DATA CAPTURE AND PRESENTATION IN THE ROMANIAN ONLINE DIALECT ATLAS

Sheila Embleton, Dorin Uritescu and Eric Wheeler
York University, Canada

RODA, the Romanian Online Dialect Atlas (Embleton, Uritescu, and Wheeler 2002, 2004, 2006, in press), is a two-stage project involving (1) the transfer of data from a hard copy atlas of the Crisana dialect of Romanian (Stan and Uritescu 1996, 2003) to an online system for general availability, and (2) the application of innovative statistical methods to the data.

Romanian, as the prime exemplar of the eastern Romance languages, has had scholarly attention, including the detailed work of Stan and Uritescu (1996, 2003) and Uritescu (1984a, 1984b) on the dialects of the Crisana region in north-west Romania. In digitizing this data to make it more broadly accessible, and in successfully digitizing a hardcopy dialect atlas of Finnish (Embleton and Wheeler 1997b, 2000), we encountered several situations worth highlighting to others who may be considering parallel projects.

1. STANDARDIZED TOOLS

The contention of Bird and Simons (2003) that such capture and presentation should be done with standard tools, such as commercial databases and Unicode encoding schemes, is laudable but not entirely practical. Standardized tools change over the years: the standard database in 1995 would not have been our choice in 2005. Hence the promise of consistency and portability is somewhat muted. We chose instead to keep data in "flat" files (with simple layouts) that can be readily imported into a tool of choice, today or in the future.

2. NON-STANDARD DATA

Furthermore, the primary data were in a field notation that used many non-standard symbols, and arranged them in a non-linear order. See Figure 1 for examples of (1) diacritics on symbols, (2) symbols placed above symbols, and (3) symbols not in the standard ASCII or readily-available Unicode fonts.

We elected to capture the data using internally pairs of standard characters (a0, a1, a2 for each variant of "a" etc., other pairs to indicate shifts of position around the base linear order of characters). The result is a notation that can be searched and processed explicitly for what the notation expresses. However, to see

©LINGUISTICA *atlantica* Nos. 27–28, 2006–2007 37–39

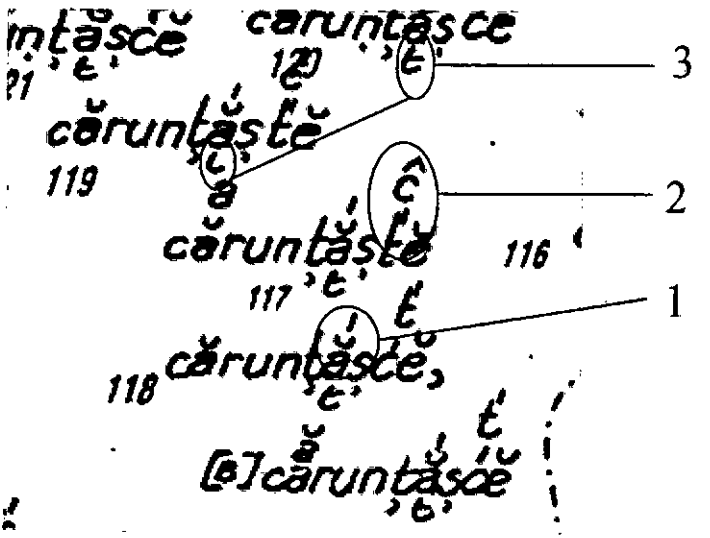

FIGURE 1

Example of field notation source for the Romanian Online Dialect Atlas

the notation in a "readable" form, we need to transcribe the two-character data as either single characters, losing some of the fine distinctions, or as images of the complete character, which requires a custom-made tool. Both the online atlas and our internal conversion tools use the second option.

3. PRESENTATION TECHNIQUES

Hardcopy dialect atlases contain a lot of data, but sometimes the selection and comparison of data (for a particular hypothesis) is difficult because the data are distributed over multiple maps or they are organized in a different way than interests the reader. We are developing the online atlas with some innovative presentation techniques such as interactive data points on each map, dynamically created maps for multi-map comparisons, and user-selected data sets.

4. APPLICATIONS

An initial application of the multidimensional scaling (MDS) method (Embleton and Wheeler 1997a, 1997b, 2000) on available data shows how the data can be used by methods external to the atlas itself, and gives some indication of the utility of MDS for dialectology.

Since then, we have tried using a prototype of the atlas to answer linguistically relevant questions about Romanian and Romance languages, for example, the retention from Latin of word-final "u", lost in standard Romanian, and the treatment of final "e" and schwa in some areas of Crisana (Embleton, Uritescu and Wheeler

2006). We found that even our prototype tools made the data much more accessible and invited us to explore many more possibilities than we might have looked at before.

REFERENCES

Bird, S. and G. Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79:557–582.

Embleton, S. and E. Wheeler. 1997a. Multidimensional scaling and the SED data. In *The computer developed Linguistic Atlas of England 2*, ed. W. Viereck and H. Ramisch, 5–11. Tübingen: Max Niemeyer.

——. 1997b. Finnish Dialect Atlas for quantitative studies. *Journal of Quantitative Linguistics* 4:99–102.

——. 2000. Computerized Dialect Atlas of Finnish: Dealing with ambiguity. *Journal of Quantitative Linguistics* 7:227–231.

Embleton, S., D. Uritescu and E. Wheeler. 2002. Online Romanian Dialect Atlas. `vpacademic.yorku.ca/romanian`

——. 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* 11:183–192.

——. 2006. Seeing words change using the Romanian Online Dialect Atlas. Presentation to the International Linguistics Association Annual Meeting. Toronto, April 2006.

——. In press. Romanian Online Dialect Atlas: Data capture and presentation. In *Festschrift for Gabriel Altmann*, ed. P. Grzybek and R. Koehler. Berlin: Mouton de Gruyter.

Rusu, V. 1984. *Tratat de dialectologie românească*. Craiova: Scrisul Românesc.

Stan, I. and D. Uritescu. 1996. *Noul Atlas lingvistic român. Crisana*. Vol. I. Bucharest: Academic Press.

——. 2003. *Noul Atlas lingvistic român. Crisana*. Vol. II. Bucharest: Academic Press.

Uritescu, D. 1984a. Subdialectul crisean. In Rusu, pp. 284–320.

——. 1984b. Graiul din Tara Oasului. In Rusu, pp. 390–399.