DORIN URITIESCU AND THE ROMANIAN ONLINE DIALECT ATLAS: CONTRIBUTIONS TO DIALECTOLOGY

Dorin Uritescu’s career-long study of Romance languages, Romanian, and in particular, the dialect region of Crișana in north-west Romania, and his colleagues’ (the authors’) early interest in dialectology and the use of information technology has produced not only the Romanian Online Dialect Atlas (RODA) but also many varied studies in the development of Romanian, the nature of phonology, the relationship of geography to dialect, techniques for the display of dialect data and relationships, and the re-purposing of RODA to work with other languages and other modes of linguistic analysis. Over several decades, the collaboration has been very fruitful. In this short note, we would like to show how the interaction of our diverse interests and backgrounds has led to this result, and present one new study that uses the RODA data.

FIRST STEPS

The relationships among a set of language varieties can be and often are portrayed using binary trees. For example, consider the Wikipedia entry for the Indo-European language family (https://commons.wikimedia.org/wiki/File:IndoEuropeanTree.svg) or a recent description of Semitic languages published in Academia by Na’ama Pat-El (Pat-El 2022). But such trees are “fragile” because 1) small changes in the underlying data set (from errors, or ambiguous entries, or new data) can lead to large changes in the tree, and 2) even different clustering methods on the same data can produce quite different trees.

Instead, Embleton (1987) proposed using a well-known statistical procedure, multidimensional scaling (MDS), to produce a visual map in which the degree of similarity in language varieties would (generally) be represented by the spacing of the points on the map. MDS maps are not fragile in the way that trees are, because small changes in the data lead only to small changes on the map (Embleton, Uritescu and Wheeler 2009).

1 The authors honour their late colleague, Dorin Uritescu (1947–2020) who was so central to our shared research work. We also acknowledge the support over many years of the Social Sciences and Humanities Research Council (Canada) and the Romanian Academy.

2 For further explanation of MDS, see Wheeler (2005) or Gabmap (nd). Gabmap is an online tool for doing dialectometry, and includes good discussions on MDS and other techniques.

3 Embleton, Uritescu and Wheeler will be cited hereafter as EUW.
We tested this idea on sets of English data (Embleton and Wheeler 1997a) and Finnish data (Embleton and Wheeler 1997b). The initial result was a map with a “cloud” of points, but points that clearly spread in a way that reflected geography: northern English counties were at one end of the cloud, and southern counties at the other end – even though the map was created from strictly dialect data, without reference to the underlying geography. But the map was hard to interpret without some appropriate labelling. An MDS map with a few data points could be interpreted directly, but in our cases, where there were hundreds of data points, more was needed. To this end, we tried labelling different geographic subgroups with distinct colours, adding label names to individual points, and drawing a polygon around a given subsets of points. But the most effective method was the idea of a “starburst”: each point in a subgroup was connected by a line to a centre point for the group. The result was a highly-visible configuration of points; it permitted subgroups to overlap when that was necessary, and it allowed for simpler labelling of a group rather than labelling each individual point. See figure (Illustration 1) from Embleton and Wheeler (1997a).

Illustration 1 An MDS picture of 40 regions from the Survey of English Dialects (SED), using starbursts, colour and group labels to visualize the MDS results

The lesson, which we carried through to RODA, was that high volumes of data in an MDS map needed effective visual presentation, but also, MDS lent itself to visual interpretation.
RODA

Dorin Uritescu, with his colleagues and students, over many years had done extensive and detailed field work on the dialects of Romanian in the Crișana region of north-west Romania. When we met him, he had the first few volumes of a hard-copy atlas that would, in time, be a five-volume set with detailed maps (initially hand-drawn and hand-labelled) covering a large range of dialect features (Stan & Uritescu et al. 1996 seq.). The challenge for us was to digitize the existing hard-copy data, and then to use modern information technology to analyze the data. As the project unfolded, we found that there was a lot more that could be done, and the outcome of this long interaction between what could be imagined and what we could actually do was (in part) the Romanian Online Dialect Atlas (RODA).

First, to make it feasible to convert hard-copy to digital data, we created special data entry software. The form of the source data (field notes, and the first volumes of the published atlas) was not simply ASCII data, but also included characters with accents in various positions around (above, below, before, after) a given character, and sometimes the superposition of one such accented character above another. To capture this, we created a virtual keyboard that not only had the base alphabet, but also various shift keys to allow for the non-standard positioning of symbols. All of this had to be represented in a custom two-character notation. Non-standard notation means that it was not immediately usable by standard software, but it did allow us to do such things as search the data for strings with particular accents or specific characters in superimposed positions. We gained control over the full range of the data, in the form it was entered in the field notes (EUW 2007b, 2008a, 2008b).

In dealing with the complicated representation of the data, we were led quite naturally to develop search techniques that gave us unprecedented access to the data: we could find not only words with a given spelling, but also all the strings that had vowels with a certain degree of raising (an accent) or that distinguished between syllabic and non-syllabic final schwa. Such searches were used to make a study that had deep theoretical implications in phonology (EUW 2006, 2013).

APPLICATIONS

With RODA, it was possible to frame research questions and seek answers that could be supported with all the available data (EUW 2004, 2007a, 2007b). Prior to this, it might be possible to find some anecdotal support for a theoretical claim, or with extensive manual work, find a larger set of supporting data. But with RODA, we could look for all the data that had a bearing on the question. Sometimes, that could put a plausible claim into doubt (see EUW 2021 for more discussion), but when used with good judgment, it became a powerful way to investigate theoretical questions.

For example (EUW 2006), some Latin nouns and verbs had a reflex ending in word final -"u that is now largely lost in Romance and completely lost in standard
Romanian and in most Romanian dialects, but has been preserved, in part, in the dialects of the Crișana region of Romania. The distribution of final syllabic -u, or its reduction to a non-syllabic form, or its loss altogether, demonstrated separate phonological changes in play.

Combined with a second study on the raising and loss of syllabicity of two other word final vowels (-e and schwa), this again showed separate processes (where, on anecdotal evidence, one might posit a single weakening process). Together, these results led Uritescu to advance theoretical claims about separate “weakening” processes in phonology (cf. Uritescu 1986, Uritescu and Darden 2005, EUW 2013).

Our research studies also led us to consider how dialect differences correlate with geographic distance. The RODA data was extremely useful here. In Crișana, we could measure direct geographic distances among the 120 locations where Uritescu had done field work, and related distances such as travel time and travel distance. We found a high correlation between the two, though not so high that it precluded other factors as well (EUW 2015, 2016a).

In a different direction, we found that the digitizing of the field notes permitted us to annotate the data with “meta” tags (i.e. tags indicating syntactic features about the texts, or sociological features such as age and gender of the respondent) that could also help frame a more refined search (EUW 2016b). Furthermore, we were able to generate the published form of the later volumes of the hard-copy Atlas, directly from the digitized data, thus saving the work (and chance of error) that comes in manually creating the publishable text.

So, RODA and the extensive field work that went into creating it has had useful applications. Here, we want to illustrate how the data has been applied in a new, and somewhat different application.

**CHARACTER PROFILES FOR DIALECTS**

In earlier work (Wheeler et al. 2020, Wheeler and Embleton 2023), we looked at a quantificational study of texts. Using a simple measure, called a Character Profile4, we were able to place a text somewhere on a two-dimensional map. Texts from one language would all cluster close together on the map, but different languages positioned far apart in general. See Illustration 2 for a sample, based on English novels by Charles Dickens, and translations. The data comes from Project Gutenberg.

However, when we looked at texts from different eras, all in one language (English), we found that the clusters strongly overlapped, although there was a noticeable spread for texts that were separate in time or even genre. See Illustration 3.

We tried to explain these patterns with a theory of “gravity”. On the one hand, we expect the texts to position anywhere on the picture, and different texts could be far apart; call this Spread. On the other hand, if the texts belong to a

---

4 For a Character Profile, we count the relative frequency of each letter in a given alphabet. These frequencies become the coordinates in a high dimensional space, which we then project down to two dimensions, as we did with our RODA MDS pictures.
Illustration 2 A picture of several similar texts in Dutch, German, English, French and Finnish. In each language, the texts cluster close together, but the different language clusters are generally far apart.

Illustration 3 Various texts from different eras of English literature, showing both overlap and clustering.
community that understands and communicates with one another, the authors must use the shared conventions of the community – including the frequency of letters. This makes texts from a given community very similar; call this Gravity. For texts from a given language, Gravity pulls them together on the picture, but it hardly affects texts from different languages so the languages cluster apart.

Clearly, as time and other factors (such as text genre: prose, poetry, plays) change the conventions of a community, we expect the effects of Gravity to be lessened by Spread, so that texts from different eras will have some overlap, but less so as time goes on.

The same logic should apply to different dialects: the more remote the dialects are, the farther apart they should be in Character Profile space. We chose to test this by using the RODA data, as digitized from Uritescu’s field notes.

Each location was treated as an author, and each location belonged to a previously identified dialect area. The data gathered at a location became the text, and the alphabet was the rather extensive alphabet used to digitize the field notes. Each letter in this alphabet was represented by two ASCII characters, based on the RODA virtual keyboard used for entering and searching the digital data.

The results were as expected. In Error! Reference source not found. we see all eight dialect regions that we usually identify in Crișana. There are recognizable clusters, although they show a lot of internal variation (as we have observed in our MDS pictures of the region). But there is also the expected spread with the south further apart from the central and northern regions.
This becomes clearer when we look at a subset of the dialects. In Error! Reference source not found., we see two dialect areas (Bihor, Sălaj) that are geographically in the centre of Crișana, and further away, the two areas from the South East and South West, which are generally recognized as distinct from the others.

The result, then, is we see the use of this very simple measure, Character Profile, showing a pattern that can be explained by Gravity and Spread, just as we could in our earlier studies, even though the nature of the underlying texts is quite different from authored texts such as a Dickens novel.

Illustration 5 Four dialect areas, showing both clustering (from Gravity) and differentiation (from Spread). The areas from bottom to top are Bihor, Sălaj, SouthWest and SouthEast

SUMMARY

Dorin Uritescu was a key member of our research team, providing not only our window onto Romanian, but also linguistic insight and research wisdom that allowed us to go into so many aspects of the description and measurement of dialect and dialect differences. The work goes on, with the study we report here, and we hope that others will be able to leverage his work for many years to come.

REFERENCES


[EUW] see below


DORIN URITESCU AND THE ROMANIAN ONLINE DIALECT ATLAS

DORIN URITESCU AND THE ROMANIAN ONLINE DIALECT ATLAS: CONTRIBUTIONS TO DIALECTOLOGY

(abstract)

Dorin Uritescu’s collaboration with the authors has produced the Romanian Online Dialect Atlas (RODA) and many other studies in phonology and dialectology. Here we recap some of this work, and introduce a new, subsequent study on the use of “Character Profiles” for measuring dialect differences.

Keywords: dialectology, multidimensional scaling, character profile, Romanian, Crișana, gravity and spread, dialect identification.

Cuvinte-cheie: dialectologie, scalare multidimensională, profil caracteristic, română, Crișana, graviitație și răspândire, identificare a dialectului.

York University, Toronto, 4700 Keele Street, Ontario, Canada, M3J 1P3
Laurentian University, Sudbury, 935 Ramsey Lake Rd, Ontario, Canada, P3E 2C6
embleton@yorku.ca

York University, Toronto, 33 Peter Street, Markham, Ontario Canada, L3P 2A5
eric.wheeler@sympatico.ca