Lessons from Digitizing a Linguistic Atlas

Sheila EMBLETON, Dorin URITESCU, Eric S. WHEELER

It is good to convert hard-copy data into a digital form because of the many useful ways the data can be processed using modern information technology. Through the internet, it also becomes the best way of making the data available to the whole community of linguists, easily and in a useful manner. However, the task of converting the data can be time-consuming, expensive, and error-prone unless it is done thoughtfully. Furthermore, the conversion task can expose ambiguities and surprises in the source data, requiring serious editorial decisions.

Part I. An Introduction to RODA

The Romanian Online Dialect Atlas (RODA) is a digitized version of a multi-volume hard-copy dialect atlas (Stan and Uritescu 1996, 2003) of the Crişana region in north-west Romania. It records responses to over 400 indirect questions at 120 locations, and shows phonetic, lexical and morphological patterns in the region.

RODA consists of a set of digitized data files in a simple "flat file" format, and an application (written in the Java programming language, so that it runs across a wide range of computer platforms) to provide sophisticated access to the data.

The RODA data and programme is available for downloading over the internet from http://vpacademic.yorku.ca/romanian/ (Embleton, Uritescu and Wheeler 2007b) and in future will also be posted at an archive site run by the York University library.

Notations and Fonts

Because the notation used in the hard-copy atlas is capturing subtle phonetic variations, it employs an extensive set of 115 basic characters and over 60 accents and other ancillary symbols (see figure 1).

The accents (primarily the last row in figure 1) can be in any of 8 positions around the base letter. What is more, characters are not only arranged left-to-right, but sometimes one character-with-accents is positioned above another.

To deal with this situation, RODA uses a two-character set of ASCII codes: The first character is alphabetic and indicates the column in the font image; the second character is numeric (0 to 9) and indicates the row. Thus, "a0" encodes a simple "a" and "a1" encodes an "a" with a circle above; "a2" encodes an "a" with a circle above and a cedilla below. Special codes "+1" to "+4" and "+6" to "+9"introduce an accent in one of the 8 available positions from top-left to bottom-right, and "+0" introduces a sequence for a whole accented-character positioned above this character (we call this "superposed").

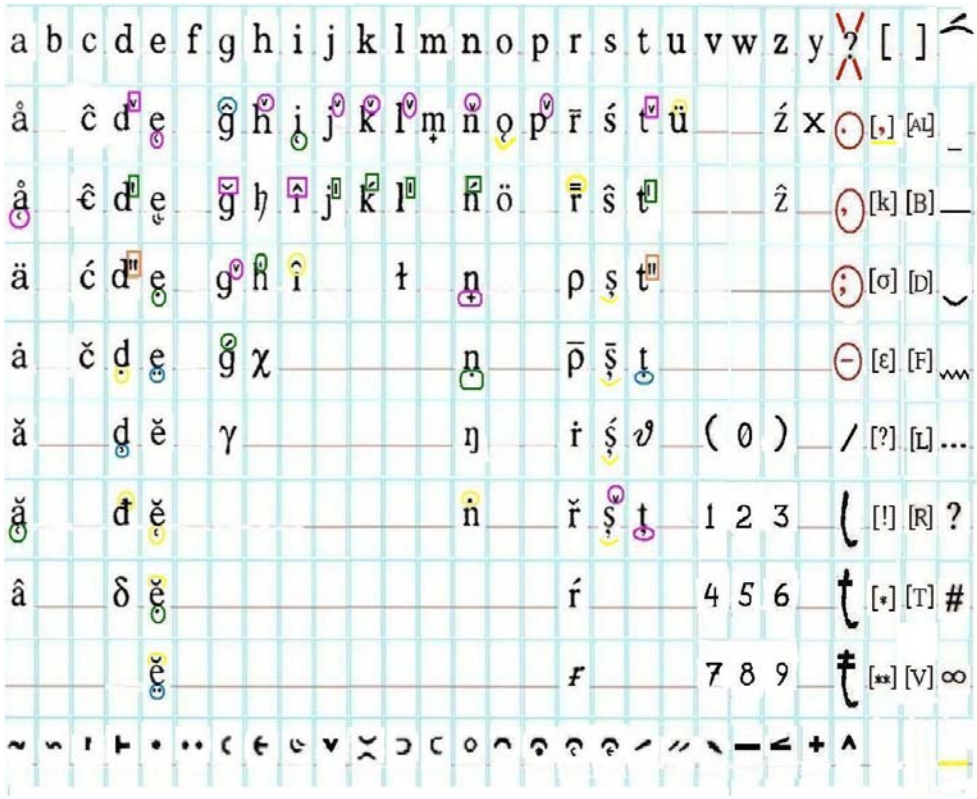

Figure 1. The font image for RODA

Thus in figure 2, a sequence from the hardcopy book becomes "a5+2c9s2c0+2j9+0t0+3c9e5%9" where "%9" is the final space.

Further, there are symbols (here, "[B]") that encode fieldworker notations such as "used by older people" or "hesitation" and indicate sociolinguistic and style variation or more subtle aspects of internal derivation (see for instance Uritescu 2006).

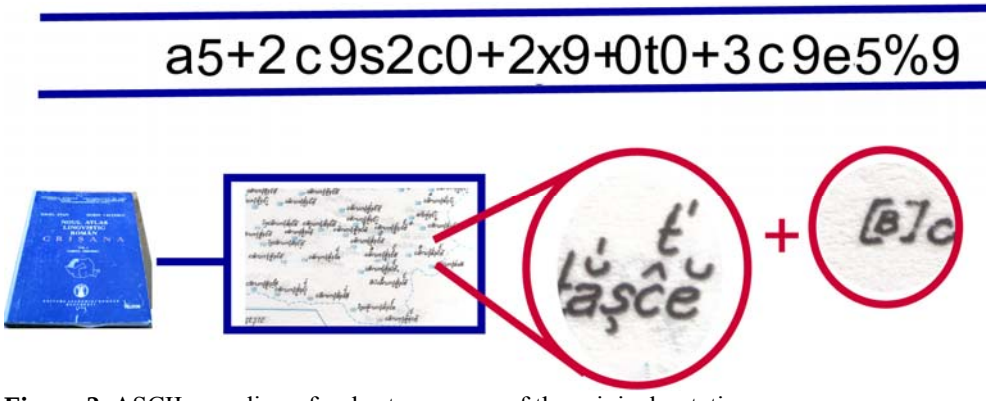

Figure 2. ASCII encoding of a short sequence of the original notation

It must be obvious that a notation as rich as this one cannot be handled by simply expanding the standard ASCII notation to the larger Unicode font set. It is necessary to encode more, and that has forced us to develop our own way of presenting this notation.

We use a .jpg image (figure 1) which our application then cuts up into individual symbols, and arranges those symbols in the appropriate left-to-right, accent-position 1 to 9, and standard-or-superposed placements.

There is a disadvantage to this approach because of the additional custom programming needed to see the digitized data (an issue we addressed in Embleton, Uritescu and Wheeler 2007a). On the other hand, there are also advantages such as:

- The capacity to make *any* symbol needed;
- The potential to annotate symbols (we have put coloured circles and squares around subtle features of the notation so that, for example, the second "a" (a1) and third "a" (a2) are more noticeably distinct, as are the third and seventh (a6). See figure 1);
- The ability to arrange symbols in non-standard positions and sizes.

In this way, we have dealt with a field notation that was developed long before the digitalization project was thought of.

Functions

RODA allows the user to have sophisticated access to an extensive data set. Because of the power of the digital technology, this access is far greater than what one would get out of a printed publication, even if the publication had an extensive index and table of contents.

With RODA, one can:

- Select the files to include, each file representing one of the elicitation forms. Thus, in one of our studies (Embleton, Uritescu and Wheeler 2008a), we looked at the evolution of the endings of Latin "oculum" (eye) and "canto" (I sing). In our region, the reflex can be either a syllabic or non-syllabic /u/, but it was helpful that we could select files representing the appropriate lexical items, Latin or non-Latin, and with the desired phonetic environment.
- View the data. One can select a data point by location and file, and see the relevant item in its presentation form. The data can be accessed from a map or a list, and the list can show items one by one, in detail, or several items at a time in a more compact form. The power of a digital format is that the same data can be viewed in more than one way, to meet different needs.
- Interpret the data. It is possible, while examining the data, to make a map with symbols on any location, to represent your own interpretation of what the data shows. Such maps can be saved for further work later (work-in-progress) or for inclusion in external applications such as web pages or word-processor documents.
- Search and count the data. For example, we looked for *all* the occurrences of front vowels coming after non-palatalized dentals (Embleton, Uritescu and Wheeler 2008a) to demonstrate that this phenomenon was more common than had been previously supposed. To do this, we had to search for more than one vowel and compare the results. The results were expressed as the number of occurrences at

each location, and were displayed as horizontal and vertical bars on a map (figure 3). As such, the count showed us that the phenomenon was more common in the Oaş area (northernmost part of our region) where it was expected, but that it also occurred widely throughout our area to some extent, which was not commonly expected.

- Review and edit searches. Whenever an automatic search is done over a large set of data, it is possible to uncover examples that were not originally anticipated. For example, in our search for word final /u/, some of the examples represented forms with the definite article which was not what we were searching for. By reviewing and manually revising the search result, we obtained a more accurate list of examples for the question we were investigating. This ability both to search-and-count automatically and then to manually review and revise is essential if the user is to get trustworthy results; we do this review naturally when we deal with small data sets, but for very large data sets, the function must be built-in to the application.
- Hear the data. RODA offers the ability to access selected sound clips by location from a map. The large size of sound data makes it necessary (at this point) for us to limit what is generally available, but the interface can access any available files. Hearing the data is yet another way for a researcher to review and check the data available in the atlas.
- Process the data with advanced methods. We have built-in a method of doing multidimensional scaling (MDS. See Wheeler 2005; Embleton, Uritescu and Wheeler 2008c, forthcoming) in which we get a map of the 120 locations based on linguistic distance rather than geographic distance. This analysis has led to some interesting observations about the nature of the dialect situation in the north-west, and about the definition of "dialect area" in general (Embleton, Uritescu and Wheeler 2008b). But, MDS is not the only possibility. With digitized data, it is conceivable that one could apply various techniques. RODA, for example, can export a (linguistic) distance matrix, and such a matrix can be used in clustering techniques and other approaches (see figure 4).

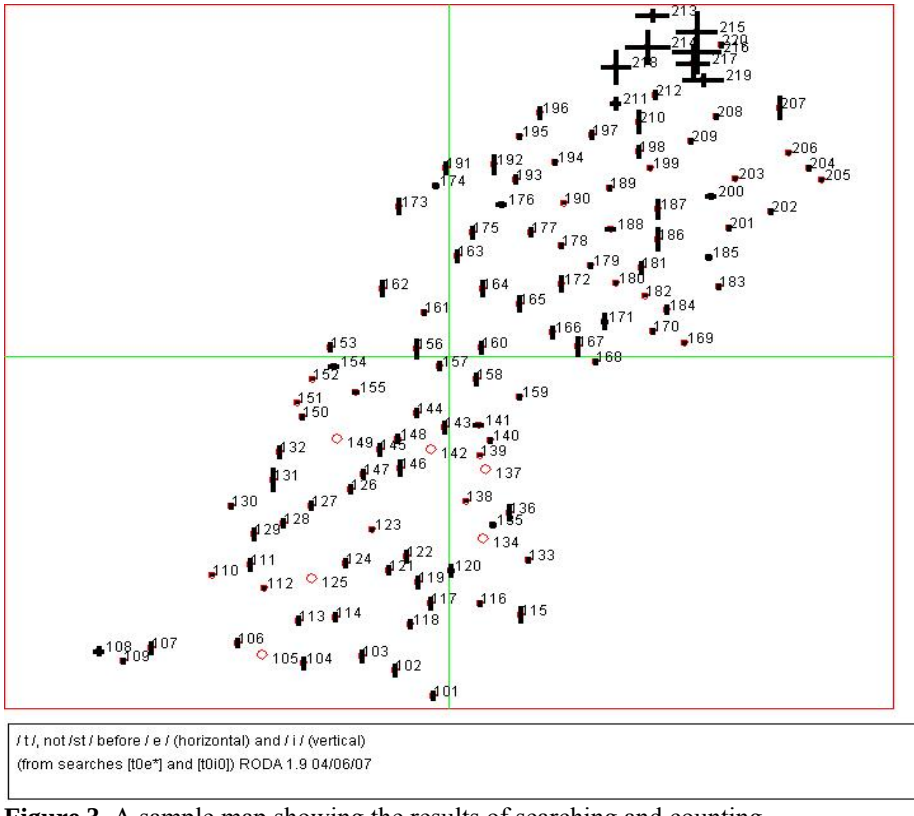

Figure 3. A sample map showing the results of searching and counting

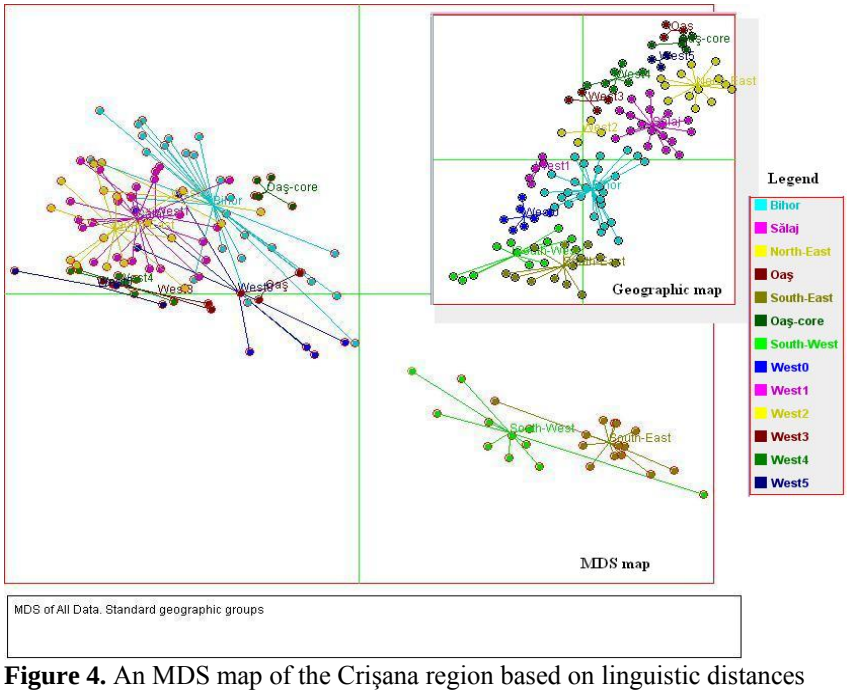

Figure 4. An MDS map of the Crişana region based on linguistic distances

Part II. Lessons Learned

At a practical level, we have learned from our RODA project (and its predecessor, where we digitized a Finnish dialect atlas based on Kettunen 1940; see Embleton and Wheeler 1997, 2000) many useful tips about the digitization of data and the online presentation of digital data, some of which may apply to future projects. These include the use of:

• A customized data entry screen, that can lessen the work required to create the original digitized data, and reduce the opportunity to introduce errors. In the case of RODA, we provided a virtual keyboard with the project-specific characters and positioning modes.

Figure 5. Virtual keyboard created for RODA data entry and application use

• "Flat files" to store data, rather than committing to a particular data base format which over time may become obsolete. From the flat file, one can create (or later recreate) the necessary forms for any particular software programme.

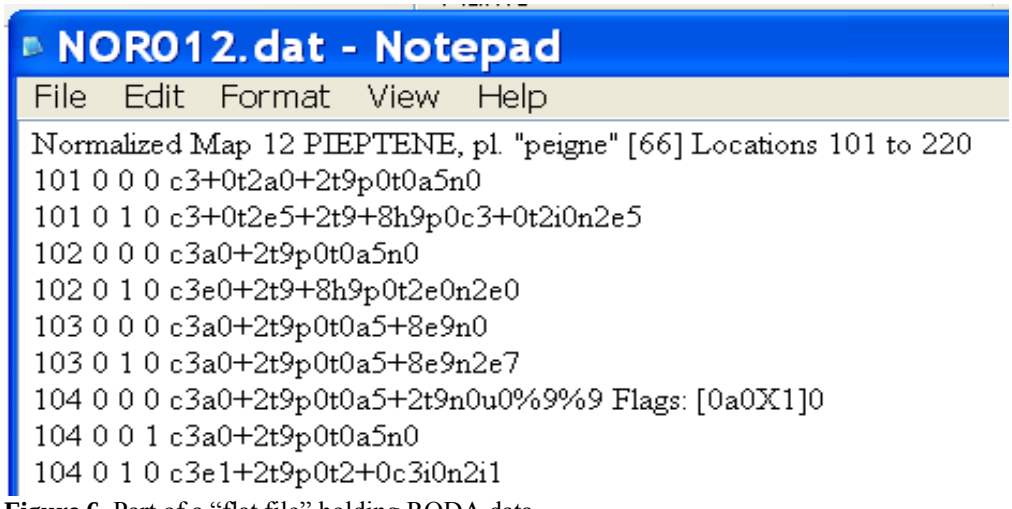

Figure 6. Part of a "flat file" holding RODA data

• A flexible development process. At any point, it may be desirable to change earlier design decisions. The process should anticipate this sort of change, and frequently test the evolving application against typical user expectations. Also, there will be unforeseen implications to any design decision. For example, in RODA, we expected to use Unicode symbols (because Unicode has a large and rich set of fonts and characters) but we had to change our approach when we could not find all the symbols we needed, when some symbols would not print with the available fonts, and when we realized the data had more than just left-to-right ordering.

• A modular application. If the application is built in small, self-contained units, it will be less complicated to understand, less error-prone, easier to maintain, and much easier to extend by adding more units. Application maintenance (i.e., fixing errors, changing the programme to fit a changing environment, and making obvious improvements to its performance and usability) is an extensive and on-going part of any software development project, and must not be under-estimated.

• A good editing team. The editors will not only have to check for the quality of the digital data (and allow for correction and re-checking) but also make "editorial" decisions about how the data will be interpreted digitally. For example, in our Finnish project, we had locations that were on the border of an area represented by a feature: do we assign the feature to that location or not? This is not a decision for a data entry person to make "on the fly", but rather, for someone who can make a decision that is appropriate and consistent across the whole project. In the case of RODA, we could consult original field notes to help resolve ambiguities. In the Finnish case, we have ambiguities that cannot be resolved easily, and that need to be flagged as such (an extension of our digital notation that we had not originally anticipated).

But, in addition to the practical lessons discovered along the way, we have gained an overall vision of what it means to make a digital dialect atlas.

First, a digital atlas is not just an electronic book of maps (see Embleton, Uritescu and Wheeler 2006). Mapping is not the challenge that we expected it to be. With simple geographic coordinates for each location, we created effective base maps by just positioning dots on a page (see figure 3, for example). To these dots, we could add labels (to name the locations), symbols (to put the location in a dialect area) or horizontal and vertical bars (to show quantification).

Rather, a digital atlas is a rich repository of data, and the challenge is to find ways of getting at the data it holds. Do you want to: know about phonetic variation across the region? or know the historical evolution of sound changes and whether or not they happened independently? or refute a claim about lexicalization? or look for morphological or morphophonemic patterns? The answers are in the data set and the computer is powerful enough to search large data sets quickly and often, if we can only provide the interface to allow you to ask for the right searches. With a large data set, you may get the relatively rare, but important examples or counter-examples you need. With quantitative measures, you get a strong indication of how much evidence there is on any given claim. And when you search the whole data base, you have the confidence that you are using the best evidence available so far.

But the challenge for us was to provide means of getting at that data. As our concepts developed, our searches became more and more sophisticated: they had to

find strings of characters, with or without accents, with or without superposed characters, with or without field-worker notations, in contexts such as word final and word initial. For morphological and lexical variation, we chose to arrange the data in such a way as to provide direct access to morphological forms or to lexical or other variants. For this, we arranged the data in different fields (for instance, field 1 for singular, field 2 for plural), each of them having potential variants, and layers of fields for lexical variants. While we cannot claim to have an ultimate solution, it is the case that access to the data was much more of a concern for us than mapping the data.

Second, a digital dialect atlas is not simply pages of isoglosses, but rather it is a multitude of ways of looking at variation and correlating it with geography. We offer maps that symbolically represent a user defined interpretation of the data (a RODA "interpretation"), or a quantitative assessment of the data (a RODA "search and count"), or a more advance processing of the data (an MDS map, or a distance matrix). The results can be seen as a map, or a list of examples, or a data set that can be sent to another application. In addition, through the digitalization of the manually created interpretive maps form the hard-copy atlas, an operation that is now under way at the Institute of Linguistics and Literary History in Cluj-Napoca, researchers will be able to apply MDS to sets of isoglosses and compare the results to the statistical analyses of the raw data. We will thus be able to see the role of discrete features and their overlap in defining dialect areas, and to study the relations between these features and the linguistic continuum put forward by statistical analyses of the basic data. As a result, it becomes possible to use the digital atlas in ways that go well beyond what anyone would have done with a hard-copy book.

Third, a digital dialect atlas is not just the tool of Dialectology, but opens up possibilities for a broad spectrum of disciplines. Ethnographers looking at the dating of human population movements may search for evidence in the digital data set (use of data that the creator of a hard-copy atlas would not have anticipated, and may not have made available). Likewise, the historian seeking signs of past events, and the sociologist or sociolinguist wanting to test theories evidenced by language can all benefit from an accessible data set.

For us, the most important lesson we learned is this idea of "accessible data", in which information technology provides the power to manipulate large amounts of data, the flexible interface allows us to seek patterns in the data we would not have looked for otherwise, and the dynamic presentation of the data (from customized maps to exportable data files) makes the results immediate and useful.

References

Embleton and Wheeler 1997: Sheila Embleton and Eric Wheeler. "Finnish Dialect Atlas for Quantitative Studies", *Journal of Quantitative Linguistics*, volume 4, pp. 99–102.

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

Embleton, Uritescu and Wheeler 2004: Sheila Embleton, Dorin Uritescu and Eric Wheeler. "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, Uritescu and Wheeler 2006: Sheila Embleton, Dorin Uritescu and Eric Wheeler. "Defining User Access to the Romanian Online Dialect Atlas". Presentation to the 5th Congress of Société Internationale de Dialectologie et Géolinguistique (International Society for Dialectology and Geolinguistics). Braga, Portugal. August 2006. to be published in 2008 in Dialectologia et Geolinguistica, vol.16.

Embleton, Uritescu and Wheeler 2007a: Sheila Embleton, Dorin Uritescu and Eric Wheeler. "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 Koehler ed. Berlin and New York: Mouton de Gruyter, pp. 87–96.

Embleton, Uritescu and Wheeler 2007b: Sheila Embleton,, Dorin Uritescu & Eric Wheeler. Online Romanian Dialect Atlas. http://vpacademic.yorku.ca/romanian.

Embleton, Uritescu and Wheeler 2008a: Sheila Embleton, Dorin Uritescu and Eric Wheeler. (forthcoming). Digitalized Dialect Studies: North-Western Romanian. Bucharest: Romanian Academy.

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

Embleton, Uritescu and Wheeler 2008c: Sheila Embleton, Dorin Uritescu and Eric Wheeler. "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 Quantitativel Linguistics.

Embleton, Uritescu and Wheeler *forthcoming*: Sheila Embleton, Dorin Uritescu and Eric Wheeler. forthcoming. The Stability of Multidimensional Scaling with Large Data Sets.

Kettunen 1940: Lauri Kettunen. Suomen murrekartasto [The Dialect Atlas of Finland]. Helsinki: Suomalaisen kirjallisuuden seura.

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

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

Uritescu 2006: Dorin Uritescu. "'Real', though not 'needed'? On a phonotactic constraint as reflected in *NALR-Crişana*". In Aronson, I. Howard, *et al.* (eds.), *The Bill Question*, Bloomington: Slavica, pp. 207–214.

Wheeler 2005: Eric S. Wheeler. "Multidimensional Scaling for Linguistics". in Reinhard Koehler, Gabriel Altmann and Rajmund G. Piotrowski. editors. Quantitative Linguistics. An International Handbook. Berlin: Walter de Gruyter. pp. 548–553.

Lessons from Digitizing a Linguistic Atlas

It is good to convert hard-copy data into a digital form because of the many useful ways the data can be processed using modern information technology. Through the internet, it also becomes the best way of making the data available to the whole community of linguists, easily and in a useful manner. However, the task of converting the data can be time-consuming, expensive, and error-prone unless it is done thoughtfully. Furthermore, the conversion task can expose ambiguities and surprises in the source data, requiring serious editorial decisions.

Toronto, Canada