Bulgarian Dialectology as Living Tradition: A Labor of Love

Quinn Dombrowski

Division of Literatures, Cultures, and Languages & Stanford University Libraries, Stanford University, Palo Alto, CA, USA; ORCID: 0000-0001-5802-6623

Ronelle Alexander

Department of Slavic Languages and Literatures, UC Berkeley, Berkeley, CA, USA

Vladimir Zhobov

Department of Slavonic Philology, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria

Bulgarian Dialectology as Living Tradition: A Labor of Love

Bulgarian Dialectology as Living Tradition (BDLT) has been one of the longest-running Slavic digital humanities projects in the United States. Initially conceived in 2008 as a series of printed volumes, the digital project was built upon the foundation of a long-term international collaboration dating to the 1970's. As BDLT nears completion in 2019, this paper reflects on the trajectory of its development and its sustainability as an unfunded digital humanities project, and the ways it can serve as both a model and cautionary tale for others who seek to undertake similar work.

Keywords: digital humanities; digital preservation; content management systems; dialectology; Bulgarian language

1. Introduction

Bulgarian Dialectology as Living Tradition (BDLT) has been one of the longest-running Slavic digital humanities projects in the United States. Initially conceived in 2008 as a series of printed volumes, the digital project was built upon the foundation of a long-term international collaboration between Ronelle Alexander (UC Berkeley) and various Bulgarian scholars dating to the mid-1970's. A serendipitous conversation in 2011 between Alexander and Quinn Dombrowski, a digital humanist with a background in Slavic linguistics, and in library and information science, transformed the focus of the project from preparing Word documents for eventual publication, to preparing data for entry into a database. Soon after that, Vladimir Zhobov of Sofia University became the Bulgarian research director of the new digital project. In 2016, Zhobov and Alexander decided to open the previously password-protected website to the public, even as basic data entry was still in progress, and in 2019, the project officially launched after many years "in beta". While data entry is not yet finished for all aspects of the project, it is rapidly nearing completion.

At eight years old, BDLT hardly ranks among the most longstanding digital projects with a focus on Slavic materials (cf. various national corpora-building efforts; manuscript markup and display environments such as http://manuscripts.ru/; the database of Russian birchbark letters at http://gramoty.ru; and dialectological databases, such as those on Russian dialects found at http://www.parasol.corpus.org/Pushkino and http://www.rureg.hs-bochu.de, a site on Bulgarian diaspora dialects at http://www.corpusbdr.info, and a comprehensive site on Polish dialects at http://www.dialektologia.uw.edu.pl/index.php?11=start, etc.) Nonetheless, the institutional and financial circumstances for US-based Slavists undertaking digital projects are vastly different than for their colleagues situated in countries where such projects can be framed and funded as valuable efforts to bolster the national language in a digital environment where English predominates. As more US-based Slavists engage with digital tools and methodologies that are transformative in their application to Slavic studies, but may not be perceived as sufficiently "innovative" on the technical level to successfully compete for national-level grant funding (e.g. from the NEH, ACRL, etc.), BDLT can serve as a replicable model for project development that depends much more on time than money.

2. Overview

Bulgarian Dialectology as Living Tradition (BDLT, http://bulgariandialectology.org/) is a searchable database of oral speech representing the full range of Bulgarian dialects. It comprises 184 excerpts (henceforth called "texts"), drawn from a large corpus of material recorded in Bulgarian villages over the period 1986-2013. BDLT is the digital embodiment of a scholarly project with two goals: first, to make both the discipline and the material of Bulgarian dialectology available to a broader, international audience; and second, to bring the focus of dialectology back to the natural, spontaneous speech

which constitutes the basic data for dialectological research.

2.1 Background and source material

BDLT emerged out of a multi-year collaboration between the American Slavist and dialectologist Ronelle Alexander and several Bulgarian dialectologist colleagues. As early as 1975, Alexander began to discuss the desirability of joint fieldwork with Bulgarian colleagues Todor Bojadžiev and Maksim Mladenov; in 1986, two additional Bulgarian colleagues, Georgi Kolev and Vladimir Zhobov, joined this conversation. Such work was not possible during the socialist period, but as soon as the government changes occurred, various members of this group took short field trips: Alexander and Kolev recorded material in the Razlog region (one village) in 1990, and Alexander and Mladenov recorded material in the Ihtiman, Panagjurište, and Velingrad regions (five villages total) in 1992. These trips were followed by longer expeditions in 1993 and 1996, which visited many more locations and gathered the bulk of the material underlying BDLT. These ventures were supported by the International Research and Exchanges Board (IREX), and the field expeditions were directed jointly by Alexander, Kolev and Zhobov. When *Bulgarian Dialectology as Living Tradition* arose as a project and publication, the material from these field trips was augmented with similar work done by members of the research team, their colleagues, students and associates, in order to increase the geographic coverage and obtain a more representative set of transcripts from Bulgaria as a whole.

In order to place primary emphasis on natural spontaneous speech, audio clips from the actual field recordings have been made available along with each text, and both they and the transcriptions are presented in as "natural" a frame as possible. The audio files have undergone very little sound editing; only certain loud and distracting noises have been removed. In the transcription, every utterance has been included

(including those by bystanders when relevant to the conversation) as well as any non-linguistic sounds when there was even the slightest possibility that they may have influenced the flow of conversation, e.g. by distracting the speaker. In addition, overlapping speech by several informants has been transcribed. Such transcription of "natural speech", therefore, makes the material available for linguistic analysis at several different levels beyond the word itself (the focus of nearly all the maps in dialect atlases). Topics which are rarely, if ever, addressed in dialectological research, such as word order, functional sentence perspective, conversational analysis, narrative structures, and intonation, could now be studied on the basis of this material.

2.1 "Revitalizing Bulgarian Dialectology"

One of the conditions of the IREX grant supporting the 1996 field expedition was the publication of a volume summarizing results of the expedition. This volume, entitled *Revitalizing Bulgarian Dialectology*[1], was published in 2004 under the editorship of Alexander and Zhobov, in association with the University of California press, as an open-access PDF manuscript available through the California Digital Library's eScholarship platform. The goal of the expedition had been to "revitalize" Bulgarian dialectology both in Bulgaria and the West by means of putting Bulgarian and American students together in the field and creating situations where they could learn not only from their teachers but also from each other. The resulting volume included not only articles by the teachers (Alexander, Kolev and Zhobov) but also research papers by each participant, student and teacher alike, based on dialect material recorded during the expedition. The volume was published in California to underscore

1 Ronelle Alexander and Vladimir Zhobov, eds.. Revitalizing Bulgarian Dialectology. 2004. University of California Press. http://escholarship.org/uc/item/9hc6x8hp

the importance of making Bulgarian dialectal data available at the international level, and it was published electronically and open-access to maximize availability, especially in Eastern Europe.

2.2 BDLT as audio-based chrestomathy

Although *Revitalizing Bulgarian Dialectology* had made public some outcomes of the most recent expedition, the ultimate goal of the research team was to devise a way to make available the actual field material gathered on this and previous expeditions, and to do so in such a way as to make this material more accessible to outsiders. Realizing that it would not be possible to transcribe the entire amount of recorded material (over 200 hours), they decided to choose representative excerpts and create an audio-based chrestomathy; in order to make the chrestomathy more fully representative of the broad scope of variation throughout Bulgaria, they also decided to include material from previous trips undertaken by Zhobov and Kolev prior to their collaboration with Alexander. The plan was not only to transcribe each excerpt but also to provide it with interlinear glosses and an English translation; each excerpt would also be accompanied by a streaming audio file, a clip from the actual field recordings. The goal of the resulting publication was to make actual field data maximally available (including in audio form) at the international level. Furthermore, since the excerpts were chosen not only for linguistic value but also for content, the volume would give a representative picture of both linguistic variation and traditional cultural phenomena throughout Bulgaria.

3. BDLT as digital humanities project

At AATSEEL 2011, Alexander discussed the audio-based chrestomathy with Quinn Dombrowski. After receiving an MA in Slavic linguistics as well as an MLIS,

Dombrowski had found employment as IT staff in the Academic Technologies unit of the University of Chicago's central IT organization. Dombrowski had experience with developing digital humanities projects across a number of fields, and at the time was on the program staff of Project Bamboo, a Mellon-funded digital humanities cyberinfrastructure initiative. Having previously attempted an XML markup project[2] to capture dialectal variation in subsets of the data published in the *Bŭlgarski dialekten atlas*[3], Dombrowski had a personal interest in working with a different kind of Bulgarian dialectology material, and making it as accessible and reusable as possible.

Alexander shared early drafts of BDLT with Dombrowski in the form of Microsoft Word documents, where each line of the text was transcribed and translated into English, and each token was annotated with linguistic metadata. Dombrowski noted that the high degree of structure in these Word files was more reminiscent of a database than a traditional scholarly monograph. Moreover, the process required to generate the Word files involved significant duplication of work, as each token would need to be glossed and annotated anew every time it occurred. Not only was this inefficient, it also increased the risk of inconsistencies. Dombrowski felt that the rigidity of a print-oriented PDF end product would also limit its audience. The transcripts touch on a wide range of topics, from folklore and traditions, to agricultural practices, to personal stories

---

[2] Andrew Dombrowski and Quinn Dombrowski. "An XML-Based Approach to Dialectological Data: The Development of Syllabic Liquids in Bulgarian." Presented at the 17th Balkan and South Slavic Conference at The University of Ohio. 2010. http://quinndombrowski.com/blog/2010/04/13/bulgarian-dialect-atlas-at-the-17th-balkan-and-south-slavic-conference

[3] Stojko Stojkov et al., ed. 1964-1975. Insitut za bălgarski ezik. *Bălgarski dialekten atlas I-IV*. Sofia: Izdatelstvo na Bălgarska akademia na naukite, 1964-1981.

depicting daily life in rural Bulgaria. These narratives could be valuable in a wide variety of contexts, within and beyond the academy, but the formatting of the Word documents -- where the narrative was visually interrupted every few words by a block of linguistically-oriented data -- significantly impeded the narrative's readability and accessibility. This could be remedied by the production of another set of Word documents that presented the narratives as continuous text, but there, too, choices would have to be made about whether to include the original Bulgarian (and how: inline, in parallel columns, or separately), and whether to include a transliteration along with the Cyrillic (again, and how)?

Converting the project's structure to a database would eliminate these issues. Tokens could be entered, glossed, and annotated once, and these token entries would then be referenced in each text where they appear. Rather than committing to a single display format, database queries could enable any number of displays, in order to accommodate various audiences' needs and interests. A linguistics-oriented view could display all the tokens and their metadata (much like the original Word files); multiple narrative-oriented views could display the text without interruption, and in any combination of writing systems. A database would allow users to not only view the linguistic metadata on tokens but also to use it as a means of querying the transcripts: e.g. pulling up all lines that include a lexeme of interest, or all lines that include a particular verb form. A database would also facilitate augmenting the transcripts with additional metadata to support discoverability and analysis -- for instance, individual lines could be tagged with thematic content, and tokens could be grouped into phrases that show noteworthy linguistic features. In short, moving from a print-oriented workflow to a database would vastly increase the research potential of the corpus, in

addition to making the content more accessible to the broadest possible international audience.

Dombrowski offered to create a prototype of BDLT as an online database. She had previously built web-based digital humanities projects using the open-source content management system Drupal, and saw it as being well suited to this project as well. At the time, Drupal had a large, international developer community creating and maintaining modules (pieces of add-on functionality for the core Drupal platform) that could fulfill the project's technical requirements of storing and querying structured metadata, storing and presenting audio files, and importing and exporting text. This would allow Dombrowski to quickly develop a complex web application that would be highly customized to the specific data model of BDLT, without writing any code. (See "Drupal and Other Content Management Systems"[4] for further discussion of Drupal and other content management systems.) Drupal 7 was released shortly before Dombrowski began to develop the pilot version of BDLT, and the Drupal development philosophy supports API-breaking changes between major versions[5]. As a consequence, there is always a delay between the release of a new major version of Drupal core, and the point when it becomes usable for complex projects, as module developers need time to refactor their code if they intend to continue supporting their modules. For that reason, Dombrowski chose to build BDLT in Drupal 6 -- a decision that had long-term consequences, even as it was unavoidable at the time.

---

[4] Quinn Dombrowski. "Drupal and Other Content Management Systems" in *Doing Digital Humanities: Practice, Training and Research*, ed. C. Crompton, R.J. Lane, and R. Siemens. 2016. Routledege.

[5] Dries Buytaert. "Backwards Compatibility". May 17, 2006. https://dri.es/backward-compatibility

It took approximately 12 hours of work to develop the initial prototype of BDLT, which included an interface for entering and editing texts and annotating tokens, a text display equivalent to the original Word files, a map display of locations, as well as data structures for linking tokens to lexemes, annotating thematic content, and browsing all tokens, organized by lexeme. For the sake of expediency, Dombrowski manually entered the data for a few example texts, but anticipated that the existing Word files could be imported into the system without much difficulty. In February 2011, Dombrowski demoed the prototype for Alexander. After conferring with Zhobov, Alexander decided to move ahead with implementing BDLT as a database, with Dombrowski acting as the project's technical staff.

4. Technical implementation

The pilot version of the site that Dombrowski developed in 2011 remains largely unchanged to this day, though a few additional displays and features have accrued over the course of the site's development. "Digital Humanities Development without Developers: Bulgarian Dialectology as Living Tradition"[6] provides a detailed description of the technical underpinnings and data model for BDLT as of 2014, and is inclusive of all of the site's major features, with the exception of the more recent "phrases" content types and displays.

4.1 Structural overview

In brief, there are seven content types (Location, Contents, Text, Token, Line, Lexeme, Phrase) and five search functions (Wordform, Lexeme, Linguistic Trait, Thematic

---

[6] Ronelle Alexander and Quinn Dombrowski. "Digital Humanities Development without Developers: Bulgarian Dialectology as Living Tradition". 2014. Proceedings of DH-CASE II (DocEng workshop). doi: 10.1145/2657480.2657481.

Content, Phrase). The results from any search query can be exported as a CSV or Microsoft Word file, and the site provides a map display as one output from linguistic trait, wordform, or phrase search queries.

- *Locations*. Each village visited is located on a map on the home page and is represented by a page of its own, accessible either by a link from a list on the home page or a tab on the map. Each Location page gives basic metadata about the village (administrative region, dialect group, date visited), and provides a lengthy prose description of the relevant dialect subgroup. Salient traits of the group are illustrated by examples taken from the site itself. Links to the text(s) representing this village are also available on this page.
- *Contents*. This page displays basic information about each of the 184 excerpts, or texts, which the site contains: text name, dialect group, duration of audio file, number of lines of text, number of tokens of informant speech, and a brief synopsis of thematic content. Data entry status for content not yet completely entered is also noted. Data can be sorted on all columns except the audio length, and texts can be accessed from the text link on this page.
- *Texts*. Each text has its own page: it contains a sidebar with a small map locating the village, a photo image of the village, and metadata (date of recording, word count, physical context of recording, name(s) of investigator(s), and synopsis of thematic content). Each text is broken into lines for ease of data retrieval; each line is numbered, coded to identify the speaker, and provided with a timecode to facilitate location of the transcribed portion within the accompanying audio file. Each text is presented in three different views: Glossed view gives a translated text with interlinear tags, comprising grammatical and lexical tags placed underneath each token; Line view gives simply the transcribed text with English

translation; and Cyrillic line view simply gives the text in Cyrillic transcription (it is assumed that Bulgarian users need neither translation nor interlinear glosses). The audio link is available in all three views, and it follows the text as the user scrolls down the page.

- *Tokens*. Each token has its own page, which lists all the tags assigned to that token, and all the lines throughout the database where that token occurs, with each line identified by text name and line number.
- *Lines*. Each line has its own page, which lists all its tokens, any thematic content tags assigned to that line, and any identified phrases associated with that line.
- *Lexemes*. Each lexeme has its own page, with links to all the tokens associated with that lexeme. Note: a "lexeme" is the lemma in standard Bulgarian associated with the dialectal token; if no such lemma exists, then one is created and tagged as a "dialectal lexemes". Lexemes are also tagged for etymological and other information.
- *Phrases*. A unique feature of this site design is the ability to isolate grammatically significant groups of words, or phrases. Each phrase has its own page, listing all the tags assigned to it, as well as the line of its occurrence and any other lines in which it occurs.
- *Wordform Search*. This search page allows users to select any combination of grammatical/pragmatic tags and/or the English translation and/or the Bulgarian lexeme, and see all the lines on the site which display the tokens so identified; the geographical distribution of the selected tokens is displayed on a map. Each selected token is displayed within the line of its occurrence; users may then follow a link to the text with the token in question to see the larger context and hear the audio.

- *Lexeme Search*. This search page allows the user to see all the phonetic representations throughout the site of any one lexical item. Users can also isolate words with particular prefixes or suffixes (using “Begins with…”, “Ends with…” buttons). Users can also search for lexemes within categories of special interest (such as dialectal lexeme, loanword source), and for instances of lexical variation (the occurrence of more than one dialectal term for a particular item or action).
- *Linguistic Trait Search*. This page, by allowing the user to search for any one of a very large number of linguistically significant traits, enabled the linguistic tagging of tokens at a much more complex level than that marked by the interlinear tags which form the basis of the Wordform search. Here, the user makes hierarchically embedded choices to isolate the trait in question; this allows very complex searches at both the synchronic and diachronic level. Each selected token is listed in the context of its line, and the geographical distribution of selections is displayed on a map.
- *Thematic Content Search*. This search page allows users to find chunks of text (identified by text and line number) where the recorded conversation concerns a particular topic. The search page allows one to locate the desired topic either through a thematically ordered ethnographic list with many subdivisions in each category, or by an alphabetical list of every single tag regardless of its place in the hierarchical listing.
- *Phrase Search.* This search page allows the user to find instances of grammatically significant groups of words at a number of levels. This is particularly useful for scholars of Bulgarian and Balkan linguistics, since many of the traits characterizing the Balkan Sprachbund must be defined in phrasal

terms. Because there was no way to mark these traits at the token level, this additional content type was devised specifically for this site. As in other searches, results give the context of the full line, and display the geographical distribution on a map.

4.2 Hosting

Hosting is a perennial challenge for web-based digital humanities projects. Digital humanities thought leader Miriam Posner has characterized obtaining server space as "the most hilariously awful problem in doing DH at a university, and almost nobody has got this figured out. I know people who are secretly running servers under their desks, buying their own server space, or running projects off Google Drive."[7] Universities continue to struggle with questions of what campus organization, if any, should be responsible for providing web hosting for digital projects. Many central IT organizations, including those at UC Berkeley, follow a model of offering inflexible, standardized services in order to reduce support costs when those services are made available to the campus as a whole. As a result, they are a poor fit for digital scholarly projects, which are unlikely to resemble standard templates for departmental websites, faculty profiles, etc. At some institutions, the library has stepped in to fill the need for hosting for scholarly projects,[8] but when web hosting is seen as an indefinite commitment, the ongoing costs of server hardware and -- more significantly -- the staff

---

7 Miriam Posner. "Here and There: Creating DH Community". September 18, 2014. http://miriamposner.com/blog/here-and-there-creating-dh-community/

8 Jennifer Vinopal & Monica McCormick. "Supporting Digital Scholarship in Research Libraries: Scalability and Sustainability". Journal of Library Administration: Vol. 53, 2013, Issue 1. p. 27-42. https://doi.org/10.1080/01930826.2013.756689

time necessary for patching and updating software, can become a drain on library resources. As a result, some organizations that take supporting digital scholarship as their mandate have retreated to a position of offering advice on commercial hosting options, with the costs (financial and technical upkeep) to be borne by the scholar[9]. Over the course of its development, BDLT has navigated three of the most common hosting scenarios for digital projects.

Dombrowski built BDLT using a general-purpose shared web hosting account already purchased for use with multiple different projects. Within four years, the site needed to be migrated to a different environment after the hosting service threatened to shut it down due to an excess number of tables in the MySQL database. The large number of tables that Drupal generates as part of creating its content types (data structures) is a frequent criticism of the system[10], and it became a technical barrier for hosting the site using low-cost, general-purpose hosting. By 2015, when the site was threatened with eviction from its hosting environment, Dombrowski had moved to the Research IT organization at UC Berkeley (Alexander's institution), and was overseeing the hosting services offered by that campus's digital humanities program. Dombrowski initially arranged for BDLT to move to the Drupal-specific commercial hosting service that had partnered with UC Berkeley's IT organization to provide hosting for Drupal sites. This move was ultimately short-lived: recognizing that hosting would disappear when the digital humanities program's funding ran out, Dombrowski and Alexander

---

[9] Sarah Kalikman Lippincott. "Digital Scholarship at Harvard: Current Practices, Opportunities, and Ways Forward". June 27, 2017. https://projects.iq.harvard.edu/files/dsi/files/harvard_ds_final-report_20170627_v2.pdf

[10] "Drupal Schema – Why this methodology?" January 21, 2011. Drupal forums. https://www.drupal.org/forum/general/general-discussion/2011-01-21/drupal-schema-why-this-methodology

took advantage of the Berkeley Language Center's offer to move the site to their server, with system-level support from that unit's sysadmin. Under this model, hosting for the site would be guaranteed for at least as long as Alexander was an active or emerita faculty member at UC Berkeley.

4.3 Technical staffing

Technical development of digital humanities projects can quickly become costly, even when reusing existing code as part of a configurable open source content management system such as Drupal. Professional technical expertise commands a premium. For that reason, self-funded projects such as BDLT particularly benefit from having a core team of personally committed collaborators that includes at least one individual with the technical expertise to implement the project. While restricting the technical scope to what a core member of the team can personally accomplish may limit the project's scholarly ambitions, the alternative involves waiting for a significant influx of funding that may not be feasible, particularly if the project involves applying established methodologies in a new domain. Some projects attempt to overcome this hurdle by hiring professional technical staff to work on the project piecemeal as smaller amounts of funding (e.g. university-internal microgrants, etc.) become available, but this approach becomes more expensive overall as the project pays for the start-up costs of professional developers re-familiarizing themselves with the project at the beginning of each new phase of work. It also risks the project being left half-completed if the scholar is unable to secure further grants.

Dombrowski has served as the primary technical developer on this site from its inception to the present day. Like Alexander, Dombrowski has never been paid for work on the project, instead contributing out of personal interest and commitment. However, particularly because BDLT represents volunteer work, Dombrowski's availability to

direct time to the project has fluctuated, and changes in institution, job, job scope, and life circumstances (including the birth of three children over four years) have all had an impact. Alexander has used her research funds to pay graduate students with technical knowledge of Drupal (including some trained by Dombrowski) to implement site configuration changes during times when Dombrowski has been unavailable. However, those graduate students themselves have taken on this work as one among many conflicting priorities, including finishing their dissertations, leading to periods where they have fallen incommunicado for weeks or months at a time. In August 2017, Alexander took a weeklong workshop on Drupal offered by Dombrowski at the Digital Humanities at Berkeley Summer Institute, with the goal of developing sufficient technical proficiency to serve as her own technical backstop for the project, and reduce the turnaround time needed to make minor configuration changes on the website.

4.4 Migrations and code changes

One disadvantage of building a project using a content management system is that it closely ties the project's lifecycle to the support lifecycle for that version of the content management system. A major version upgrade is a non-trivial undertaking on any such platform, but Drupal's API-breaking design philosophy further exacerbates these challenges. Building BDLT in early 2011 necessitated the use of Drupal 6, but this choice guaranteed that the site would have to be migrated to a new version of Drupal within the medium term, when the Drupal open source project stopped providing security updates for that version.

In Alexander and Dombrowski 2014, the authors anticipated a migration directly from Drupal 6 to Drupal 8, with the expectation that version 8 -- which had not yet been given a release date -- would provide more robust technical underpinnings for the site in the long term. Instead, the Drupal project's decision to jettison much of Drupal's own

architecture and replace it with the enterprise PHP framework Symfony[11] had the effect of alienating many smaller-scale developers, including those who typically work on digital humanities projects. The resulting lag in module availability has been tremendous, and many modules with significant adoption in digital projects across a wide variety of disciplines (e.g. Biblio, which provides a data structure for importing, exporting, storing, and displaying bibliographic references) have still seen no significant movement towards a Drupal 8 port as of 2019[12].

By summer 2015 the release of Drupal 8 was imminent, and Drupal 6 would only be given a three-month grace period after its release before security updates were no longer provided. Discussions in the developer forums suggested that many general-purpose modules would not be available concurrently with Drupal 8's release, to say nothing of scholarly-oriented modules. In light of this, Dombrowski advised Cammeron Girvin, a graduate student working with Alexander, on a site upgrade to Drupal 7. While Girvin had served as a project manager for BDLT for some years, the upgrade was his first experience interfacing directly with the technical underpinnings of the site (i.e. the filesystem and MySQL database). The upgrade was difficult, requiring multiple attempts and a downtime spanning the entire summer before the site was again available online; furthermore, it took nearly an additional six months to resolve all the bugs related to the upgrade.

By 2015, Drupal 7 had seen significant uptake among digital humanists, and all of the widely used Drupal 6 modules were available for Drupal 7 at that point, or

---

[11] Dries Buytaert. "Why the big architectural changes in Drupal 8?" September 9, 2013. https://dri.es/why-the-big-architectural-changes-in-drupal-8

[12] Bibliography module – Issues – Drupal 8 port. March 20, 2015. Drupal module issue queue. https://www.drupal.org/project/biblio/issues/2456591

replaced with improved alternatives. Unfortunately, early in the development of BDLT, Dombrowski had selected a niche module called Editview as the primary interface for data entry. That module had been abandoned by its developer after Drupal 6, despite user requests for a Drupal 7 version starting in 2010[13]. Rather than completely reconceptualizing data entry for the site, Alexander contracted with Agile Humanities Agency (http://agilehumanities.ca/), a digital humanities-oriented development firm created by former English professor Dean Irvine, to write a Drupal 7 version of Editview. This piece of technical development work represented a significant financial investment for BDLT, but it also has served as a locus of broader impact for the project within the digital humanities community. The Drupal 7 Editview module has subsequently been adopted by other projects with similar tabular data entry needs, including the *George Washington Financial Papers Project* (http://financial.gwpapers.org/).

5. Data entry and labor

For BDLT and similar projects, the amount of time dedicated to developing the technical infrastructure is dwarfed by the enormity of the task of data entry. Dombrowski's expectation that data could be parsed from the 2011 Word files and imported into Drupal to seed the database was quickly shown to be overly optimistic. Despite consulting with developer colleagues at the University of Chicago who offered elaborate examples of using regular expression syntax to capture some of the words and linguistic annotation, it was ultimately too error-prone to use and all the lines, tokens, and annotations had to be manually entered into Drupal.

---

[13] Editview module – Issues – D7 port of Editview. April 7, 2010. Drupal module issue queue. https://www.drupal.org/project/editview/issues/764882

In some respects, entering data into Drupal was not dissimilar from work Alexander and Zhobov already anticipated undertaking in Microsoft Word as part of their audio chrestomathy. It may have been easier, as there was no need to fuss over table spacing and formatting in Word for the linguistic annotations. The work was, nonetheless, slow, and became slower as the database grew. The growing number of annotated tokens led to an increasing lag in the site's autocomplete functionality, which was necessary to ensure that new texts were able to reference existing tokens, rather than creating new database entries. The possibility of additional metadata beyond what the Word documents would have supported represented another data entry task, and the ways that the database throttled the speed of data entry (e.g. through waiting for the autocomplete) represented a significant increase in the overall time needed to put the material in its final format.

The challenge of data entry at scale is endemic to digital humanities projects. The need for large-scale, low-cost data entry has led projects to adopt practices vis-à-vis undergrad labor that have drawn critique from others in the field[14]. A survey described in "Student Labour and Training in Digital Humanities"[15] shows that the vast majority of digital humanities projects are funded by federal and/or institution-internal grants, which are necessary to offset the many costs of developing these projects, not least among them the cost of paying student workers. Only three of the 40 projects surveyed indicated that they received no funding.

---

[14] Spencer Keralis. "Disrupting Labor in Digital Humanities; or, The Classroom Is Not Your Crowd". In *Disrupting the Digital Humanities.* Dorothy Kim and Jesse Stommel, eds. 2018, Punctum Books.

[15] Katrina Anderson, Lindsey Bannister, Janey Dodd, Deanna Fong, Michelle Levy, and Lindsey Seatter. "Student Labour and Training in Digital Humaniteis". Digital Humanities Quarterly, 2016, vol. 10, no. 1. http://www.digitalhumanities.org/dhq/vol/10/1/000233/000233.html

The first point of the Student Collaborators' Bill of Rights[16] states that "As a general principle, a student must be paid for his or her time if he or she is not empowered to make critical decisions about the intellectual design of a project or a portion of a project (and credited accordingly). Students should not perform mechanical labor, such as data-entry or scanning, without pay." For BDLT, the lack of project funding combined with the City of Berkeley's steep increases in minimum wage over the course of the project ($15/hour as of 2018, up from $11/hour in 2015) made hiring undergraduates for data entry unfeasible. Instead, Alexander has worked with cohorts of students through a longstanding UC Berkeley program, URAP, which connects undergraduates with faculty members doing research. While it may not be the ideal solution, Alexander has devoted significant thought and energy towards collaborating with those students in ways that align with the Collaborators' Bill of Rights.

5.1 URAP program

Since 1991, UC Berkeley has offered the Undergraduate Research Apprenticeship Program (URAP) as an institutional framework "to assist faculty in reconciling their commitments to research with their responsibilities for undergraduate education. By promoting faculty-student research collaboration, URAP works to invigorate undergraduate education and to contribute to the sense of intellectual community on campus."[17] Faculty who wish to participate in URAP submit a project description to an online portal, and students can submit a statement of interest to up to three different

---

[16] Haley Di Pressi, Stephanie Gorman, Miriam Posner, Raphael Sasayama, and Tori Schmitt. "A Student Collaborator's Bill of Rights". June 8, 2015. UCLA HumTech. https://humtech.ucla.edu/news/a-student-collaborators-bill-of-rights/

[17] "What is the Undergraduate Research Apprenticeship Program?" URAP website. http://urap.berkeley.edu/program-intro

projects. Faculty members interview the students and can select any number of them to collaborate on the project.

The BDLT project was ideally suited for this framework. The project description Alexander submitted to the portal outlined the nature of the project, stressing both its linguistic and ethnographic aspects, stated that knowledge of basic linguistic structure was highly desirable but not required, and that knowledge of Bulgarian, or indeed any Slavic language, was not necessary. During interviews with interested students, Alexander gave students an overview of the site and explained how data entry was done. Both the instructor, and the students who decided to choose this project, then completed a "learning contract": the instructor committed to providing a research experience for the students and the students committed to a minimum number of work hours per week throughout the semester, for which they could receive course credit (one credit per three hours of work a week, up to four credits per semester).

The project was first listed with URAP in January 2013, and has been listed every semester since then (except for fall 2017 when Alexander was doing research abroad the entire semester). The largest number to join the project in any one semester was eight, and the smallest was two. Because of the enormous amount of data to be entered, there was never a lack of work for students. Students did data entry on their own time, keeping track of their work hours, and then participated in regular group meetings for discussion of research goals.

The Student Collaborators' Bill of Rights states that "Course credit is generally not sufficient 'payment' for students' time, since courses are designed to provide students with learning experiences." URAP is one of a few programs at UC Berkeley that provides course credit for non-traditional work; another, DECal, grants course credit for student-run courses on topics ranging from "Decode Silicon Valley Startup

Success" and "Sign Language in Healthcare" to "Cal Pokémon Academy" and a master course in the board game "Settlers of Catan"[18]. Providing course credit specifically in exchange for work on faculty research makes the nature of the exchange clear to the student upfront, in contrast to traditional departmental courses that incorporate student labor as a class assignment. Data entry, for all its tedium, is a very authentic research experience, and one that the project directors also engaged with as part of data preparation. In order to make the project available to as many students as possible, regardless of their knowledge of Bulgarian, the project directors provided the data to the students in plain text files with all the tags for coding -- not unlike the original Word files of the audio chrestomathy.

In order to make the project as meaningful as possible, Alexander met twice monthly with student apprentices as a group. In addition to discussing any problems with data entry, these meetings were an opportunity for students to learn more about the history and development of the project, and its importance for Bulgarian dialectology as a research field. Since one of the goals of the very minimal BDLT research budget has been to bring the Bulgarian project director (Zhobov) to Berkeley once a year, some of the student apprentices have been able to meet with him as well and to learn first-hand about the Bulgarian aspects of the collaborative project. This aligns with the Student Collaborators' Bill of Rights principle that "At a minimum, internships for course credit should be offered as learning experiences, with a high level of mentorship."

Students have also been able to participate in project development: their input has been sought on certain aspects of project design, and there was more than one occasion when a student volunteered an idea that led to a particular breakthrough. Whether such contributions amount to students' being "empowered to make critical

---

[18] Decal Courses. Spring 2019. https://decal.berkeley.edu/courses

decisions about the intellectual design of a project or a portion of a project" is arguable, but the project directors' willingness and enthusiasm for reworking the site in response to solicited and unsolicited student input has given these students more agency in the project than simply doing data entry.

5.3 Impact on students

To date, thirty-one undergraduate students have worked on the project through URAP. Their importance to the project is inestimable, a fact of which they are reminded at the celebratory dinner at the end of each semester. In more lasting terms, their contributions are acknowledged on the site's Project Team page (http://bulgariandialectology.org/project-team), which gives a small list of the names of "active apprentices" and an ever-growing list of the names of "alumni apprentices". This is in alignment with the Student Collaborators' Bill of Rights point #4, which states that "If students have made substantive (i.e., non-mechanical) contributions to the project, their names should appear on the project as collaborators".

Many of the apprentices keep in touch after graduation. Two have gone on to graduate work in linguistics, listing participation in this project as a major deciding factor in their career choices. Of the graduate students who have worked on the project, two were specialists in Bulgarian, and project directors created shorter field expeditions in Bulgaria with them in mind, so that each was able to get first-hand experience of the process through which field data are acquired. In addition, the graduate student who was most involved in project design was able to cite his work on this project as an important qualification for his current alternative-academic career path.

The students who have been least satisfied working on the project are those in their final year as computer science majors. It is understandable that they wish to be doing cutting-edge technical work, rather than staying at the level of using (and not

even modifying the code for) a PHP/MySQL content management system. Their dismay and frustration, while somewhat disorienting for other students, has been instructive as a concrete illustration of the ways in which those in the humanities do research with very limited financial resources.

5.4 International collaboration

The BDLT project has been international from the outset, since it grew out of collaborative work between one American scholar and a group of Bulgarian scholars, and has been maintained over the last decade through collaboration between the two project directors, one American (Alexander) and one Bulgarian (Zhobov). The two are in constant electronic contact, consulting over issues of data preparation, data entry, and site design issues (especially with the most recent design additions, concerning "phrases"). They visit each other's universities frequently for purposes of on-site collaboration; Zhobov's visits to Berkeley are especially useful for student apprentices to learn more about international aspects of the project. They have also presented joint research papers about the project at various venues in Europe and Russia.

Most data entry takes place in Berkeley, because of faster computer speed and more modern equipment. Some types of data entry need Zhobov's specialized knowledge, however, and must be done in Bulgaria, despite the fact that it takes place more slowly as a result of network delays, and computers with less memory and older browsers that often perform poorly with the site's AJAX-based data entry interface.

6. Research outcomes

Scholars worldwide have become aware of the rich data resource which BDLT provides, and several research projects are currently utilizing data from the BDLT site. In particular, both Zhobov and Alexander have recently produced major research papers

drawing on material in the BDLT site.[19] Although Zhobov's work on dialectal vocalism could have been prepared directly from the original field tapes, the choice to focus his analyses on texts from the site, and to cite only examples which could then be consulted directly on the site, increase the value of his work to other researchers. Alexander's work, by contrast, on accentuation in certain word groups, derives directly from the data organization in the "phrases" section of the site. Indeed, it is anticipated that this part of the site will be especially valuable to Balkan linguists once the full set of data is entered: they will be able to access dialect data about word order sequences, pronoun reduplication, instances of "evidential" usage, and similar topics. Before the availability of the BDLT site, scholars could only collate data on these topics by laboriously combing through whatever dialectal "texts" had been included as supplementary material to published dialect descriptions: now they will be able to easily search for such material due to the site's search interface.

From a certain angle, it is difficult to argue that the project itself is research. While the investigation of any one research question would necessitate a focused subset of the data preparation involved in creating BDLT, those research questions require scholars to limit the scope of their curation and annotation, and move on to analysis and write-up. In the name of developing a resource of value beyond any one inquiry, or even any one person's research agenda, the project directors have spent the majority of the last eight years focusing on data curation and annotation, which has unavoidably come at a cost with regard to their scholarly output vis-à-vis the kind of dialectological work

---

[19] Zhobov, Vladimir. New Approaches to Bulgarian Dialectal Vocalism; Alexander, Ronelle. Bulgarian Dialectal Accent, A New Approach. Both articles are slated to be published in the forthcoming monograph: Alexander & Zhobov, *Bulgarian Dialects, Living Speech in the Digital Age*.

that is the focus of their pre-2011 scholarship -- work that can now resume fully as BDLT nears completion (nevertheless, the completion of the two major research ventures noted above proceeded alongside work on BDLT). For a scholars who are principally interested in disciplinary research, but who are drawn to the promise of what a resource equivalent to BDLT in their field could provide, the reality is that they will get substantially more research done if they engage in the traditional research practices of focusing their data collection and curation to those materials that contribute directly to a specific inquiry. Building a site such as BDLT is an act of hope, and of generosity for the future scholars who will receive all of the benefit without sacrificing years of their professional lives to data preparation. It is not an undertaking for early-career scholars who can ill afford the impact on their publication rate. At the same time, late-career scholars who are giving thought to the nature and impact of their legacy may find that a project like BDLT, which generates richly annotated data that can jumpstart research for subsequent generations of scholars, is a meaningful gift to the future that reaches far beyond any new monograph.

7. Sustainability

For the project's potential to be realized, the materials need to remain available. While the primary value is in the texts and linguistic annotations, the interface has been specifically designed to facilitate access to the data, reducing the effort necessary to extract the subsets of the corpus relevant for particular research questions. The recent change in licensing terms for the Google Map API, which broke the map-based navigation that the site had used since its inception and necessitated a complete rebuild of the site's geospatial functionality, was a stark reminder of how BDLT is vulnerable to decisions made by large corporations whose interests and priorities diverge from those of the project team. Collaborating with students has added an ethical dimension to

the question of sustainability; per the Student Collaborators' Bill of Rights: "Senior scholars should recognize that projects on which students have collaborated represent important components of students' scholarly portfolios. Senior scholars should thus make every reasonable effort to either sustain a "live" project or, failing this, either transfer its ownership to student collaborators or distribute to students an archived version or snapshot of the project." With the data entry phase of BDLT winding down, and the full picture of the contents of the project becoming clear, the project team is taking a multi-pronged approach to sustainability.

7.1 Website

The Drupal project has announced end-of-life for Drupal 7 in November 2021[20]; ironically, this will also be the end-of-life for Drupal 8, aligning with the end-of-life for the version of the Symfony PHP framework that replaced Drupal's previous technical underpinnings. Moving directly from Drupal 6 to Drupal 8, as BDLT initially planned, would not have bought the project any additional time between migrations.

The digital humanities community's response to Drupal 8 has not been enthusiastic. In addition to the increased difficulty for coders whose skill set does not align with "enterprise PHP development" to build Drupal modules and themes, the server requirements for Drupal 8 to perform adequately outstrip what is typically available in shared low-cost commercial hosting environments. As a result, many digital humanities projects built in Drupal 7 will face a decision point in the next few years, and Drupal 9 is not an obvious choice. One compelling alternative that has emerged in the wake of Drupal 8 is Backdrop CMS (https://backdropcms.org/), a fork of Drupal 7 aimed at nonprofits and small businesses that prioritizes stability and a positive user

[20] Dries Buytaert. "Drupal 7, 8 and 9". September 12, 2018. https://dri.es/drupal-7-8-and-9

experience for non-programmers who build such sites, over technical advancements in the core APIs. The skill set of Backdrop's target audience, and the project's overall priorities, align well with the needs of digital humanities projects. Backdrop has incorporated many Drupal modules that previously had to be maintained and updated separately into its core code, and Backdrop core includes an option to enable automatic updates in order to reduce ongoing support costs and minimize the risk of the site being hacked as a result of delayed installation of security updates. While additional work is needed on Backdrop ports of some of BDLT's modules, Dombrowski has been working with the Backdrop developer community to ensure this functionality is in place in time to migrate BDLT to Backdrop before Drupal 7's end-of-life.

The Berkeley Language Center sees it within their purview to provide access to BDLT indefinitely via through their server and sysadmin, but the scope and priorities of organizations such as a language center are subject to change, particularly in a context of public disinvestment in higher education. While it is not a substitute for the full functionality of a live database that can support any combination of queries, Dombrowski and Alexander are working with digital preservation specialists at UC Berkeley to capture and preserve the site with web recording software once data entry is complete. This will generate a moderately interactive surrogate of the site that can be used in perpetuity for some kinds of information retrieval needs, even if the site itself ceases to exist online.

7.2 Data

The BDLT website includes functionality for exporting any search query as a CSV. With an eye towards capturing the full extent of the data for potential use in computational research, and/or in other interfaces if the website is no longer available, Dombrowski has generated a set of CSV files that include all fields from all content

types, and will update these files with new versions once data entry is complete. Drupal automatically generates a unique ID for each node (instantiation of a content type), and stores references between nodes (e.g. a pointer from a token to a lexeme, or a line to a token) using that unique ID. Including this ID in all exports will make it possible to reconstruct the network of relationships between the various content types in future data analyses and interfaces.

While UC Berkeley does not have a track record of providing an institutional repository for data, the project team anticipates depositing the data sets in such an environment if it is established. The Tromsø Repository of Languages and Linguistics (https://site.uit.no/trolling), which is affiliated with the CLARIN European Research Infrastructure, is appealing as a disciplinary repository. In addition to formally accessioning the data to these data repositories, Dombrowski has followed the common digital humanities practice of putting the CSV files, along with some example analysis code, on the code repository platform Github (https://github.com/quinnanya/bdlt-data).

7.3 Print

The "texts" on the site are valuable pieces of data, even in the absence of the metadata that enables the various search options. To make sure that these texts are preserved at multiple levels, print copies will be produced of all three versions of each text: that with the grammatical and lexical glosses, that in Latin transcription with English translation, and that in Cyrillic transcription. It is particularly important to have both of the latter, since Bulgarian dialectology uses a different set of transcription symbols than those currently used in the West.

8. Conclusion

Over the course of eight years, the *Bulgarian Dialectology as Living Tradition* project

has navigated the full digital humanities project lifecycle, from idea to archiving, without support from external funding. While others may disagree with any of the decisions made during the course of the project's implementation, this paper has served to explicate the motivations, context, and constraints that informed those decisions, to serve as a point of discussion for the development of future digital humanities projects within the broad field of Slavic studies, and beyond. As BDLT transitions from a project to a scholarly resource, the directors hope that this undertaking can lay the foundation for the emergence of a richer understanding of Bulgarian language and culture, and that scholars working in other areas may be inspired to undertake similar endeavors to make materials available digitally -- but only at the right time and place in their careers where such work becomes feasible.

9. Acknowledgements

The authors would like to express their deep appreciation for all their collaborators on this project as of April 2019: senior associate research team member Georgi Kolev; associate research team members Roslyn Burns, Cammeron Girvin, Snejana Iovtcheva, Kea Johnston, Eric Prendergast, Vesela Simeonova, Traci Speed, and John Sylak; apprentice research team members Jessica Adams, Richa Bhandal, Zuhra Bholat, Gabrielle Bozmarova, Nina Chang, Jessica Chapman, Lana Cosic, Katie Crowe, Stephanie DeLeon, Naomi Francisco, Austin Frenes, Dimiana Georgieva, Emmanuella Hristova, Siyana Hristova, Andrew Kuznetsov, Kathleen Lamont, McKayla Major, Kelsey Mota, Grace Newsom, Jerry Nikolaev, Nadia Nizetich, Siyao [Logan] Peng, Stella Petkova, Charles Rosencrans, Elizabeth Sawyer, John Sockolov, Jeffrey Stock, Aleksandrina Stoyanova, Vanessa Taylor, Milena Tintcheva, and Emma Wilcox; fieldwork core team members Georgi Kolev and the late Maksim Mladenov; fieldwork contributors Elena Uzeneva and Georgi Mitrinov; fieldwork assistants Krasimir

Mirchev, Radko Shopov, and Ivan Vankov; fieldwork student apprentices Cammeron Girvin, Marieta Nikolova, Traci Speed Lindsey, Matthew Baerman, Jonathan Barnes, Tanya Delcheva, Elisabeth Elliott, Kamen Petrov, and Petŭr Shishkov. For a complete and current list of project collaborators, see http://bulgariandialectology.org/project-team.

This paper is dedicated to the late Maksim Mladenov, who has been the project's guiding light and guardian angel.