General Online Dialect Atlas

Sheila Embleton, Dorin Uritescu, and Eric S. Wheeler

York University, Toronto, Canada

His voice was rather rich and dark; the accent was Middle Western, but underneath the nasalities there was something soft and furry that came from the South.

-- Mary McCarthy. 1942: 67

A dialect atlas is a book of maps containing data from a linguistic area or selected geolinguistic interpretations illustrating dialect relationships, but an online dialect atlas is much more: it is a repository of data, a broad means to access and process the data, and a dynamic presentation of results in multiple forms.

Modern Information Technology has changed the way we present and use dialect atlases. Here is a proposal for a general online dialect atlas, and an account of some of the experiences that have led us to this proposal.

Background

Traditionally, a dialect atlas is a set of hard-copy maps, each of which displays data (the "data set") related to one or more points of linguistic interest (the "prompt") at a set of geographic locations (the "location"). In some cases the prompt is based on the field worker's elicitation question (e.g. "What do you call the building that you keep the cows in?") and the data set holds the "raw" answers received. In other cases, the map and the prompt reflect a point of analysis (e.g. " /k/ vs /k'/ in word-final position) and the data set has some summary or interpretation of the raw data .

In any case, the reader of the atlas has ready access to the information that the author has selected to present, but much less access (if any) to other matters, even if they are related (e.g. "Is <byre> (cowshed) one syllable or two?" or "What is the relative frequency of /k/ vs /k'/ ?). (In the case of some multinational linguistic atlases, such as 'Atlas linguarum europae' or 'Atlas linguistique roman', access to some of these aspects is given via broader linguistic analyses of each map.) Even if one published large amounts of raw data, the effort to find just the relevant parts means that the access is limited.

Modern Information Technology (such as the digital representation of data, data bases and other software applications for storage and retrieval, mapping programmes for displaying data, and the internet for publishing and sharing data) provides some relief from these constraints. They make it possible to store data in a "raw" form, in large quantities, and still access only a select subset of the data as needed; the computer does the tedious work, and the user benefits from the greater access.

We have been involved in three projects using digital dialect data, each leading us to a greater understanding of what is possible. The projects are:

- The Computer Developed Linguistic Atlas of English (Viereck and Ramisch. 1997. See Embleton 1987, 1993; Embleton and Wheeler 1994, 1997a) in which we used a digital data set to perform a Multidimensional Scaling analysis (MDS) of the dialects of English in England.
- Finnish Online Dialect Atlas, in which we digitized a classic hardcopy dialect atlas (Kettunen 1940) and then did MDS on the data. (Embleton and Wheeler 1997b, 2000). Recently, we have been able to go further with this data, as described below.

- Romanian Online Dialect Atlas (Embleton, Uritescu, and Wheeler 2002, 2007a, 2003, 2004, 2006, 2007b, 2008a-e, 2009), in which we digitized an extensive hardcopy atlas (Stan and Uritescu 1996, 2003), for a region of Romania that is critical to Romance linguistics. We then prototyped an application for using the digital data set, in effect creating a digital online dialect atlas.

The Progression of Requirements

In the Romanian Online Dialect Atlas (RODA), our concept of what was needed and possible progressed from a simple idea to a much more elaborate system, as we developed the application.

- At first, the system was intended to be a digital version of the book. It could have been as simple as a set of scanned images of the hard-copy maps. As such, the data set would be relatively easy to prepare, as accurate as the original, and yet relatively easy to share via the Internet.
- However, to apply software to the data, the data set needed to be in a digital form, i.e. the hard copy text needed to be represented as digital text encoded in ASCII (standard keyboard characters) or Unicode (extended keyboard characters) in some form, so that programmes can read it and process it. The effort to digitize a data set can be a heavy, labour-intensive task, and the possibility of transcription errors is always a concern. The development of custom entry tools (which in the case of RODA included a customized "virtual keyboard") can go a long way to mitigate these problems.

  The digitization led us to a major "design choice". The hardcopy text was more elaborate than normal left-to-right text: it had characters that were not available in either ASCII or Unicode; it positioned symbols (typically, accents) in eight different positions around a character, sometimes putting multiple symbols in a position; and it could have one fully accented character positioned above another. In short, the source text was not conventional, so we encoded it in a custom encoding using pairs of ASCII characters. But, that impacted how we entered data, processed data and presented data, and put constraints on how other projects could use our application.
- With a digital data set, we could present the data on maps that we created dynamically (i.e. as the user requested them). Not only could we give the user the data as it was in the hard-copy, but we could allow the user to specify a selection of data: first by selecting which data files to use (e.g. we selected lexical items with a Latin origin in a small study of what happened to the ending on those words in Romanian; see Embleton, Uritescu and Wheeler 2006, 2008a); second, by selecting a particular search string in context (e.g. to display the dentals that did not palatalize before high front vowels; see Embleton, Uritescu and Wheeler 2008a)
- Searching for particular patterns led us to count the occurrences of patterns. Now, instead of displaying a data item beside a location on a map, we could display the number of occurrences of a search pattern, either as a digit or graphically with a bar proportional to the number. Comparisons can be made between two searches using vertical and horizontal bars. The result is a map that displays the comparison with high, immediate visual impact (such visual impact is, of course, one justification for using dialect maps rather than tables of dialect data) .
- From counting, which is really an analytical technique, albeit a simple one, it is an easy step to propose that there be means of incorporating other analytical tools. In RODA, we built-in our multidimensional scaling (MDS) procedure so that we could display a map based on linguistic

similarity rather than on geography. But there is no reason to limit a system to any particular technique; we have accommodated other approaches by exporting the similarity matrix we generated for MDS.

- We have added the ability to represent "interpretations" of the raw data. Thus, in addition to having the collected responses to a prompt, we also have data representing the editor's assessment of the raw data, usually in the form of a map showing the presence or absence of a set of linguistic features (these interpretive maps were also in the hard-copy atlas). The system provides the users the ability to review the raw data, and create (save, display) their own "interpretations".
- From an early stage, there was a requirement that RODA present samples of the audio field tapes. It became relatively easy to use a copy of the online map of the region as an interface, with links by location to selected sound clips. However, the sound clips represented large amounts of data, and we made them be an optional "add-on" so that the base programme and data are not too large.

Development by Prototyping

The RODA application has progressed from a simple presentation of a hard-copy book to a function-rich system of searching, analyzing and displaying digital data sets. The route to achieve this current (but not ultimate) system was one of repeatedly going through a cycle of requirements-design-build-test-and-assess, with each cycle leading to something more useful than what we had before. It is a truism of software design and development that some such process needs to be followed when you do not initially know what you want to end up with (see for example McConnell 1996).

However, the prototyping process also leads to design choices, some of which may not be optimal for the system at later stages, but which the system is committed to. There comes a point when it seems good to stop the prototyping, and start over from the beginning, bearing in mind the lessons that have been learned from the prototyping. Yet such a move can be costly, because it may involve a lot of rework: for example, our custom encoding works well for our Romanian data, but will it work well for other Romanian atlas projects (there are several underway) and will it work for projects generally?

Finnish Online Dialect Atlas

The Finnish Online Dialect Atlas consists of a data set representing a number of hard-copy maps, each of which shows the distribution of one or more dialect features across the region of current-day Finland and some areas beyond. As such, the data is not "raw data" (i.e. not the direct response of an informant) but an analysis of such data showing phonetic and other distinctions, and corresponds to the interpretive maps of the RODA project.

We have been able to recast the Finnish data set in the form of RODA interpretive maps. (Such interpretive maps do not use the custom encoding system, and the conversion is relatively easy.) The result is that we can use a slightly modified version of the RODA software to view the Finnish interpreted data, and to apply our MDS procedure to it. On the other hand, there is no "raw data", and the various functions of RODA that use the raw data (such as viewing, searching, counting) are not available. Part of the adaptation of RODA to Finnish is a matter of turning off functions that are not meaningful.

However, because RODA was not designed to have these functions turned on and off, the actually customization is more ad hoc than we might want. Again, prototyping leads to a solution, but perhaps

not the optimal solution.

Proposal

We have looked at a number of dialect projects with an online presence (see http://ericwheeler.ca/atlaslist for a list of online atlases and related works). They range in function from those that only describe a given project, to those that provide dynamic access to the underlying data. From our experiences developing RODA, and adapting it to the Finnish data, we are led to consider what a modern, and general, online dialect atlas might be (at a minimum).

General data

By a "general" atlas, we mean one that can use a wide range of language data. Existing digital data may need to be re-stated in a particular form; newly collected data may be captured in a prescribed form.

XML is a recognized standard for expressing data that is to be shared between applications. For example, a RODA custom encoded character could be represented in XML as:

<character>

<glyph>a3<accent position=2>b9</accent> </glyph>

<superposition><glyph>a1</glyph> </superposition>

</character>

This describes the character that we represent as a3+2b9+0a1, being one of the forms of the character "a", with an accent, and another form of "a" positioned over it. Our notation is much shorter, and easier to process when our convention is understood, but it is a relatively simple matter to convert from our notation to XML and back, as needed. Internally, a general atlas might use one notation, and externally a more explicit equivalent.

XML conventions, though, do not say what to represent. Certainly, we want the atlas to be able to handle data expressed as simple strings of characters (whether ASCII or Unicode), in various orders (right-to-left, left-to-right, right-to-left embedded in left-to-right, etc.). Such will account for a lot of language data. Going beyond that, we have heritage data that might consist of field notes written in ad hoc notations, and in ad hoc directions, or data from situations that are not naturally sequential, such as gestural communications or body language. There are two approaches, either of which might help with these cases.

- Structure the data as if it were a sequence of complex symbols; allow each symbol to specify special processing such as a different direction, or a special inventory of symbols. Thus, for a stream of gestures, there is a sequence (dictated by time) of complex gestures, and each complex gesture consists of more specific gestures, not necessarily in sequence, but perhaps divided by body part employed and motion type. To the system, it is a sequence of data that can be stored, searched, and presented according to the specific needs of the project.
- Use multimedia, including embedded links to images of the hard-copy data. Thus, for example, a set of idiosyncratic field notes can be coded in plain text (for search purposes) but presented as images (for display purposes).

Online Atlas

The ability to share data with users is greatly facilitated by having an online application, accessed on the Internet, and maintained on one server. Alternatives include making a stand-alone application that can be downloaded to a user (the current form of RODA). The "best" way seems to change with the changing economics of internet technology. An online atlas that is to persist for some time needs to be adaptable to new technologies, but the specifics of that adaptability are hard to pin down. Perhaps it is best to say that no application should make itself wholly dependent on any one technology.

Furthermore, persistence of data is being addressed by archiving sites (e.g. York University Library is offering such an archive, which is where RODA now resides), and a general online atlas needs to have an archive-level home.

Functions

Given that we have a data set, keyed to prompts and to geographic locations, there are several functions that we would expect a general, online dialect atlas to permit a user to do:

- View a prompt to see what it was intended to elicit (e.g. Prompt 1 "What is the opposite of yes?")
- Select a set of data, by prompts. This could be as simple as: "Give me prompts 2, 4 and 7"; or it could be a case of letting the user classify prompts ("prompts 2-45 are phonology"; "prompts 2, 4, 6, 8 are Latin-based" etc.) and combine such requests in some Boolean combination (e.g. "all phonology prompts, excluding Latin-based words, plus prompt 9").
- Display the raw data (from selected prompts) by location, in a map (the classic atlas function) or in a table (for human viewing) or in a file (for further processing, perhaps in XML format).
- Search for the occurrences of a pattern. Given that the data is a string of symbols, the search pattern could be defined as a possible substring of the data (e.g. all occurrences of /ti/), but one quickly wants to go beyond this simple idea:
    - Combinations of strings e.g. /ti/ or /te/
    - Symbols defined by characteristic, e.g. /t + HighFrontVowel/
    - Search strings in context e.g. /ti/ at word-end only.
    - Where symbols are complex, search strings that contain symbols with or without certain features, in contexts with or without certain features, e.g. /t/ with or without accents, in superposition over another symbol, at word end.
- Count the occurrences of a search pattern, and display the result (as a map, table or file) by location.
- Compare counts from more than one search.
- Review the data that created a count at a location, and permit the user to interactively change the count. (e.g. if location 123 has an occurrence of /ti/ that is a known exception to the phenomenon under study, the user edits the result to exclude it.
- Define "similar" for two data items and allow the creation of a similarity matrix comparing locations to locations, based on how similar they are. The definition of "similar" could be as simple as "having the same data" – but in practice, there will be different representations of data that may be considered equivalent (e.g. the order of accents on a symbol may not be significant;

or the quality of a vowel may, by the user's choice, not be significant). Similarity matrices can be created over all the data, or over a selected subset of prompts, or by a subset selected by a search for certain properties in the data. The similarity matrix can be output to a file for further processing, such as multidimensional scaling or factor analysis.

- Create interpretive maps, based on a review of the underlying raw data, or the results of a search, or the results of a count. Interpretive maps could be created automatically, in some cases, from a count by marking a feature as "present" whenever the count at that location is above a given threshold. An automatically generated interpretive map could then be manually revised to reflect a linguist's more sophisticated understanding of the actual linguistic situation.
- Present data as maps, which in turn can be saved in various formats for printing or embedding in electronic documents. There are many aspects of a map that could be customized, such as the inclusion of titles, legends, location labels, data-based labels (i.e. the data value at a given location), colours, size and resolution. Our RODA maps also allow the user to zoom in on a subset of the map, for greater clarity and focus.
- Use a map as an interface to the data. For example, we can access sound files, raw text data, and search pattern matches by location; a click on a map provides the easy way to specify that location.

Data Entry

For projects that include digitization, it is useful to have data entry tools.

- An editor that facilitates the entry of the data in the format that is appropriate to the project. For example, on our Finnish digitization project, we created an editor that sequentially went through the 530 locations, and that presented the data entry options for each of the hard-copy maps; the operator could simply point-and-click to the correct data item or items for the current location (after consulting the hard-copy) and the system moved on to the next location. Data entry was faster and less prone to error than if the data had been entered as text.
- For the Romanian project, we created a virtual keyboard that displayed images of the characters as they were presented in the hard-copy; the keyboard entered our custom encoding into the database. In this way, the data entry operator never had to know what the underlying encoding was, and only dealt with data as it was seen in the hard-copy. The virtual keyboard (an image, on which one pressed keys using the mouse) allowed us to have as many keys as we needed (about 280, though not all were used)
- On several projects, it was necessary to read data in one format, parse it, and rewrite it out in a different format. Alternatively, we have written "meta descriptions" of data that tell programmes about the data as it is, without physically rewriting the data. In any case, it is important for a general atlas to be able to map data in an existing form into the format that the atlas needs.

Summary

A modern general online dialect atlas is more than a traditional book. It is a system of accessing a digital data set, allowing the user to:

- Find the relevant data
- Process the data appropriately

- Present the results as maps or otherwise

RODA offers a prototype of such a system, and RODA can be adapted to other projects and data.

However, it may be appropriate (if the funding is available), to rebuild the system from the ground up so that it more effectively serves a broader community. In either case, we have a model of what a general online dialect atlas can be.

References

Embleton, Sheila. 1987. Multidimensional Scaling as a Dialectometrical Technique, in *Papers from the Eleventh Annual Meeting of the Atlantic Provinces Linguistic Association*, ed. Rose Mary Babitch. Pp. 33-49.

Embleton, Sheila. 1993. Multidimensional Scaling as a Dialectometrical Technique: Outline of a research project, in *Contributions to Quantitative Linguistics, Proceedings of the First Quantitative Linguistics Conference, September 23-27, 1991*, ed. Reinhard Köhler & Burghard Rieger. Dordrecht & Boston: Kluwer. Pp. 267-276.

Embleton, Sheila & Eric Wheeler 1994. *Dialect Project: Technical Report*. York University, Toronto, Department of Languages, Literatures & Linguistics.

Embleton, Sheila & Eric Wheeler. 1997a. Multidimensional Scaling and the SED Data, in *The Computer Developed Linguistic Atlas of England 2*, ed. Wolfgang Viereck & Heinrich Ramisch. Tübingen: Max Niemeyer. pp. 5-11.

Embleton, Sheila & Eric Wheeler. 1997b. Finnish Dialect Atlas for Quantitative Studies, *Journal of Quantitative Linguistics*, volume 4, pp. 99-102.

Embleton, Sheila & Eric Wheeler. 2000. Computerized Dialect Atlas of Finnish: Dealing with Ambiguity, *Journal of Quantitative Linguistics*, volume 7, pp. 227-231.

Embleton, Sheila, Dorin Uritescu & Eric Wheeler. 2002, 2007a. *Online Romanian Dialect Atlas*. http://vpacademic.yorku.ca/romanian (now at http://pi.library.yorku.ca/dspace/ under the "dialectology" community, "RODA" collection)

Embleton, Sheila, Dorin Uritescu & Eric Wheeler. 2003. "Romanian Online Dialect Atlas". International Colloquium of IQLA – International Quantitative Linguistics Association, University of Georgia, Athens, Georgia, May, 2003

Embleton, Sheila, Dorin Uritescu & Eric Wheeler. 2004. Romanian Online Dialect Atlas. An exploration into the management of high volumes of complex knowledge in the social sciences and humanities. *Journal of Quantitative Linguistics*. 11.3. 183-192. December 2004.

Embleton, Sheila, Dorin Uritescu & Eric Wheeler. 2006. Seeing Words Change using the Romanian Online Dialect Atlas. Presentation to International Linguistics Association. Annual Meeting. Toronto. April 2006.

Embleton, Sheila, Dorin Uritescu & Eric Wheeler. 2007. Romanian Online Dialect Atlas: Data Capture and Presentation. *Exact Methods in the Study of Language and Text*. (Quantitative Linguistics, 62.) G. Altmann Festschrift. Peter Grzybek, Reinhard Köhler (eds). Berlin and New York: Mouton de Gruyter. Pp 87-96.

Embleton, Sheila, Dorin Uritescu & Eric Wheeler. 2008a. *Digitalized Dialect Studies: North-Western Romanian.* Bucharest: Romanian Academy Press.

Embleton, Sheila, Dorin Uritescu & Eric Wheeler. 2008b. Defining User Access to the Romanian Online Dialect Atlas, *Dialectologia et Geolinguistica*, 16, pp. 27-33.

Embleton, Sheila, Dorin Uritescu and Eric Wheeler. 2008c. Identifying Dialect Regions: Specific features vs. overall measures using the Romanian Online Dialect Atlas and Multidimensional Scaling. Leeds, UK: Methods XIII Conference. August 2008. (to be published)

Embleton, Sheila, Dorin Uritescu and Eric Wheeler 2008d. "Data Management and Linguistic Analysis: MDS applied to RODA". Presented to the Trier Symposium on Quantitative Linguistics, Trier, Germany, December 2007. To be published in Journal of Quantitative Linguistics.

Embleton, Sheila, Dorin Uritescu and Eric Wheeler. 2008e. "Lessons from Digitizing a Dialect Atlas". International Conference on Linguistic, Literary and Ethnolinguistic Communication in the New European Context, Iasi, Académie roumaine, Institut de Philologie roumaine 'Al. Philippide', September 2008.

Embleton, Sheila, Dorin Uritescu and Eric Wheeler. 2009. "The Stability of Multidimensional Scaling over Large Data Sets: Evidence from the Digitized Atlas of Finnish", in *Mélanges en l'honneur de Juhani Härmä*, Mémoires de la Société Néophilologique de Helsinki, ed. Eva Havu, Mervi Helkkula and Ulla Tuomarla, May 2009, pages 207-214.

Kettunen, Lauri. 1940. Suomen murrekartasto [The dialect atlas of Finland]. Helsinki: Suomalaisen kirjallisuuden seura.

McCarthy, Mary. 1942. The Man in the Brooks Brothers Shirt. The Company She Keeps. Harcourt, Brace (1965 Penguin Books).

McConnell, Steve. 1996. Rapid Development. Redmond Washington: Microsoft Press.

Stan, Ionel, & Dorin Uritescu. 1996. *Noul Atlas lingvistic român. Crişana. Vol. I*. Bucharest: Academic Press.

Stan, Ionel, & Dorin Uritescu. 2003. *Noul Atlas lingvistic român. Crişana. Vol. II*. Bucharest : Academic Press.

Viereck, Wolfgang, and Heinrich Ramisch, 1997 ed. *The Computer Developed Linguistic Atlas of England 2*. Tübingen: Max Niemeyer.