Approaches to Dialect Diversity

Ronelle Alexander
University of California, Berkeley

Introduction

Dialectology, which many in the past viewed simply as the collection of lexical and phonetic curiosities, is now a serious subfield of linguistics. The richness of dialectal diversity provides highly valuable information not only for historical and typological linguists, but also for other aspects of language study. Given that a "dialect" is most commonly defined as pertaining to a particular region, it is clear that the most effective way to portray this diversity is by picturing it in cartographic terms, or as "linguistic geography."

The tool of choice for this portrayal has long been the dialect atlas. It is commonly thought that "dialect atlas" as a concept is straightforward: a questionnaire is administered to speakers of a number of different dialects, and the diversity of their responses is depicted on a series of maps in the form of individual symbols, isoglosses or both. But in fact there are many different approaches to the cartographic representation of dialectal diversity: just as in almost any scientific endeavor, choices must be made, and these choices unavoidably channel both the investigation, and the presentation of its results, in different directions. These choices in turn inevitably influence the eventual presentation and interpretation of the data.

Dialect Atlases: The Beginning

Such a difference can be seen already through a juxtaposition of the first two major projects of dialect cartography in Europe, one devoted to German dialects and the other to French dialects. The German effort, begun in 1876, resulted in the publication of what was in fact the world's first linguistic atlas, *Der Sprachatlas des deutschen Reichs* (Wenker 1881), while the French-Swiss effort, begun in 1896, resulted in the thirteen-volume *Atlas linguistique de la France* (Gilliéron and Edmont 1902-1910). Each of these undertakings attempted to depict the dialectal diversity of the language spoken within the political unit whose name corresponded to that language (a political unit which was assumed in each case to be monolingual in that language). The collection methods were very different, however.

The initiator and sponsor of the German undertaking, Georg Wenker (1852-1911), devised a questionnaire containing around 40 short sentences. In 1876, he sent this questionnaire by mail to some 1,500 schoolmasters in the Rhine valley, requesting each to translate the sentences into the local dialect. Encouraged by the responses, he then extended this survey throughout Germany (over the period 1877-1887), eventually collating over 45,000 responses. Although the responses were of variable quality, the survey had the obvious advantage of thoroughness of coverage, since by involving local individuals who held responsible positions (schoolmasters), he was able to elicit responses from each and every German village which had a school. At the same time, this method has obvious disadvantages, since there is no direct contact with actual dialect speakers. All Wenker had were the written responses of the various schoolmasters, whose awareness of dialect (and reliability as evaluators of dialect speech) he had to take on trust.

The French effort, by contrast, relied on direct elicitation. In 1896, its initiator, Jules Gilliéron (1854-1926), devised a questionnaire of around 1,500 items. Over the next four years (1896-1900), he sent fieldworkers throughout the villages of France to administer this questionnaire directly to dialect speakers. In fact, the vast majority of the material was recorded by a single investigator, Edmond Edmont. This indefatigable researcher visited 639 localities and recorded in all 700 interviews. The clear advantages of this method are obvious: there were not only the facts of a broader questionnaire and direct contact with informants, but also (for the material collected by Edmont) the consistency of interview methodology and of phonetic transcription.

Dialect Atlases of Slavic Languages

Dialect atlases in the Slavic-speaking countries appeared considerably later: work on the vast majority of them was undertaken only after WWII (with one notable exception: the sub-Carpathian dialect atlas published by Polish scholars [Małecki and Nitsch 1934]). The creation of questionnaires and the use of them to canvas rural regions can be dated to considerably before that, however: the Moscow

Dialect Commission, working in the years prior to WWI, produced what they called an "attempt at a dialectological map of Russian" (Durnovo *et al.* 1915). Other statements of dialect geography were also published in these early years, such as Aleksandar Belić's "dialect map of Serbian" (Belić 1905). The key point, however, is the use of a questionnaire on a consistent basis. This is the factor which distinguishes the Moscow-based effort.

The post-war socialist period saw an explosion of accomplishment in dialectology, thanks in part to extensive support from socialist governments. Intended as an overarching effort, the *All-Slavic Linguistic Atlas* (*Obščeslavjanski lingvističeskij atlas*, or "OLA") was formally initiated by Soviet linguists in 1958, who directed the formation of dialectological commissions in each of the separate countries to work according to a common, standardized questionnaire. Although a fully complete atlas may never be attained, many valuable publications have come out under this rubric, including several volumes of maps. Within the individual Slavic countries, the fullest coverage was achieved in Bulgaria, Poland and Ukraine, but much work was done in all Slavic countries, though some of it still remains unpublished (for a listing of achievements up to 2003, see Alexander 2006:35-38). Although there were differences in the presentation of the material on maps, the unifying factors underlying the construction of these research tools were the consistent use of a single questionnaire and the dual focus of these questionnaires. One of these goals was the search for dialectal forms providing evidence for historical Slavic phonology and morphology, and the other was the collection of dialectal vocabulary. The fact that answers to both these types of far-ranging questions could be exemplified by single words made it possible to construct highly detailed maps displaying a plethora of information.

The field research underlying these gargantuan efforts was done in the traditional manner, by sending fieldworkers out to interview dialect speakers in the village context; records were made by hand in field notebooks and transferred later onto small cards or slips of paper, which were then collated to provide the base data for the maps in each atlas. When it became practicable to take portable tape recorders to the field, mechanical recordings of actual speech were made. At the beginning these recordings were intended simply to supplement the primary data collection and to provide samples of connected speech; as time went on they became more and more central to the enterprise. This created an extra layer of work, of course, since the recordings must first be transcribed before individual data items could be excerpted from them. At the same time, this additional layer led to the discovery of much finer distinctions, especially at the phonetic level, since it allowed for multiple listenings and collaborative analysis.

Dialectology in the Internet Age

With the advent of the internet, a completely new level of dialectological research came into being. Now, with the possibility of direct access to audio recordings of field data, it was these audio files that took central focus. Readers who had earlier consulted a print version of a dialect atlas (or the print versions of individual dialect descriptions) now could interact with a website that allowed them direct access to the audio recordings of (presumably) unedited stretches of natural dialect speech recorded in the informant's home environment – either inside his or her actual home, or somewhere in the village that was a natural place for conversation to occur. These texts provide the first opportunity for researchers who are interested in elements that transcend the word level (such as syntax, intonation, word order, discourse structure and the like) to study these elements in spoken dialectal speech. Furthermore, users of such sites could work with primary material, rather than trusting the choices and analyses of the authors or compilers of print sources. Such material is also valuable at a more existential level, in that it brings the focus of dialectology back to its origin by allowing users to encounter any one dialect not as disembodied isoglosses or word lists, but as a real, natural human speech system which is the primary means of communication for its speakers.

However, this very ability of the internet to provide so many more possibilities means that many more choices must be made. To illustrate the range of these choices, I shall survey seven sites, four of which are structured around audio files of dialect speech recorded in context, and three more of which, although they also provide access to audio files, do not apparently see this as their main goal. All sites provide, in addition to audio files, supplementary materials of various sorts. Of the four sites in the first group, two present recordings of Russian dialects and two of Bulgarian dialects; all four represent collaborative efforts between Western scholars and those from the country in question. The two sites devoted to Russian dialects are the *The Language of the Ustja River Basin: A Corpus of North Russian Dialectal Speech* (covering dialects in a relatively compact area south of Arxangel'sk, henceforth "URBC") and the *Russian Regions Acoustic Speech Database* (covering a wide range of dialects throughout Russia, henceforth "RuReg"); and the two sites devoted to Bulgarian dialects are the *Transdanubian Electronic Corpus* (covering Bulgarian dialects spoken in Romania, henceforth "BDR"), and the website *Bulgarian Dialectology as Living Tradition* (covering dialects throughout the current borders of Bulgaria, henceforth "BDLT").

The three in the second group present material about Polish, Bulgarian and Macedonian dialects, respectively. All are the result of work exclusively by native scholars within the countries in question, and each covers the full dialectal range of the country in question. That devoted to Polish dialects bears the title *Dialekty i gwary polskie, kompendium internetow* (henceforth "DGP"), that devoted to Bulgarian dialects bears the title *Karta na dialektnata delitba na bŭlgarskija ezik* (henceforth "KDD"), while that devoted to Macedonian dialects appears on the page bearing the title *Digitalna zbirka na tekstovi od makedonskite dijalekti* and is called simply "Mapa na dijalekti" (henceforth "MMD").

Sites to Be Discussed: Overview

URBC [http://www.parasolcorpus.org/Pushkino/login.php] represents a collaborative effort between Ruprecht von Waldenfels, formerly of Zurich University and now at the University Oslo, and Michael Daniel and Nina Dobrushina of the Higher School of Economics in Moscow. It is structured around audio files of dialect texts recorded by joint Swiss-Russian field teams in the summers of 2013, 2014 and 2017.

RuReg [http://rureg.fh-bochum.de/de/] is a project headed by Christian Sappok of the Ruhr Universität Bochum, with the collaboration of scholars from Moscow, Kirov, Oxford, Leipzig and Bergen. It is structured around audio files of dialect texts recorded by different groups of the collaborating scholars in many different areas of Russia between 1991 and 2007.

BDR [http://www.corpusbdr.info] is a site created by Olga Mladenova of the University of Calgary and Darina Mladenova of Sofia University. Its full title is "Transdanubian Electronic Corpus: Supplement to *Bulgarian Dialects in Romania* by Maxim Mladenov." It is structured around audio files of dialect texts recorded by Mladenov and collaborators between 1962 and 1975 in Bulgarian-speaking communities within southern Romania.

BDLT [http://bulgariandialectology.org/] represents a collaborative effort between Ronelle Alexander of the University of California, Berkeley, and Vladimir Zhobov of Sofia University. It is structured around audio files of dialect texts recorded by a joint Bulgarian-American team headed by Alexander, Zhobov and Georgi Kolev (also of Sofia University) between 1990 and 2013 with supplemental material recorded by Zhobov and Kolev between 1986 and 1989.

DGP [http://www.dialektologia.uw.edu.pl/index.php?11=start] represents a massive project headed by Halina Karaś of Warsaw University. Although it includes numerous files audio files of dialect texts representing all major dialects of Polish, its aim is as the title states: to provide a full “compendium” of information about Polish dialects and dialectology.

KDD [http://ibl.bas.bg//bulgarian_dialects/] was produced by the Dialectology section of the Institute for Bulgarian Language of the Bulgarian Academy of Sciences. It is a full-color map of the distribution of Bulgarian dialects. For every dialectal region one can click on a link to a short pdf file summarizing that dialect’s main features; for some of these, one can also click on a link leading to a brief audio track of a recording made in that region.

MMD [http://ical.manu.edu.mk/Map/Map.html] was produced by the Research Center for Areal Linguistics of the Macedonian Academy of Arts and Sciences, under the direction of Marjan Markoviḱ. It is a Google map of Macedonia with tabs for every village from which a text has been transcribed, plus eleven tabs which lead to an audio recording from that village. With some detective work one can access elsewhere on the site fourteen more audio recordings.

As might be expected, the four sites in which Western scholars have played major roles present basic instructions and all supplementary materials in English (except for RuReg, which presents some of its material in German only), the intent being clearly to make materials of Slavic dialectology more accessible to the Western scholar. In some cases native-language versions are also available. In the case of the other three sites, everything is in the language in question (Polish, Bulgarian or Macedonian). As with nearly any site on the internet now, however, translation is available at a single click. Since the software which produces such translations relies on dictionaries of the standard languages, this translation technology can be helpful when it concerns descriptive material or scholarly analysis written in the standard language; it would not make sense to rely on it for translation of dialect texts, however.

The four collaborative projects are “works in progress”: each has ambitious goals, and although each has made sufficient progress towards these goals to warrant opening up the site to the general public, each is incomplete in a number of ways. It appears clear in each instance, however, that those responsible for the site are continuing work towards their respective goals. The three in-country projects, by contrast, seem each to be more or less content with the representation given of the dialectal landscape in question, though each also does indicate the possibility of further work.

For instance, the Polish site (DGP), by far the most thorough and extensive of the three, gives quite a full picture of Polish dialects. The authors do note, however, at the end of the page describing the structure and content of the compendium that the nature of an internet site allows for it to be supplemented by further dialect and ethnographic materials. This indicates that although they consider the current version (dated 31.XII.2010) to be complete as is, they do not rule out the creation of an expanded version. As to the Bulgarian map (KDD), the accompanying commentary states that it is the first stage of a project for an electronic interactive map; the implication is that many more links of the same type (one link to a single page summary of traits and another to an audio sample) will be added to the map. Commentary to the Macedonian map (MMD) simply states that the 25 existing audio files are accessible from the map: since in fact only eleven are currently accessible from the map, it would seem that the next step would be to post links on the map to the remaining fourteen. Elsewhere on the site, however, is a note referring to the existence of many more recordings of dialect speech that are presumably waiting to be digitized.

Description of the Seven Sites

Since the feature common to all these sites is the immediate availability on the internet of audio recordings of dialectal speech, a comparison of the ways in which each site makes these files accessible to the user (beyond simply providing playback) is useful. Here, too, the division between collaborative projects (the first four) and in-country projects (the latter three) is striking, in that each site in the first group conceives of itself as an electronic corpus or database and processes texts in some way so as to allow researchers to interact with them directly. By contrast, the other three sites simply use the audio-visual capabilities of the internet to present a traditionally-oriented description of the relevant dialectal landscape in more user-friendly ways. Audio samples are provided as illustration, but the major focus in each case is on visual representation. I shall discuss these three in-country sites first.

(1) ***KDD****: Karta na dialektnata delitba na bălgarskija ezik*

The central focus of the Bulgarian undertaking (KDD) is the large and detailed map, which intentionally uses color coding to emphasize the major division between eastern and western dialects, and carries this coding through in the several subdivisions. Within each of these subdivisions on the map there is a small icon of an open book: this takes one to a short pdf file describing one particular village's dialect situated within that group, according to a standard seven-line template. The first line names the village and gives its identification number with the files of the massive *Bulgarian Dialect Atlas*; the next five identify the form of the most basic indicators of dialect grouping in Bulgaria (reflexes of the back *jer*, of the back nasal, of *jat* before back vowel, of *jat* before front vowel or soft consonant, and the form of the future tense particle); and the final line gives a sample transcribed utterance in the dialect. If a small icon of a sound speaker is present, this leads one to a short audio file. There is no visible relation between the text and audio files. Usually they represent different villages altogether, though in the very few instances when the two icons present material from the same village, the content of the audio file bears no relation to the transcribed utterance on the pdf file. No transcription of material on the audio files is given, nor is anything translated into a Western language. In short, it appears that the aim of KDD is to present the basic information about Bulgarian dialects in capsule form on a visually attractive map, and to allow speakers of Bulgarian to hear brief samples of various sorts of dialectal speech.

(2) ***MMD:*** *Digitalna zbirka na tekstovi od makedonskite dijalekti, Mapa na dijalekti*

The central focus of the Macedonian instance (MMD) is the volume of transcribed texts to which the map is keyed. The map itself gives no information about dialectal divisions; it is simply a background upon which to place tabs that direct one to a printed text exemplifying the dialect located in that spot. Clicking on a tab takes one to the relevant page in the lengthy pdf file of the print volume of dialect texts referred to in the site's title (Vidoeski 2000). This is a compendium of dialect texts written down by various hands over the period of a century or so. Most are drawn from sources published between 1890 and 1998 (sources which are documented in the book's index), but a number are from Vidoeski's unpublished field notes. The contribution of the site is to match a particular village, seen in visual geographical terms, with a text representing its dialect. As far as can be judged, there is no relationship between the contents of any one of the few audio files that have been posted and any of the transcribed texts. No transcription of the material on the audio files is given, nor is anything translated into a Western language. Here, too, it appears that the aim of MMD is to present the most detailed compendium available of transcribed Macedonian dialect texts in a form that allows users to coordinate texts with villages and regions and to allow speakers of Macedonian to hear brief samples of various sorts of dialectal speech.

(3) ***DGP****: Dialekty i gwary polskie: Kompendium internetowe*

Each of the above two sites consists essentially of a single main page displaying a map, from which additional information can be gained by clicking on particular tabs. By contrast, the Polish site (DGP) takes full advantage of the layering capacities of the internet to present extensive information about all aspects of Polish dialectology: it includes a general introduction to dialectology as a science, an overall map of major dialect divisions and detailed maps of each separate region, ethnographic information and extensive dialect descriptions. Within each section devoted to an individual dialect type are one or more pages of texts. Each of these pages contains a single excerpted text from a recorded session with a particular informant: both the village and the informant are identified fully, and – if s/he had given permission – a photo of the informant is provided. The audio file which is the central focus of that page is accompanied by a full transcription (a list of phonetic symbols used is given in the introductory section of the site). Additionally, dialectal words of interest are annotated within the text by means of floating windows accessed by hovering over the word in question. Clearly, the aim of the very large team which brought this site to fruition is to use the multimedia capabilities of the internet to give a full account of Polish dialectology and dialects. The fact that the many individual audio files are transcribed and that each word of interest to a dialectologist is provided with annotations makes the site very valuable to researchers interested in linguistic questions of analysis at the phrase, sentence or discourse level. At the same time, everything (including the annotations) is in Polish; nothing has been translated into a Western language.

As noted above, the four sites which represent collaboration between Western and native scholars differ considerably in that the main goal of each is to provide researchers with the raw material of recorded dialectal speech and to shape this material in such a way as to give researchers different types of access to the contexts of each segment.

*(4) **BDR**: Transdanubian Electronic Corpus*

Of the four, BDR is the most traditional, which is consonant with its stated goal to "supplement" a major print study of Bulgarian dialects in Romania (Mladenov 1993). The site contains well-researched and documented pieces about diaspora dialects, the sociolinguistic situation of Bulgarian speakers in Romania, Bulgarian dialectology in general, dialectal phonetics and the like. With respect to the data, the main organizing principle is location. Each village in which Mladenov and colleagues did their field work is described in some detail (accompanied by a gallery of photos), and all of the texts recorded in that location are available under the main location tab. Texts themselves are presented together with the audio link: as in DGP, the audio link is directly above the transcribed text. Each text is broken up into sections corresponding to the major topics discussed, and the audio is likewise broken up into these sections. The transcription of the text is in Bulgarian Cyrillic only; no translation is provided, nor are any of the individual forms annotated in any way.

The extensive texts in BDR can be searched in three different ways. The first is standard in corpus linguistics: searching for a lemma brings up each instance of that lemma in a context of specified length (one can choose to see six, 12 or 24 words on either side). One can search the entire site, or one can limit one's search to a particular location (or locations). The list of results gives only the words in context, but the relevant coordinates (locality, theme ID number and line number in the text) can be found in a floating window above the form. The second search is by thematic content: all portions of the texts have been tagged for content according to a quite detailed list. Searching for any one of these content tags brings up a list of indicators as to which lines in which text speak to that content. The third search is related to the fact of frequent code-switching between Bulgarian and Romanian, since all speakers are bilingual. All Romanian speech segments appear together with translations in Bulgarian; users can search for any words in such sections by either Romanian or Bulgarian gloss.

User access to BDR is incomplete. It appears that all texts have been digitized and that most if not all have been transcribed and tagged for content. Very few of them have actually been uploaded; however, until recently there has been no way for users to tell which are the village pages for which material has been uploaded other than to try each one in the hopes of finding text and audio material. A recent and very welcome addition to the site is a page with this information.

*(5) **RuReg**, Russian Regions Acoustic Speech Database*

Of the four sites, RuReg is the one with the most extensive range of information and the broadest geographical coverage. It is organized according to expeditions. The home page displays a map of Russia with thirteen points on it marked in red, each corresponding to one of the team's expeditions. In fact, however, the site includes data from many more expeditions. Under the tab Regions, one finds two sets of maps, one of European Russia and one of Siberia; each displays many more points, and the list below identifies each of the points and itemizes briefly expeditions made in that region. Detailed information about expeditions is found under the tab Expeditions, where 37 different expeditions are listed. Clicking on any one in principle takes one to a fascinating page, which describes the expedition in some detail, together with illustrations, under the subsections Planning, Motivation, Route, History, Region, People, Speech, Culture and Info. As of this writing (December 2017), however, only two such pages have been posted, and it is highly unfortunate that there is no indication to the user as to which two, other than the fact that one of the two is used as an illustration in the section under Instructions.

The access to audio files in RuReg is extensive; however, there are presently 59 recordings in the database, each quite lengthy and each divided into tracks. They are grouped by region, and then by expedition within region. Clicking on the track name takes one not just to the audio track, but also to an oscillogram showing the waveform of the audio. Even more valuable is the fact that one can select and download specific segments of the audio, and give each segment its own referential code, which will then allow anyone to find that segment again within the database (the authors of the site call this an “acoustic citation”). The texts appear to have been annotated on a number of different levels, possibly more than the search page as currently constituted allows access to (at least according to Sappok 2010, a link to which is on the site). The search function page allows three types of interconnected searches, by keyword, by social factors (such as region, and the age, education and occupation of the informant) and discursive factors, the use of which is not immediately evident (especially as none of the promised subdomains appear to have been activated). Search results give the lemma in context, spell out the descriptors and give the identification tag of the keyed utterance, which is also displayed in oscillogram format.

It is apparent from search results that texts have been transcribed; it is unclear, however, how one might access these transcripts in any form other than the individual short segments generated as search results, or as samples in the Speech section of any one expedition’s descriptive summary, where speech samples are given together with translation. There does not appear to be any way to access the full running text of any of the transcripts, except for the 15 expeditions which are labeled BFF on the Expeditions page. This initially confusing abbreviation is deciphered at the bottom of the page by links to “BFF texts,” which are pdf files of publications in the series *Bjuletin’ fonetičeskogo fonda russkogo jazyka,* whose subtitle, *Zvučaščaja xrestomatija*, suggests that the publications either were originally accompanied by CD discs, or that they referred to the existence of cassette tapes which were to be made available in some form.

There is no indication as to how many more audio files remain to be posted in RuReg, nor as to whether there is a way for users to access complete text transcripts other than those in the BFF series. It is certainly to be hoped, however, that the remaining expedition pages will be posted, as the two descriptions now online offer extremely valuable insight into the process of fieldwork. Indeed, they make the expeditions come truly alive, something that very few resources on dialectology are able to do.

*(6) **URBC**: Ustja River Basin Corpus*

The other site dealing with Russian dialects, URBC, covers a more modest region than RuReg. Furthermore, it offers very little explanatory material about the regions visited other than maps of the area and a spreadsheet giving basic biographical information about each of the informants (plus an index noting how much speech was recorded from each informant). Instructions about use of the site are very scant on the site itself, although the site does include a link to a paper which describes it in more detail (von Waldenfels *et al*. 2014).

The focus of URBC is on the data, and the site is organized in terms of searches. Visitors to the site can search the data in one of three ways. The basic search is for any attestation of a lexeme, the advanced search allows one to specify a grammatical part of speech and also to search for two words at variable distance from one another, and the complex (CQP) search allows one to formulate queries in that specialized format. Results are in the form of a list: each entry gives the highlighted form in brief (one line) context, the audio of that line, the code number of the speaker (which links one to the spreadsheet of metadata), and the date the file was entered into the system. There are also buttons for CSV-export view of that line and a button that displays the cited form in broader context; this last view has its own URL for later retrieval. Full texts can be viewed by researchers who have registered with the site personnel.

Given the short time that URBC has been in existence, the amount of material it contains is remarkable. There are 124 texts (most quite lengthy), all of which have been transcribed (in Russian), annotated and provided with audio. A link to the full audio of the text is at the top, but because the texts have been entered in ELAN format, there are also links at the end of each line of text allowing one to jump to the audio of that line. The transcriptions are in standard Russian; the site authors defend this transcription choice by stating that it allowed their team to transcribe and post material at maximum speed, and that researchers interested in phonetic detail can listen to the audio and make their own more detailed transcriptions. Registered users are also given the opportunity to submit corrections in the transcription to the authors, and users are advised to listen to the audio before quoting any transcribed passage.

*(7) **BDLT**: Bulgarian Dialectology as Living Tradition*

The final site to be discussed, BDLT, returns the discussion to Bulgarian dialects. Of the four, this site is the one most directly aimed at Western users, in that it provides English translation of all dialect texts. The organizing principle of this site is the individual text (each of which is named for the village in which it is recorded, with an additional identifier in the case of multiple texts from any one village). In contrast to the other three sites surveyed above, which post the audio of field tapes in their entirety, the 181 BDLT texts are carefully chosen excerpts from field tapes. The total amount of audio material available is thus considerably less than the other three sites discussed; what distinguishes this site is the correspondingly greater detail in the processing of these texts. Not only is each available both in Latin and Cyrillic transcription, but the Latin transcription is also available in two different views, one with interlinear annotations visible beneath each token and the other with simple text (to enable “distraction-free” reading for content). Both Latin transcripts are provided with a line-by-line English translation, and all three views are accompanied by a floating audio link.

Texts can be searched in a number of different ways. One can ask for the occurrence of single words (either by English gloss or Bulgarian lemma); one can ask for a combination of grammatical and pragmatic traits, drawing from a fairly extensive list; one can ask for all phonetic instances of any one lemma; and one can ask for a variety of lexical traits, including the provenance of loan words. Finally, one can ask for words which display any one of a number of “linguistic traits” (elements which the site authors determined were or could be of interest to one studying Bulgarian dialect diversity); this latter option allows one to do complex searches which place optional conditions of environment and realization. Most searches give the results by displaying the token in the context of its occurrence; the “linguistic trait” search also plots results on a map. All search results are tagged in such a manner as to allow users to move immediately to the text and its audio, to see and hear the form in its spoken context.

The criteria governing the excerption (from the full corpus of field recordings) of texts which appear on BDLT were two in number: each text should illustrate as many of the salient features of the dialect as possible, and each text should constitute a self-contained chunk of discourse whose content reflects traditional life. Site authors are currently coding the texts for different features of thematic content. When this is complete, users will also be able to search the database for chunks of text by content.

Comparison of the Sites

The factor common to the seven sites surveyed above is the ability of users to access directly audio recordings of dialectal speech. The sites fell into two different groups, those which utilized audio clips as secondary illustrative material within the overall representation of the dialectal facts of a particular language, and those which viewed these audio recordings as basic research data and organized the site around them. Within the first group, the Bulgarian and Macedonian sites (KKD, MMD) consist basically of a single map each, with links to textual information available elsewhere but now correlated visually with geographical points, and to brief audio clips with no transcription. The Polish site (DGP) gives a much more detailed and complex overview and includes with each audio file a transcript, informant metadata and annotation of dialectal forms.

Within the second group, three of the four (BDR, URBD, RuReg) took the corpus approach as their basic principle and uploaded field tapes in their entirety, while the fourth (BDLT) chose to limit the material to selections from the field tapes, but in recompense to provide much more information about each excerpt than the other sites give for their material. All of the four provide a means for users to focus on specific sections within the audio material and to search the database on various parameters, but the possible searches are quite different in nature. All allow searches for individual lemmas but only the two Bulgarian sites (BDR and BDLT) allow searches for thematic content as well. One Russian and one Bulgarian site (URBC and BDLT) allow searches to specify particular grammatical tags; tags in the former allow one to specify only part of speech whereas those in the latter give a broader and more detailed list of choices. Both Bulgarian sites also allow additional types of searches: BDR, in response to the bilingual nature of the data, allows searches on Romanian glosses as well, while BDLT allows searches for numerous other factors of interest to researchers, such as the source of loanwords, the reflexes of particular old Slavic vowels and consonants and many others.

All sites allow a means for one to navigate within the audio files and to locate specifically the point within the audio file that contains material of interest. Here RuReg offers something the others do not: an oscillogram which displays sound waves and which furthermore allows a user to capture a specific segment of these sound waves and create an individual file from it. URBC, by using the ELAN notation system, allows users to locate individual pieces of speech with a similar degree of precision, but not to make individual captures of these segments. BDLT indexes each line of text with the time-code of its place within the audio file, thus allowing users to find (and hear) in context any token that appears in any of the search lists. URBC bypasses this intermediate step by making the audio clip of any one searched line available directly on the list of search results. BDR is the least precise from this point of view: mousing over any result of a search gives one the text name, thematic chunk and line number; however, one must then listen to the entire audio of that chunk in order to hear the form in question.

As stated earlier, the central focus of each of these sites is the audio files taken directly from field recordings, but the degree of this focus varies among the sites. It is the most “rigid” in RuReg, which offers not only the audio but also a visualization of the audio in the form of oscillograms. Other than the brief textual excerpts which appear within search results, or in the “speech” samples found on the summary “expedition” pages, RuReg provides no transcription at all. The other three sites do provide transcriptions of the audio files, but there is considerable variation as to the form. BDR and URBC provide transcription in Cyrillic script only in the native language (Bulgarian and Russian, respectively), while BDLT provides transcription both in the Latin and Cyrillic scripts. URBC deviates from tradition by transcribing dialect texts using only standard orthography (for commentary, see von Waldenfels *et al.* 2014). The two Bulgarian sites use the phonetic transcription accepted among Bulgarian dialectologists in their Cyrillic transcriptions; BDLT uses in addition a Latin script which is a modified version of IPA (marking vowels with IPA symbols, but consonants and accent with symbols accepted in Slavic transliteration). Of the four sites, only BDLT also provides consistent translation into English of all transcribed material.

Finally, the degree of metadata is variable, as is the presentation thereof. Both BDR and RuReg offer detailed descriptions of various aspects of fieldwork, amply illustrated by color photos; BDR also includes short articles on various issues relating to Bulgarian dialects in Romania with extensive bibliography. Both these sites, and BDLT, include pages describing the use of the site in some detail (some additional background on BDLT is given in Alexander 2015, and a thorough description of RuReg is given in Sappok 2010). URBC, by contrast, offers little or no information about use of the site, though it does give a link on the home page to an article containing some of this information (von Waldenfels *et al.* 2014) and requests users to cite both the site URL and this article. On the other hand, URBC is the most thorough of the four when it comes to documentation about informants: its spreadsheet (accessible from the home page) gives extensive information about each informant. BDR also includes a page listing all informants with relevant metadata, though in somewhat less detail, and names each informant on the appropriate village page. RuReg also gives this information, though only indirectly, as part of individual search results. BDLT is the least informative in this respect, simply identifying informants as to locale and gender. It does, however (as do all sites), specify the date of recording and members of the recording team.

Conclusion

The internet, with its significant increases in technology and vast ability to reach nearly everyone, has changed many undertakings radically, and dialectology is no exception. The ability to make field recordings available to anyone who wishes to hear them and to provide all sorts of additional research tools along with these audio files has given users immediate and direct access to something which used to be available only to field researchers themselves: that which is encompassed in the full title of BDLT: “Bulgarian dialectology as living tradition.”

It is clear that internet technology is changing very quickly, indeed so quickly that in a few years – perhaps even soon after this report appears – there may be new sites and new methods on the horizon. The goal of this report has been to record what is available at this point (if only as historical record) and to observe again how the choices one must make in approaching one’s data affect the presentation, perception and ultimate use of those data.

References

Alexander, Ronelle. 2006. “Dialectology,” *Slavic Linguistics 2000: The Future of Slavic Linguistics in America*, Franks, S., *et al*. (eds), 52 pp. Available online at [http://slaviccenters.duke.edu/sites/slaviccenters.duke.edu/files/media_items_files/8alexander.original.pdf].

________. 2015. “Bulgarian Dialectology as Living Tradition: A Digital Resource of Dialectal Speech,” *Balkanistica* 28, pp. 1-13.

Alexander, Ronelle, and Vladimir Zhobov. 2011-2016. *Bulgarian Dialectology as Living Tradition*. Available online at [http://www.bulgariandialectology.org/], last accessed December 30, 2017.

Belić, Aleksandar. 1905. “Dialektologičeskaja karta serbskago jazyka,” *Sbornik po slavjanovedeniju* 2. St. Petersburg: Tip. Imperatorskoj Akademii Nauk.

Daniel, Michael, Nina Dobrushina and Ruprecht von Waldenfels. 2014. *The Language of the Ustja River Basin: A Corpus of North Russian Dialectal Speech.* Bern, Moscow. Electronic resource. Available online at [http://www.parasolcorpus.org/Pushkino/login.php], last accessed December 30, 2017.

Durnovo, N.N., N.N. Sokolov and D.N. Ušakov. 1915. *Opyt dialektologičeskoj karty russkago jazyka v Evropě s priloženiem očerka russkoj dialektologii*. Moscow: Sinodal'naja tip.

Gilliéron, Jules. 1902-1910. *Atlas linguistique de la France*, vols. 1-13. Paris: H. Champion.

Gilliéron, Jules, and Edmond Edmont. 1902-1910. *Atlas linguistique de la France*, 35 fasc. de cartes.

Institut za bălgarski ezik. 2014. *Karta na dialektna delitba na bălgarskija ezik.* Available online at [http://ibl.bas.bg//bulgarian_dialects/], last accessed December 30, 2017.

Karaś, Halina (ed.). 2010. *Dialekty i gwary polskie, kompendium internetowe.* Ministerstwo kultury i dziedzictwa narodowego. Warsaw. Available online at [http://www.dialektologia.uw.edu.pl/index.php?11=start], last accessed December 30, 2017.

Małecki, Mieczysław, and Kazimierz Nitsch. 1934. *Atlas językowy polskiego podkarpacia. Cz. 1: Mapy, Cz. 2: Wstęp, objaśnienia, wykazy wyrazów*. Kraków: Polska Akademia Umiejętności.

Markoviḱ, Marjan (ed.). 2017. *Digitalna zbirka na tekstovi od makedonskite dijaleti, Mapa na dijalekti.* Research Center for Areal Linguistics, Macedonian Academy of Sciences and Arts. Available online at [http://ical.manu.edu.mkMap/Map.html], last accessed December 30, 2017.

Mladenov, Maksim. 1993. *Bălgarskite govori v Rumănija.* Sofia: Izdatelstvo na BAN.

Mladenova, Olga, and Darina Mladenova. 2001-2013. *Transdanubian Electronic Corpus. Supplement to Bulgarian Dialects in Romania by Maxim Mladenov*. Available online at [http://www.corpusbdr.info/], last accessed December 30, 2017.

Sappok, Christian. 2010. “Russische regionale Varietäten und Dialekte – eine akustische Datenbank mit diskursiven Annotationen,” *Wiener Slawistischer Almanach* 65, pp. 163-90.

Sappok, Christian, Alexander Krasovistskij, Ludger Paschen, Katrin Brabender, Andreas Koch and Nadja Kühl. 2016. *RuReg: Russische Regionen. Akustische Datenbank/RuReg: Russian Regions. Acoustic Database*. Available online at [http://www.rureg.hs-bochu.de/en/], last accessed December 30, 2017.

Vidoeski, Božidar. 2000. *Tekstovi od dijalektite na makedonskiot jazik*. Skopje: Institut za makedonski jazik “Krste Misirkov.”

von Waldenfels, Ruprecht, Michael Daniel and Nina Dobrushina. 2014. “Why Standard Orthography? Building the Ustya River Basin Corpus, an Online Corpus of Russian Dialect,” *Komp’juternaja lingvistika i intellektual’nye texnologii, po materialam ežegodnoj Meždurarodnoj konferencii “Dialog”* (Bekasovo, 4-8 ijunja 2014 g.), vyp. 13 (20). Moscow.

Wenker, Georg. 1881. *Sprachatlas des deutschen Reichs*. Marburg, London.

Wenker, Georg, und Ferdinand Wrede. 1895. *Der Sprachatlas des deutschen Reichs: Dictung und Wahrheit*. Marburg: N.G. Elwert.