulating the search, and viewing the results are handled by three separate applications. The results of a query are disassociated from much of their metadata, making it a challenge to derive the proper citation. Finally, it is impossible to extract a piece of interlinearized text for re-use in a paper or elsewhere because ANNEX, as an Adobe Flash program, does not allow text to be copied or extracted easily. The fact that Flash itself is being deprecated also puts the longevity of these programs in jeopardy.[6] A tremendous wealth of linguistic data, while technically accessible, cannot be easily utilized, as evidenced by the degree to which citations of traditionally published work far and away outnumber direct citations of similar data in archival deposits. Although it is possible to download time-aligned transcriptions, translations, and full IGT from archive deposits, it is only with considerable effort that an entire corpus of such data can be searched as a unit by outside researchers. As a result, this type of unpublished data tends to be cited only by the linguists involved in its collection and analysis.

Another important tool in this area is EOPAS (Schroeter & Thieberger 2006), an open-source program developed by a team led by Nick Thieberger and shown in Figure 1. EOPAS allows for online presentation and annotation with good browsing

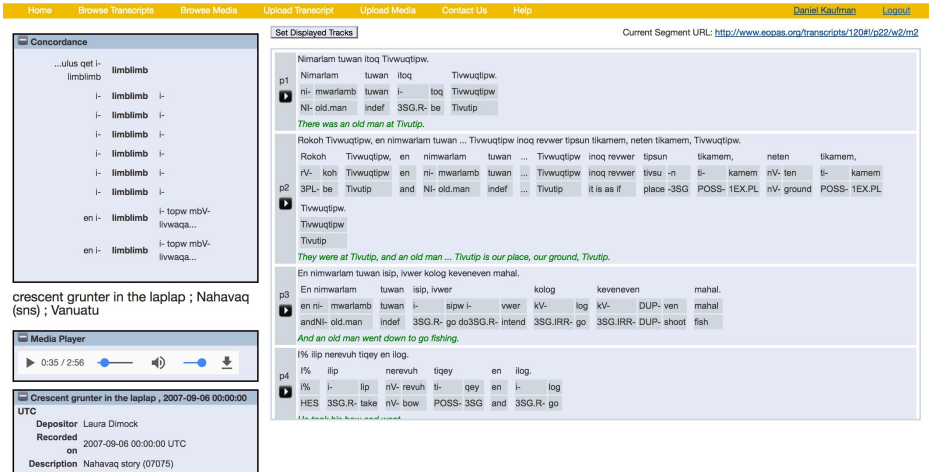
[4]https://tla.mpi.nl/tools/tla-tools/trova/.

[5]https://tla.mpi.nl/tools/tla-tools/annex/.

[6]A reviewer notes that the backend of TROVA is written in Java rather than Flash and thus may survive with a different frontend implementation if Flash becomes obsolete.

capabilities. It can currently take files produced in ELAN, Transcriber, and Toolbox as its input, which it translates into its own XML schema. It allows users to browse a text with associated media and offers excellent support for citation such that a user can cite data on the level of the phrase or even a particular morpheme. It also offers a concordance feature triggered by clicking on a particular annotation in the text, which is then queried across a particular collection. The concordance feature, however, does not yet allow for direct navigation between the various search results, and there is no possibility for the user to create a customized query. Although EOPAS is already an excellent tool for browsing and citing, the searching and concordance features are still at an early stage and do not yet provide the functions required for verifying analyses and corpus linguistics applications.

Figure 1. EOPAS (http://www.eopas.org/)

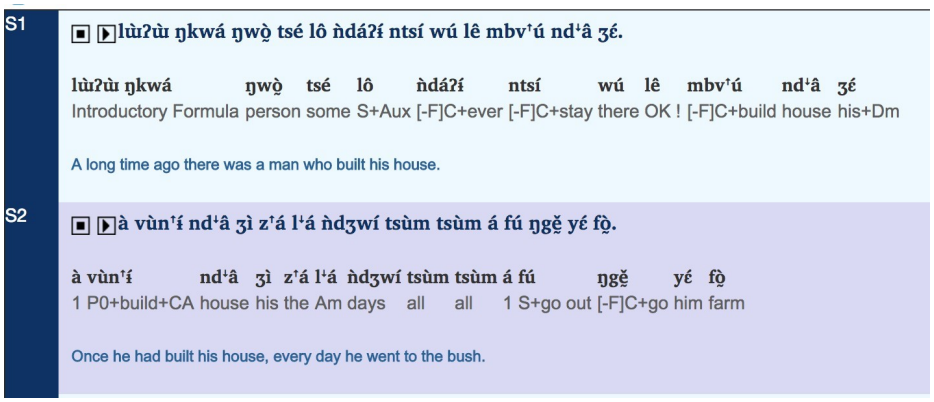


Finally, Pangloss (Michailovsky et al. 2014), developed by Lacito at CNRS, offers a somewhat similar set of functions to EOPAS. It converts IGT data from different formats into a custom XML format for display. Like EOPAS, it offers a good facility for browsing IGT data with media, a sample of which is shown in Figure 2. However, it offers no capabilities at all for searching or concordancing, rendering it of limited value for improving the accountability of linguistic analyses.

In sum, there remains a real gap between the stand-alone tools available to researchers building IGT corpora and a method for sharing the resulting databases in a fully dynamic form. Projects created in FLEx, the most widely used and sophisticated tool for building IGT corpora,[7] can be used to create print collections of IGT, but there are no means for sharing FLEx's search functions online. More generally, there are few browser-based tools that facilitate complex searches over any kind of IGT.

[7]The FLEx users' online forum has over 500 members and over 250 language projects are registered, although this probably represents only a fraction of the entire number.

Figure 2. Pangloss (http://lacito.vjf.cnrs.fr/pangloss)



2. Kratylos The tool introduced here, Kratylos,[8] attempts to bridge the gap between XML-formatted IGT and the technology available for displaying and searching such material on the web. Kratylos can be considered "middleware"; it takes as its input XML-formatted linguistic data from programs such as FLEx and ELAN, as well as popular legacy formats produced by FLEx's predecessors, Toolbox and Shoebox, and the TextGrid format employed by Praat, in addition to several lesser-used formats. Kratylos does not take an entire FLEx database as its input, but is meant instead to process exports from a project's lexicon and text corpus. As XML has emerged as the preferred format for interchange and long-term storage of linguistic data, and HTML is the best-established format for online presentation, the need for middleware tools such as Kratylos will only increase.

Kratylos does not alter the underlying database, and therefore does not need to communicate directly with any other software for updating the source of data. If the administrator of a project decides to make the project public, comments and corrections can be submitted by logged-in viewers with regard to particular records. Kratylos then presents this information to the project administrator, who can make alterations to the underlying database (in FLEx, Toolbox, etc.) if appropriate. This scheme is different from the model employed by Webonary (SIL International 2017), in which a public website is fully integrated with a piece of stand-alone software (FLEx). With Webonary, FLEx sends (lexicon) data directly to the website for electronic publication. An integrated web-based model of this sort for texts is not, as far as we are aware, being planned by the FLEx developers. Kratylos also does not attempt to replace archiving software, which must deal with a different set of difficult challenges that are beyond the scope of our work (e.g., enforcing adherence to standards, permanence, improved discoverability, versioning; see Holton et al. 2017 and Cassidy 2008 for discussion of the latter). There is, however, an unavoidable overlap in concerns when it comes to the problem of maintaining the stability of citations to documentation that is regularly revised. Here, we find a minor technical challenge and a more daunting issue of rights and permissions.

[8] http://www.kratylos.org/.

Storing entire data sets for all versions of a FLEx project with associated media requires a considerable investment in memory when we consider a repository with many projects. This can be overcome by incorporating a "reverse delta" repository, which economizes memory by only storing changes from one version to the next. The more serious problem for maintaining absolute stability of citations is that it would require researchers to relinquish some control over their material. Transfer of control is appropriate in the context of institutional archives, especially when those archives are tied to the sources funding the projects themselves (e.g ELAR, MPI Language Archive), but it is undesirable if not impossible for a piece of software like Kratylos to enforce terms of access. Citation stability, in other words, can only be guaranteed by people and institutions. Software can only help facilitate access.[9]

Optimally, researchers could simply point Kratylos to their archival deposits and let the program do the rest. Such an approach would avoid the need for researchers to again upload data and metadata that may already be archived elsewhere online. Versioning would thus be handled by the archive while research functions would be handled by Kratylos. We would be eager to explore such integration with an archive. It should be noted, however, that duplication of effort is still very much the norm in language documentation. Linguists who deposit their recordings in state-of-the-art archives in most cases must still upload them independently to media-sharing sites and social media in order to better reach the linguistic communities being documented.

2.1 Behind the scenes Any user who has created an account in Kratylos can upload data. The user must classify uploaded material by language (matches based on Glottolog are automatically suggested but not enforced) and can decide whether to make data private or public. The user is then asked to provide citation information and given the option to group files into collections (typically, an annotation file together with its corresponding media files). Users can furthermore add collaborators so that a private project can be viewed by multiple users without having to share it with the general public.

Kratylos scrutinizes the individual component files of an upload to determine their type and rejects any files that it cannot identify. Kratylos currently identifies FLEx texts, LIFT, TextGrid, EAF, Toolbox, Xigt, and CSV formats.

[9]In practice, the danger of losing a citation may be more remote than is at first apparent. In the normal course of developing a FLEx project, for instance, unanalyzed words in one version may be glossed in the next version. These analyses can be revised in further versions in any number of ways, including changing the glossing terminology and the identity of the morphemes themselves. It is considerably rarer, however, to make major modifications to the baseline or to erase a text entirely. If a text is erased, it is likely due to a permissions issue. If an analysis is revised, on the other hand, it is important to know that the linguist no longer subscribes to the analysis in the original citation, information which could be lost in a system that treated older versions of a project on par with current versions. The likelihood for a citation to become completely lost in an existing project due to revisions is somewhat unlikely. Even if the title of a text is changed, it should be possible to locate a record using a string from the baseline text or translation. Nonetheless, researchers should think twice before changing filenames or text titles in public projects and should not erase texts without considering possible consequences for citations. While this answer may not be very satisfying, it is in line with how other electronic resources are cited.

Kratylos subdivides data into records (typically lexemes, for dictionaries, sentences/utterances for FLEx texts, and annotation units for EAF and TextGrid files). Kratylos converts uploaded data, if necessary, into a new XML format. For example, the EAF format, although in XML, is not divided into records, so Kratylos reformats it into records that contain all the tiers (such as headword, part of speech, and gloss) and a reference to the media file.[10] Kratylos then applies a template to convert the XML into a Qddb representation (Herrin II & Finkel 1991; 1995). This representation stores all data in Unicode Normalization Form D (Canonical Decomposition). The template is format-specific and coordinates the following information:

1. The XML fields, described as XPath expressions,
2. The Qddb representation of those fields, which is hierarchical,
3. The formatting that the linear display should employ for those fields, which involves Cascading Style Sheets (CSS).

Although Kratylos could use Qddb itself to match queries, the databases are small enough that it can successfully apply a complete search. Some of the most complex parts of the software use the template to convert a matched entry into a displayable form.

The web server, Apache2, invokes Perl scripts using the Common Gateway Interface (CGI). The scripts use several standard modules: CGI (and submodules Carp, Session, and cookie), HTML::Template, Digest (submodules SHA and MD5), JSON, and Unicode::Normalize. The web pages that Kratylos presents to the user use the Bootstrap and JQuery libraries to format pages. The query results page also contains JavaScript code that converts entries on the fly. JavaScript handles vertical alignment across tiers, responsiveness to user interaction, and export into various formats. Finally, Kratylos maintains a MySQL database coordinating projects with their owners and other information.

Kratylos does not transcode various formats into a single semantic schema (cf. Bow et al. 2003) and, in this sense, is perhaps less ambitious than EOPAS (see §1.1 above). The advantage of this approach is that users do not have to make their data conform to any particular template. The disadvantage is that Kratylos will not know that a gloss tier from FLEx contains the same type of information as a gloss tier in ELAN or Praat, so these tiers cannot be targeted as a unitary class for a single query (Bow et al. 2003).[11]

[10]ELAN allows users to create various types of hierarchical relations between tiers and users often duplicate an entire tier structure for multiple speakers. This practice leads to greater indeterminacy when compared to FLEx data. As of now, Kratylos takes a simplistic approach based solely on time codes when creating records from ELAN or Praat files. Kratylos makes individual records from annotations that share a beginning and end time in ELAN and Praat. This algorithm handles basic cases quite well but will require further development to handle subordinate tiers that are not coterminous with their parent. This occurs, for instance, when users employ a point tier in Praat to annotate pitch accents, or another independent tier in ELAN to annotate gesture. Currently, individually annotated speakers do not present a problem for Kratylos, even when overlapping. Each annotation and its coterminous associates yields a separate record.

[11]There are several possible solutions that do not necessitate a full-fledged interchange format. Project administrators can be given the option to classify tiers according to standard levels of analysis as part of

Kratylos will be made open-source and accessible to the public through a GitHub repository at the end of the current grant period. Kratylos is built entirely from open-source software itself and transcodes proprietary media formats into the open-source codecs Ogg Vorbis (for audio) and Ogg Theora (for video).

2.2 The Kratylos display interface In the two screenshots shown in Figures 3 and 4, we can compare how one line, from a Wakhi text in Shaw (1876), appears in FLEx and in Kratylos. The various tiers of analysis are highlighted in FLEx in Figure 3, and the relevant tiers are shown as displayed by Kratylos in Figure 4. FLEx manages a complete database that contains links between IGT, a lexicon, a grammar module, and notes. Kratylos, on the other hand, simply takes XML exports produced by FLEx as its input and transforms them into a format that is quick and easy to search. The key tiers of analysis in FLEx are shown in Table 1.

Table 1. Tiers of analysis

Word	Orthographic representation
Morpheme	The allomorph employed in the current instance
Lexical entry	The basic or underlying morpheme
Lexical grammatical information	Morphological category, position in morphological template
Word gloss	Gloss for the entire word, used primarily in cases where word meaning is non-compositional
Word category	Lexical category of entire word

Figure 3. FLEx screenshot

5.5	Word	jaw	ðaj	wəzdəj			,	təm		prit	nə	njəştəj		
	Morphemes	jaw	ðaj	wəz	-d	-əj		ta	-m	prit	nə=	njəş	-t	-əj
	Lex. Entries	ja$_2$	ðaj	wəz	-t$_2$	-i$_1$		ta	-m	prit	nə=	niwiz	-t$_2$	-i$_1$
	Lex. Gloss	DEF/3SG	man	come	PST	PST		LOC.UP	PROX	front	NEG	exit	PST	PST
	Lex. Gram. Info.	det	n	v	v:Past1	v:(Pst2)		Loc/Dir	det:(Proximity)	n	<Not Sure>	v	v:Past1	v:(Pst2)
	Word Gloss	***		***				***			***	***		
	Word Cat.	pro	n	***				Loc/Dir		n	***	v		
Free	Her husband came, she went not forth into his presence.													

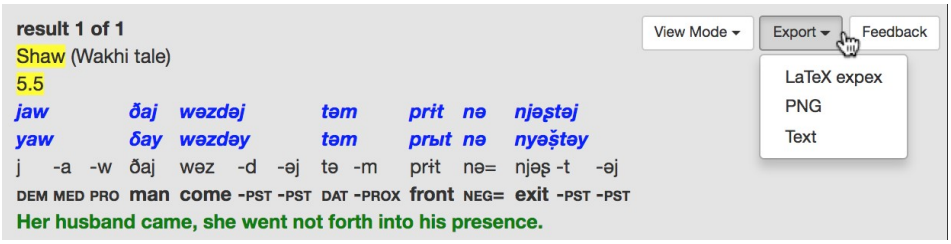
Several features of Kratylos are visible in Figure 4. First, Kratylos has added an additional orthography tier. The Wakhi project, described in more detail below, began by using IPA as the primary writing system but it was deemed essential that the corpus be displayed in a writing system with more currency among Wakhi speakers (although there is no widely accepted standard orthography). The second line in blue is an automatic transformation from the IPA writing system to a popular orthogra-

the intake process, as a kind of post-hoc template. Alternatively, we can allow for queries to target multiple fields across data formats using whatever names these fields were given in the original file. Currently, Kratylos aggregates all the fields in the data selected by the user from the projects table and lists these fields in the drop-down "filter" menu. The user then has the option of querying over a particular field/tier or over all fields (the "no filter" option).

phy. Kratylos thus allows for automatic transformations from one orthography into another for search and display, without altering the original database.[12]

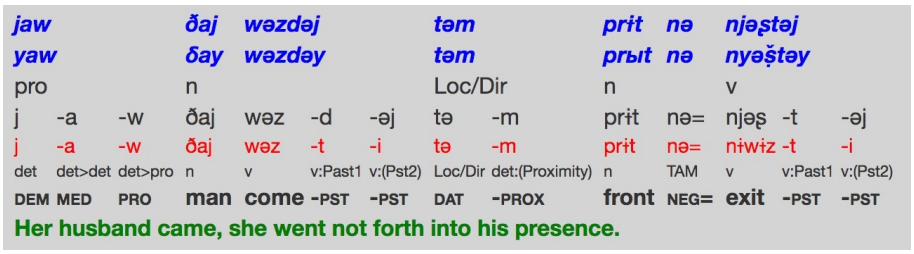
Second, Figure 4 shows export buttons. We have made it a priority to provide an easy way to extract and cite data from any corpus, a function which is not well developed in FLEx itself.[13] Kratylos presents several options for export: a PNG image of the record as seen in the browser, plain text (words and glosses unaligned), and a LaTeX format designed for the ExPex package.[14]

Figure 4. Kratylos screenshot – default view



The latter export can then be typeset by LaTeX with a standard result as in (1). The user may also interactively hide tiers in the browser. Tiers hidden by the user in the browser are not exported, so the amount of information in the source IGT can be easily customized for presentation. Figure 4 displays the default view of the FLEx tiers: word, lexical entries, lexical gloss, and free translation. With all tiers displayed, the same example appears as in Figure 5.

Figure 5. Kratylos screenshot – all tiers



(1) *j-a-w ðaj wəz-d-əj tə-m prɨt nə=njəʂ-t-əj*
DEM-MED-PRO man come-PST-PST DAT-PROX front NEG=exit-PST-PST
'Her husband came, she went not forth into his presence.'
(Shaw 1876, Wakhi tale 5.5)

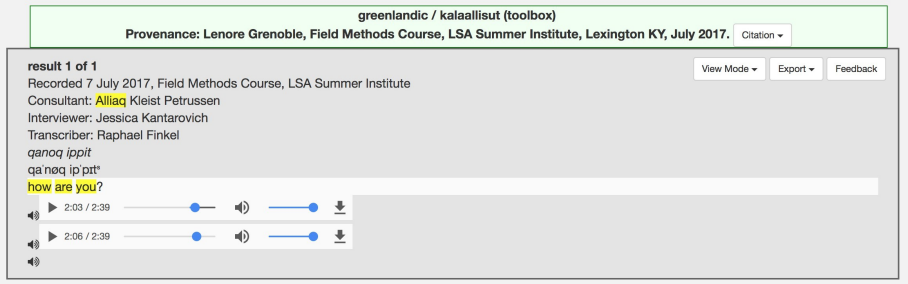
[12]This facility addresses some of the concerns voiced by McEnery & Ostler (2000:412), among others, regarding writing systems. Such automatic transformations only work if there is a consistent and predictable relation between orthographies.

[13]Copying and pasting an example consisting of IGT is, in fact, impossible in FLEx without first exporting the entire text into a format that can be read by a word processor.

[14]https://www.ctan.org/tex-archive/macros/latex/contrib/expex.

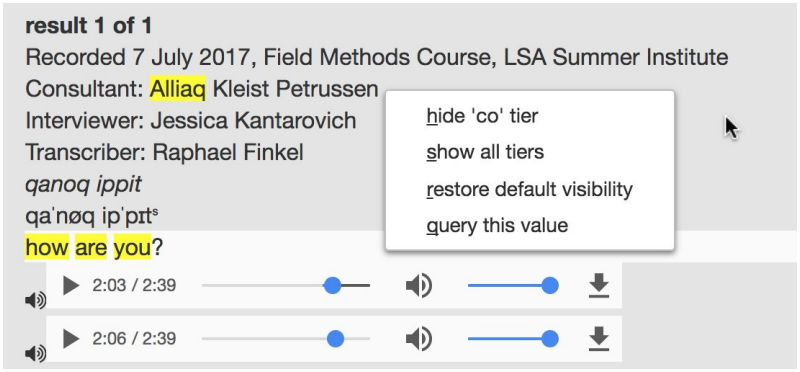
Kratylos also accepts Toolbox data, as shown in Figure 6. In this example we also see how audio is presented. The locations relevant to the record (indicated in the original toolbox file with the code \tc) can be played by clicking on the speaker icon. The audio control then appears so the user can listen to any selection within the same recording. The portion relating to the record can be replayed by clicking on the icon. As with all other types of data, the record can be exported in various formats.

Figure 6. Displaying Toolbox data



As shown in Figure 7, clicking on an element in the text opens up a new menu, which can customize display by hiding or displaying particular tiers. This menu also facilitates a search on the selected element. By clicking on the highlighted tier and selecting 'query this value', the user can browse through all the records that match the selected content. Similarly, in one click, it is possible to search over all instances of a word, translation, consultant, interviewer, or any other information included in a given tier.

Figure 7. The tier/element menu

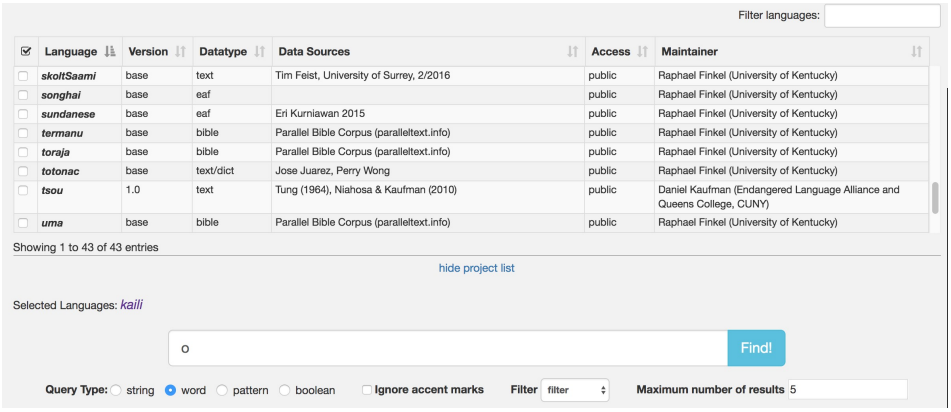


Kratylos allows search over multiple projects, which are selected from the table shown in Figure 8. This table consists of all public projects together with private projects that the user has access to. The table displays various types of information about the projects and can be filtered by any field.[15]

[15]Searching over multiple projects typically entails searching over multiple sets of glossing terminologies and even alternative analyses of the same language. There is no way for Kratylos to address the roots

Within the search options, the user has a choice to search for a partial string, a full word, a (Perl) regular expression, or a Boolean search. Any search can target a particular tier. The filter menu is populated by all the fields present in the collections selected for search.

Figure 8. Searching in Kratylos



2.3 Popular vs. research interfaces: the Wakhi project We demonstrate the further use of Kratylos with a project on Wakhi based at the Endangered Language Alliance[16] and also supported by NSF DEL Grant #1500753. Wakhi is an Iranic language of the Wakhan corridor of Afghanistan and adjacent areas in Tajikistan, Pakistan, and China, spoken by roughly 50,000 people. The bulk of the Wakhi IGT data comes directly from FLEx. Currently, it consists of a substantial set of texts collected by Pakhalina (1975) and Grünberg & Steblin-Kamensky (1988) in addition to two semesters of field work in the context of a field methods course taught by the first author, a smaller number of published texts by other linguists, and a number of texts collected by a Wakhi-speaking research assistant in Tajikistan.

(2) *kuj j-a-w dif-t j-a-w mərz jo ʧax*
who DEM-MED-PRO know-3SG.NPST DEM-MED-PRO hungry or thirsty
'Who knows whether he is thirsty or hungry.'
(ELA ruboi - satkək bə jıwət buj, 7)

of this ontological problem but regular expressions can aid in getting around it, provided that the user is already aware of the terminological issues when searching. For example, Philippine languages are analyzed alternately as ergative or accusative, among other options (Kaufman 2017). A regular expression query can easily generalize over both analyses using a term such as (ABS)|(NOM), which matches either ABS (for absolutive) or NOM (for nominative). Alternative abbreviations such as PERF and PRF for the perfective aspect or PAST and PST for the past tense can be captured by regular expressions such as PE?RF and PA?ST where ? indicates an optional character preceding.

[16] http://www.elalliance.org/.

Figure 9. Wakhi ruboi: the popular interface (YouTube)

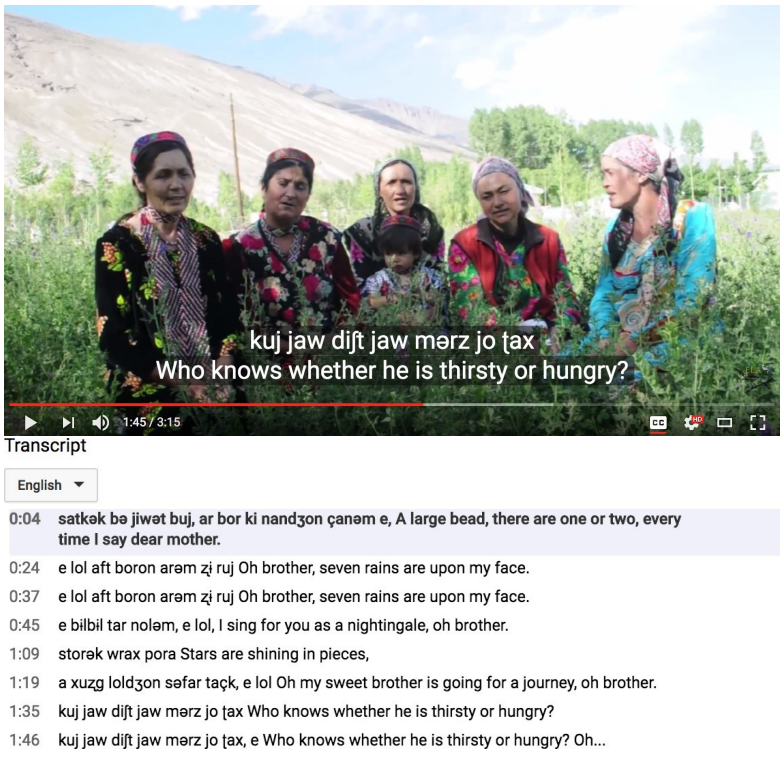
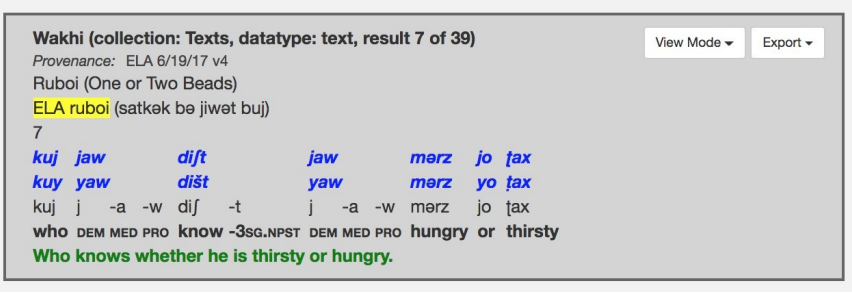


Figure 10. Wakhi ruboi: the research interface (Kratylos)



Kratylos is currently best suited for linguists and academic research, but does not yet facilitate wider dissemination to a language community. In the Wakhi project, our fieldwork recordings are displayed as searchable IGT in Kratylos as well as independently on YouTube and Facebook for the community at large. As shown in Figure 9, the YouTube videos, in this case, a traditional Wakhi *ruboi* song, can be navigated by a rolling time-aligned annotation that contains both the transcription and translation.[17] On YouTube, these transcripts can be copied and the videos can be shared

[17]These subtitles can be produced natively in YouTube, which has good facilities for transcription and export, or in ELAN, which can export annotations in the SubRip (SRT) subtitle format. Annotations in the SubRip format can then be uploaded to YouTube, Facebook, and other video-sharing services.

and commented upon. Facebook, on the other hand, does not allow viewing or navigating through an entire transcript, but in the context of the Wakhi project, it has facilitated far wider viewership and community engagement (across Tajikistan and Pakistan) than would have been possible with YouTube alone. These platforms have thus proved indispensable for sharing our annotated media with communities that have some level of Internet connectivity. In Figure 10, we see how Kratylos displays a record of IGT corresponding to the same material in Figure 9. Finally, in (2), we see how this information can be exported for use in a publication. This approach retreats from the challenge of creating a one-size-fits-all solution that caters to a popular audience and linguists at the same time. In our experience, the language community prefers tools that they are already familiar with and use on a regular basis, such as YouTube and social media. What is important to this set of users is the ability to view, share, and comment upon an entire video uninterrupted. What is of most importance to linguists is citability and the ability to carry out complex searches. It strikes us as mistaken to attempt all of these things in a single tool. However, it should be possible to implement a stronger link between Kratylos and popular interfaces for any piece of media, parallel to the above-mentioned goal of linking Kratylos to archival deposits for stability and versioning.

2.4 Kratylos as a corpus exploration tool There are few examples of how a descriptive linguist can make use of a corpus to yield generalizations in an understudied language.[18] It seems, in fact, that few field linguists employ corpora beyond their basic functions. This may be due to an unfamiliarity with corpus linguistics or a seeming disconnect between the topics commonly investigated in corpus linguistics proper (e.g. phraseology, polysemy, frequencies of words and collocations) and fieldwork-based descriptive linguistics (e.g. grammatical relations, subordination, pragmatic conditions on word order variation, anaphora).[19] Quantificational corpus linguistics is often presented as the true empirical approach to understanding language data, but given a large enough collection of text, the corpus can also play a vital role in hypothesis formation and the search for qualitative evidence without regard to frequency. We thus focus here on the corpus as a guide rather than a measurement tool.

One obstacle to better accountability, which Kratylos attempts to overcome, is a division between the type of corpus created by field linguists and that used by corpus linguistics for syntactic research. The former type, described above, consists of IGT with associated media and notes. The latter type typically consists of treebank formats in which the relations between words and phrases are coded explicitly. For

[18]See Cox (2011) for a discussion of other fundamental distinctions between the fields of corpus linguistics and language description and how these can be bridged. Mosel (2014) discusses the importance of corpus tools to discover hidden generalizations in field data, exemplifying with her own documentary and descriptive work on Teop. Rice & Thunder (2017) discusses corpus-building as a community-led project.

[19]Fillmore's (1992) caricature of an encounter between a corpus linguist and a theoretical ("armchair") linguist remains apt 25 years later:

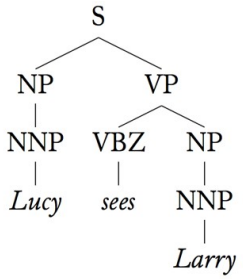
> "[…] the corpus linguist says to the armchair linguist, 'Why should I think that what you tell me is true?', and the armchair linguist says to the corpus linguist, 'Why should I think that what you tell me is interesting?'" (Fillmore 1992:35)

instance, the simple sentence "Lucy sees Larry" would be represented in IGT as in (3) and in Penn treebank notation as in Figure 11. The IGT format minimally presents the text in a phonetic or phonemic representation, followed by a morphologically segmented representation, followed by a free translation. The Penn treebank notation (Taylor et al. 2003) contains a hierarchical structure where the entire string is identified as a sentence (S), "Lucy" and "Larry" are identified as proper nouns (NNP), and "sees Larry" is identified as a verb phrase (VP). The traditional syntactic tree on the right is simply a notational variant of the bracketed structure on the left (omitting a node for the final punctuation).

(3) *'lusi 'si-z 'læɹi*
Lucy see-3SG.PRES Larry
'Lucy sees Larry.'

Figure 11. Wakhi ruboi: the research interface (Kratylos)

```
(S (NP (NNP Lucy))
   (VP (VBZ sees)
       (NP (NNP Larry)))
   ( .))
```



How can well-annotated IGT stand in for these types of representations, where constituency can be referenced directly by a query? If our goal is hypothesis testing, then landmarks such as case marking and position within the clause can serve as proxies, albeit imperfect ones, for syntactic structure. Case-rich languages like Finnish or Latin wear their argument structure on their sleeves (despite various phenomena that obscure these relations). On the other hand, grammatical relations in languages with little case marking but relatively fixed word order, like English, can be partially derived from linear order in relation to predicates and clause boundaries.[20] The field linguist using a corpus to test a syntactic hypothesis must often mimic the hearer in finding ways to attribute hierarchical structure to a "flat" string of words.

A key function of Kratylos is its ability to perform complex regular expression searches over large amounts of structured data. The benefits of this feature in the realm of morphology and phonology have been discussed by Mosel (2012), who examines reduplication patterns. We explore below some applications of regular expressions queries for syntactic description.

[20]Certain dialects of Wakhi present an interesting challenge here, which we explore in separate work, in that they display relatively free word order and a peculiarly uninformative case marking system.

2.5 Regular expressions Regular expressions consist of a vocabulary for pattern matching across a string of characters.[21] Some of the more important expressions for the purposes of querying IGT are shown in Table 2.

Table 2. Regular expressions

. any character	\b word break	\s whitespace	\S non-whitespace
* zero or more times	+ one or more times	? one or zero times	{x} *x* times
{x,y} at least *x* and at most *y* times	^ beginning of line	$ end of line	
[vcd] *v*, *c* or *d*	[^vcd] not *v*, *c* or *d*	dog|cat dog or cat	
(?!abc) not followed by *abc*	(?<!abc) not preceded by *abc*		

As seen above, some expressions generalize over characters. For instance, the symbol . generalizes over all characters and \S generalizes over all characters that are not whitespace. The second row in the table shows quantifiers, which follow their argument. The simple expression .* matches zero or more instances of any character. This pattern lets us search for multiple elements in a string.

For example, the following regular-expression query searches for records containing the gloss NOM followed by OBL, regardless of any intervening material.[22]

```
NOM.*OBL
```

Regular expressions can also anchor a query to the edge of a line (which typically corresponds to an utterance in our own FLEx projects). The symbol ^ represents the beginning of a line, and $ represents the end of a line. Thus, an expression like the following will match the suffix -OBL in utterance-final position.

```
OBL$
```

Table 2 also shows the use of square brackets and the vertical bar, which handle alternations of characters and strings, respectively. The following expression is identical to the previous one except that it matches words with either -OBL, -DAT or -ABL

[21]The term "regular expression" originated in the "regular languages" of mathematics, but this is now understood to be a misnomer, as regular expressions of the modern variety exceed the power of regular languages by allowing for context-sensitive matching.

[22]Currently, all searches in Kratylos are bound by records. A record is defined as a line in a FLEx text and a top-level annotation in other text formats (e.g. TextGrid, EAF). A reviewer brings up the important point that records in FLEx, which are generally syntactic units such as the sentence, will be quite different from those coming from time-aligned annotations, which are often syntactically incomplete utterances. Searching within the sentence domain thus requires searching across records in time-aligned formats. We plan to address this limitation by allowing an option for querying across a set number of adjacent records. In the absence of marked syntactic boundaries (e.g. punctuation), this facility would yield false positives when searching for intrasentential patterns, but this problem seems unavoidable. On a positive note, such a facility will also allow searching for local cross-sentential patterns in FLEx records.

at the end of a line.[23]

```
-(OBL|DAT|ABL)$
```

Kratylos places individual morphemes and their glosses into separate columns. If we want to target an utterance-final word with a particular morpheme, say -PST (past tense), we can use the following expression, which allows for any number of suffixes (demarcated by -) or clitics (demarcated by =) to intervene between -PST and the end of the line. The expression (-|=)\S+ represents a string beginning with - or = followed by one or more non-whitespace characters.[24] With the following * quantifier, the suffixes and clitics matched by this expression can occur zero or more times before matching the end of the line.

```
-PST( (-|=)\S+)*$
```

Finally, the expressions on the last row of Table 2 extend querying power considerably by being able to filter out material that precedes or follows a search target. These "look-around" expressions allow us to narrow down our results by making queries context-sensitive. The following query, for example, matches a record with -OBL preceded by -DAT but not followed by -DAT. The latter filter is expressed by the sub-expression (?!.*-DAT), which disallows a following string -DAT, regardless of any intervening material (indicated by .*).

```
-DAT.*-OBL(?!.*-DAT)
```

As mentioned earlier, regular expressions afford us some power to probe syntactic structure in a syntactically unparsed corpus. If we were to examine word order variation in Wakhi, we could begin by searching for all examples in the corpus that deviate from the unmarked SOV order, as in (4), where the object, *ʐi nani* 'my mother', precedes the subject, *tu* 2SG.NOM.

[23]Note also that the parentheses groups expressions together. Strings which are grouped in this way are treated as variables which can be referred back to by the expressions \1, \2, etc. for the first "captured" group, the second "captured" group, etc. See Mosel (2012) for how this feature can be used to explore reduplication patterns.

[24]It is typically necessary to use \S (the set of non-whitespace characters) to refer to a linguistic string instead of \w (word characters) because the latter set does not include special characters such as accents marks.

(4) Scrambling with the nominative-accusative pattern
kal-i *gowbon* *j-a* *ðaj-i* *wɨdɨr-d* *kɨ* [*zɨ*
bald-OBL shepherd DEM-MED man-OBL hold-3SG.NPST COMP 1SG.GEN
nan-i] [*tu*] *ʃi-t=ət*
mother-OBL 2SG.NOM kill-PST=2SG
'Then the bald shepherd grabs that man, (screams): "You killed my mother!"'
(PKH tsɨbɨr vrɨt, 2.10)

In a derivational theory, the syntactic structure of the relevant clause would be represented roughly as in the tree in Figure 12.

Here, the object phrase has moved from the verb phrase to a topic position in the left periphery of the clause. Given a treebank with such fine-grained structures, it would be a simple matter to search for a dislocated noun phrase (or DP) dominated by the Topic Phrase (TopP). In the syntactically unparsed Wakhi IGT corpus, we can approximate this type of query by looking for an instance of oblique case (OBL) preceding nominative case (NOM) preceding a verb. By employing a negative lookahead expression, we can eliminate many potential false positives where the two case-marked words do not belong to the subject and object phrase of a single predicate. We ensure that other verbs do not intervene between the OBL and NOM and that another OBL marker does not intervene between the NOM and the verb. That is, we can search for the string in (5), which excludes the starred elements. (Further exclusions could also be made to ensure that the initial oblique marked phrase is not embedded in an adpositional phrase or adjunct.)[25]

(5) OBL (*V) NOM (*OBL) V

The results, in this case, include simple elicited examples, such as (6), as well as far more complex examples from naturalistic speech, as in (7), where a long utterance contains an OBL argument, *zik-i* 'tongue', fronted across a nominative argument *wuz* 1SG.NOM.

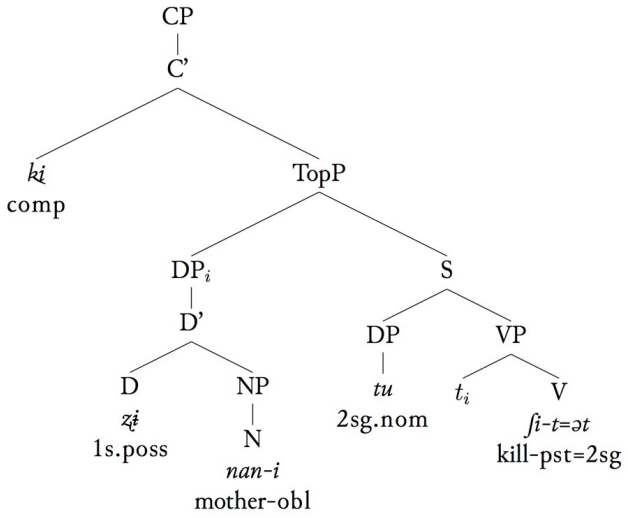
(6) *to=ʂ* *wuz* *win-əm*
2SG.OBL=PROG 1SG.NOM see-1SG.NPST
'I see you.' (Columbia fieldmethods, Fall 2011, 401)

[25]The actual regular expression used in this query, which searches for OBL not directly followed by -DAT or -ABL but followed by NOM without any intervening verbal elements or clausal boundaries (i.e. PST, COMP, PRF, NEG) is far too complex for users to create from scratch:

(OBL) ?(?! ?(-DAT|-ABL)) (?!((?!NOM).)*(PST|COMP|PRF|NEG=?))((\S+ ?)*(\S*NOM)).

Kratylos will allow users to filter out interveners between two search targets without having to work directly with regular expressions. The user will input the search terms into text fields which will then be plugged into a general regular expression template for certain types of complex searches.

Figure 12



(7) *agar tsum dɨnjo peʃravi tsar-t agar tsum taraqi*
if how.much world progress do-3SG.NPST if how.much develop
tsar-t zik-i wuz xojiʃ tsar-əm kɨ çat=bə
do-3SG.NPST tongue-OBL 1SG.NOM desire do-1SG.NPST COM REFL=also
toza niga dɨr-əm çɨ=zman-v-ər=bə
clean keep LIGHTV-1SG.NPST REFL.POSS=child-OBL.PL-DAT=also
zabon-i waxigi jəçk tsar-əm
language-EZF Wakhi accustomed do-1SG.NPST
'No matter how much the world is developing, I want to keep the language pure and pass it on to my children – I will teach them the Wakhi language.'

(ELA Davlatmo Bairambekova, 10.8)

An accurate quantitative picture of the OBL>NOM versus NOM>OBL orders would be very difficult to obtain from our dataset. However, regular expression queries allow us to find good examples of argument scrambling in a very large collection of text. These examples can then be examined more closely together with the original media and used as a basis for further elicitation. Although we must postpone for forthcoming work a more rigorous demonstration of how regular expression queries can resolve questions in Wakhi syntax, we hope the preceding examples give the reader an idea of the search power afforded by Kratylos and its potential.[26] For

[26]There are certain limitations on searching that will we will address in the near future. As of now, Kratylos allows for regular expression searches on any single tier. It also allows for Boolean searches (using any combination of the operators AND, OR and NOT) across all tiers. Ultimately, Kratylos will also allow for regular expression queries targeting multiple tiers specified by the user. This type of search would be necessary, for example, in querying structures with a specific morpheme or word immediately followed by a particular lexical category. As noted by a reviewer, ELAN currently allows for such searches over EAF files.

syntactically unparsed corpora, the user must devise strategies for probing syntactic structure. We have observed how case and verbal morphology can be used as syntactic landmarks in Wakhi but, clearly, each language's landmarks, whether they are morphological or positional, must be deduced individually.

3. Conclusion We have presented here Kratylos, a new online tool for creating linguistic databases from diverse data types. Kratylos fills a crucial lacuna in the current digital ecology of language documentation and description. There exist several tools for the creation of searchable online dictionaries but no such tools for sharing databases of IGT. We further demonstrated how Kratylos facilitates complex searches on IGT and suggested that such searches can even probe syntactic structure in a syntactically unparsed corpus. The general approach we suggest takes a two-pronged solution to sharing the products of language documentation. On one hand, there should be a public portal in a familiar digital venue where community members and others can share and comment. On the other hand, there should be a specialized tool for those who want to engage with the texts in a more technical manner. Even if our solution entails some duplication of effort (or even triplication for those who maintain a separate archival deposit and a popular media account), it circumvents the vexing problem of creating a single interface useful for linguistic research, popular purposes, and long-term preservation. Although many features are still in development, we encourage readers to experiment with Kratylos using their own data so that we can continue adapting it to different users' needs.

References

Abney, Steven. 2011. Data-intensive experimental linguistics. *Linguistic Issues in Language Technology* 6(2). 1–27.

Bender, Emily M. & D. Terence Langendoen. 2010. Computational linguistics in support of linguistic theory. *Linguistic Issues in Language Technology* 3(2). 1–31.

Berck, P. & A. Russel. 2006. ANNEX – A web-based framework for exploiting annotated media resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

Bird, Steven & Gary F. Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3). 557–582.

Black, Andrew H. & Gary F. Simons. 2008. The SIL FieldWorks Language Explorer approach to morphological parsing. In Gaylord, Nicholas, Stephen Hilderbrand, Heeyoung Lyu, Alexis Palmer & Elias Ponvert (eds.), *Texas linguistics society 10: Computational linguistics for less-studied languages*, 37–55. CSLI Publications.

Boersma, Paul & David Weenink. 2017. Praat: Doing phonetics by computer [Computer Program]. Version 6.0.35. http://www.fon.hum.uva.nl/praat/.

Bow, Cathy, Baden Hughes & Steven Bird. 2003. Towards a general model of interlinear text. In *Proceedings of the EMELD Language Digitization Project Conference 2003: Workshop on Digitizing and Annotating Texts and Field Recordings*. http://emeld.org/workshop/2003/bowbadenbird-paper.html.

Cassidy, Steve. 2008. A RESTful interface to annotations on the web. In *Proceedings of the 2nd Linguistic Annotation Workshop (The LAW II)*, 56–60.

Cox, Christopher. 2011. Corpus linguistics and language documentation: challenges for collaboration. In Newman, John, Harald Baayen & Sally Rice (eds.), *Corpus-based studies in language use, language learning, and language documentation*, 239–264. Amsterdam: Rodopi.

Fillmore, Charles J. 1992. Corpus linguistics or computer-aided armchair linguistics. In Svartvik, Jan (ed.), *Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991*, Vol. 82, 35–60. Berlin: Walter De Gruyter.

Gries, Stefan Th. & Andrea L. Berez, 2017. Linguistic annotation in/for corpus linguistics. In Ide, Nancy & James Pustejovsky (eds.), *Handbook of linguistic annotation*, 379–409. Dordrecht: Springer.

Grünberg, Aleksander Leonovich & Ivan M. Steblin-Kamensky. 1988. *La langue Wakhi [The Wakhi language]*. Vol. 1: Corpus de littérature orale. Paris: Maison des Sciences de l'Homme.

Herrin II, Eric H. & Raphael A. Finkel. 1995. Schema and tuple trees: An intuitive structure for representing relational data. *Computing Systems* 9(2). 93–118.

Herrin II, Eric H & Raphael A. Finkel. 1991. An ASCII database for fast queries of relatively stable data. *Computing Systems* 4(2). 127–155.

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

Himmelmann, Nikolaus P. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation and Conservation* 6. 187–207. http://hdl.handle.net/10125/4503.

Holton, Gary, Kavon Hooshiar & Nicholas Thieberger. 2017. Developing collection management tools to create more robust and reliable linguistic data. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 33–38. Stroudsburg, PA: The Association for Computational Linguistics.

Kaufman, Daniel. 2017. Lexical category and alignment in Austronesian. In Travis, Lisa, Jessica Coon & Diane Massam (eds.), *Oxford handbook of ergativity*, 589–628. Oxford: Oxford University Press.

Lewis, William D. 2006. ODIN: A model for adapting and enriching legacy infrastructure. In *Proceedings of the eHumanities Workshop*. (Held in cooperation with e-Science 2006: 2nd IEEE International Conference on e-Science and Grid Computing).

McEnery, Tony & Nicholas Ostler. 2000. A new agenda for corpus linguistics – Working with all of the world's languages. *Literary and Linguistic Computing* 15(4). 403–418.

Michailovsky, Boyd, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre Francois & Evangelia Adamou. 2014. Documenting and researching endangered languages: The Pangloss Collection. *Language Documentation and Conservation* 8. 119–135. http://hdl.handle.net/10125/4621.

Mosel, Ulrike. 2012. Advances in the accountability of grammatical analysis and description by using regular expressions. In Nordhoff, Sebastian (ed.), *Electronic grammaticography*, 235–250. Honolulu: University of Hawai'i Press. (*Language Documentation & Conservation* Special Publication No. 4, http://hdl.handle.net/10125/4537).

Mosel, Ulrike. 2014. Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language. In Nakayama, Toshihide & Keren Rice (eds.), *The art and practice of grammar writing*, 135–157. Honolulu: University of Hawai'i Press. (*Language Documentation & Conservation* Special Publication No. 8, http://hdl.handle.net/10125/4589).

Pakhalina, Tatyana N. 1975. *Vaxanskij jazyk* [Wakhi Language]. Moscow: Nauka.

Palmer, Alexis & Katrin Erk. 2007. IGT-XML: An XML format for interlinearized glossed texts. In *Proceedings of the Linguistic Annotation Workshop*, 176–183. Stroudsburg, PA: Association for Computational Linguistics.

Rice, Sally & Dorothy Thunder. 2017. Community-based corpus-building: Three case studies. Paper presented at the 4th International Conference on Language Documentation and Conservation (ICLDC4), Honolulu, February 26–March 1, 2015. http://hdl. handle.net/10125/42052.

Schroeter, Ronald & Nicholas Thieberger. 2006. EOPAS, the EthnoER online representation of interlinear text. In Barwick, Linda & Nicholas Thieberger (eds.), *Sustainable data from digital fieldwork: Proceedings of the conference held at the University of Sydney, 4-6 December 2006*, 99–124. Sydney: Sydney University Press. http://hdl.handle.net/2123/1297.

Shaw, R.B. 1876. On the Ghalchah languages (Wakhi and Sari-koli). *Journal of the Asiatic Society of Bengal* 45(1). 139–278.

SIL International. 2014. LexiquePro [Computer Software]. http://www.lexiquepro.com.

SIL International. 2017. Webonary [Computer Software]. http://www.webonary.org/.

Stehouwer, H. & E. Auer. 2011. Unlocking language archives using search. In Vertan, C. & M. Slavcheva, P. Osenova & S. Piperidis (eds.), *Proceedings of the workshop on language technologies for digital humanities and cultural heritage, Hissar, Bulgaria, 16 September 2011*, 19–26. Shoumen, Bulgaria: Incoma Ltd.

Taylor, Ann, Mitchell Marcus & Beatrice Santorini. 2003. The Penn Treebank: An overview. In Abeillé, Anne (ed.), *Treebanks: Building and using parsed corpora*, 5–22. Dordrecht, Netherlands: Springer. https://doi.org/10.1007/978-94-010-0201-1_1.

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

Xia, Fei, William D. Lewis, Michael Wayne Goodman, Glenn Slayden, Ryan Georgi, Joshua Crowgey & Emily M. Bender. 2016. Enriching a massively multilingual database of interlinear glossed text. *Language Resources and Evaluation* 50(2). 321–349.

Daniel Kaufman
daniel.kaufman@qc.cuny.edu
orcid.org/0000-0003-0971-8409

Raphael Finkel
raphael@cs.uky.edu