Romanian Linguistic Atlases in Digital Format – A New Approach[1]

Silviu-Ioan BEJINARIU[*], Florin-Teodor OLARIU[**]

Keywords: *geolinguistics, linguistic databases, Linguistic Atlases Publishing System (LAPS), Geographic Information System (GIS)*

1. Introduction

Romanian dialectologists' experience in linguistic cartography is rich – Romanian language is the only one currently benefiting from three generations of linguistic atlases: WLAD, ALR, and ALRR. *Synthesis*. One of the major directions of further development in Romanian geolinguistics is related to the use of information technology in this research field, currently evolving in two directions: the direction of developing and expanding the computerized applications for existing linguistic atlases digitalization at national level, or that of configuring new applications to facilitate the creation of linguistic atlantography works.

The latter direction is the initiative of the researchers at the Iasi Academy research centre on the implementation of a software system that allows the acoustic analysis and automatic generation of phonetic transcriptions; it can be applied for the (semi)automatic transcription of the dialectal audio material using the phonetic transcription system specific to ALR / NALR projects. Beginning in 2015, the research in this direction has reached the point where the application configured specifically for this purpose can propose a number of possible transcription variants, and the user has to choose between the displayed solutions (Botoşineanu et al. 2016).

The most significant achievement in the cooperation between Romanian geolinguists and computer scientists is the *Romanian Online Dialect Atlas* (RODA) application, based on recent studies in the field of dialectometry and quantitative linguistics (statistical techniques such as multidimensional scaling). First used in the development of NALR. *Crişana*, the system is also currently used to publish one of the most recent Romanian geolinguistics projects, namely the *Romanian Linguistic Atlas. Spoken Languages between Morava, Danube and Timoc* (ALR – MDT), as mentioned in a recent article by Dorin Uriţescu, one of the authors and promoters of the RODA applications (Uriţescu 2014: 371). Research within the Iasi Branch of the

[*] Institute of Computer Science, Romanian Academy, Iasi Branch.

[**] "A. Philippide" Institute of Romanian Philology, Romanian Academy, Iasi Branch.

[1] *Acknowledgements.* This work was done under the research grant "*The Audio-Visual Linguistic Atlas of Bukovina (ALAB). The Second Stage*", PN-II-RU-TE-2014-4-0880.

Romanian Academy is currently developing towards the creation of a new architecture for the initial application, which has been used since 2005 to publish the NALR. *Moldova and Bukovina* (NALR-MB), allowing the design of linguistic atlases regardless of the dialectal area studied (Moldova, Banat, Wallachia, etc.). The impulse for this new approach came from two directions: a) the editing of the phonetic transcription in the second stage of the *Audiovisual Linguistic Atlas of Bukovina (ALAB)* project, which involved access to a publishing system similar to that used in the publication of the previous volumes (III and IV) of the NALR. *Moldova and Bukovina*; and b) the request from the team working on the *Linguistic Atlas of the Aromanian Dialect (ALAR)* to develop a software application that would allow the digitization of this atlas (Olariu 2017: 86).

By cumulating the requests from the two teams of dialectologists, the computer science researchers decided to move the first application to a higher degree of generalization, reconfiguring its original architecture. The most important element in this process is the use of the Geographic Information Systems (GIS), which is a first in Romanian geolinguistics research. The system's major advantage is the use of the "geographical position as a common factor, allowing interconnection of data and the joint analysis of information from different domains, including geolinguistics and dialectology" (Bejinariu et alii 2016: 39). The GIS system features are used both in generating linguistic maps and in creating databases that can easily handle large amounts of information. The final goal of this research is to "create a standard model for all Romanian linguistic atlases, allowing future quantitative analysis of geographic variables" (*ibidem*: 44). In the following sections, we will present the structure and functioning of this new *Linguistic Atlases Publishing System* (LAPS), focusing on the linguistic and geographical databases model and AlrMaps application's main functions, namely the design of linguistic (analytical and synthetic) maps and uncartographed materials.

2. Data Model

The main goals of the data model redesign were (1) to eliminate the hard-coded information (Bejinariu et alii 2009) which restricted the previous version of LAPS for the preparation of NALR-MB maps only and (2) to provide the possibility of introducing other types of information into the linguistic database.

The proposed model is project oriented and it is based on the use of two interconnected databases which contain the linguistic and geographical information. As presented in Fig. 1, the two working databases are specified in the project definition, which includes also the particular settings of the map layout (colour, text size, phonetic transcriptions size, etc.).

The connection between the two databases is made through the inquiry point standard codes, which offers the opportunity to use the same linguistic database in different projects. This is the first step to create a unique dialectal database of all Romanian regions, a major research project that Romanian geolinguistics will have to assume for the future.

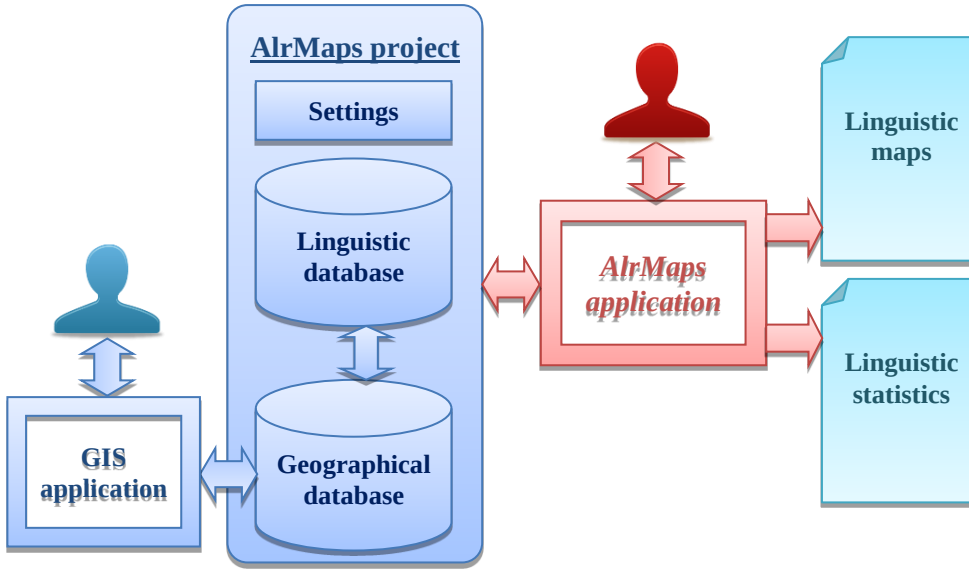

Figure 1. *Structure of linguistic atlases publishing system*

3. Linguistic Database

In the previous version, the linguistic information was stored in three different files (title words, inquiry points and phonetic transcriptions) which made the information contained in different databases difficult to correlate and unify. In the new model, the information is stored in a single file. Moreover, the phonetic transcriptions and their groups are included in the title word descriptions. Thus, the new model was designed as two tables: title words and inquiry points, both being included in a single dictionary file.

3.1. "Inquiry points" description

In contrast to the previous version, the "Inquiry points" description contains more localization information as: region, country and information about the interviewee. Through this redesign of the application and database, extralinguistic information, which in the classical printed version of the linguistic atlases was included in the introductory volume entitled *Information about localities and interviewees*, is now brought to the main page of the project; this change will contribute to a much better contextualized analysis of the linguistic data included in the project database.

The inquiry points are organized as a list of records, each containing the following fields:

– code of the inquiry point;

– ALR and WLAD code of the inquiry point, to preserve the correspondence with other Romanian linguisic atlases;

– name of the inquiry point;

– name of the commune / city to which the inquiry point belongs;

– name of the county;

– name of the region – used in case of Romanian regional linguistic atlases, for an easy classification of the inquiry points according to the historical region in which it is included;

– name of the country – this information is useful in case of ALAR (*Aromanian Linguistic Atlas*) that includes inquiry points from different countries (Albania, Bulgaria, Greece and Macedonia);

– local dialect – used in case of ALAR to specify the local language variety used in each inquiry point; this information can also be specified for the Daco-Romanian dialect when linguistic atlases for the entire North-Danubian area will be created;

– information about the interviewee: age, gender, ethnicity, level of education, occupation; this information group offers, on the one hand, the possibility of further analyses related to the diastratic (possibly diaphasic) variation present within the studied local dialect and, on the other hand, the possibility of introducing in the database several phonetic transcripts for the same title word and inquiry point, but for several categories of speakers; this information is required in the case of the ALAB project, focusing on the recording of diasexual and diagenerational variations in the multiethnic area of Bukovina, where, besides the Romanian language from the region, the Houtzoul, Polish, Lipovan and Ukrainian languages are also spoken in that ethnolinguistic space;

– year of the interview and other useful information.

In the figure below, the interface for the inquiry points information editing is presented.

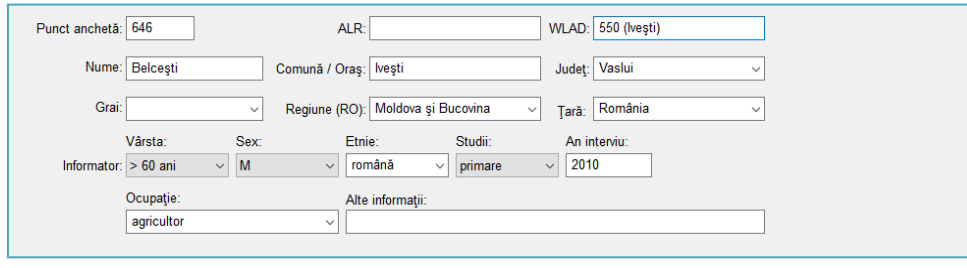

Figure 2. *Interface for "Inquiry points" information editing*

All data related to the inquiry points are text type, directly editable or selectable from a set of predefined values. The internal representation includes also a unique identifier used to link the inquiry points to the table that contains the title words descriptions.

Excepting the Code and the Name of the inquiry point, all other information is optional, not mandatory for the generation of linguistic map, but useful for other type of linguistic information processing. Also, it must be said that the identity of the interviewee is not included in the database as a measure of personal data protection. An exception to this rule is the ALAB project where the consent of the interviewees was obtained for the use of their personal data (name, professional training, age and especially their image) in order to publish the video interviews on the project site.

3.2. Title Words description

As was mentioned before, the title word description includes the phonetic transcriptions and their groups in the same data structure. The advantage of this new data structure is that it reduces the possibility of errors in importing and exporting data between different dictionaries. The interface used for title word definitions is presented in Fig. 3. The following information can be stored:

– the title word;

– the question to which the title word is the answer and the question number;

– the notes (I and III) associated to the title word;

– associated image – there is the possibility to define a set of associated images, one of which is the default image and is displayed on the linguistic map;

– number of page and number of linguistic plate in the linguistic atlas; if specified, this information is presented in the linguistic map;

– mapped material – this information is used for the automatic generation of the pages in which the lists of mapped and unmapped material is presented.

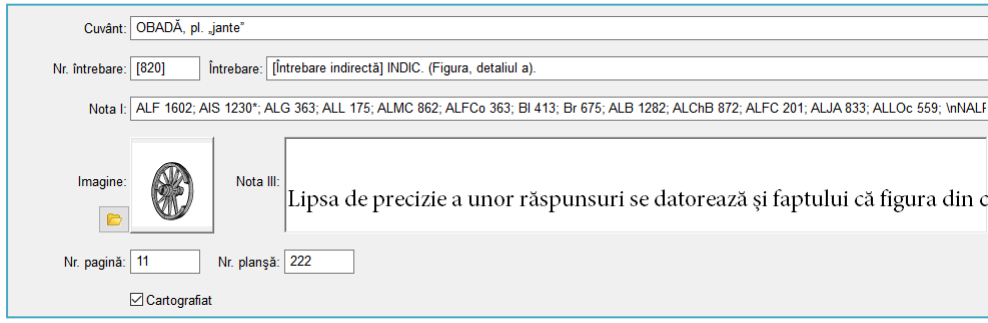

Figure 3. *Interface for "Title Words" information editing*

3.2.1. Phonetic transcriptions

The title word description also includes the list of phonetic transcriptions associated for all defined inquiry points. The link between words and points is made on the basis of unique identifiers, in a transparent manner for the user, using the interface presented in Fig. 4.

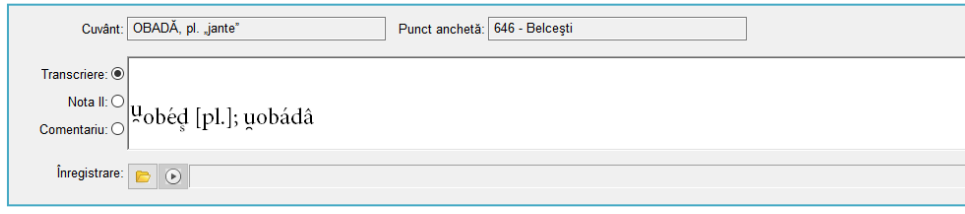

Figure 4. *Interface for "Phonetic transcription" editing*

For each pair *title word – inquiry point*, the following information can be edited using the phonetic transcription format:

– the phonetic transcription of the answer, using the ALR transcription system;

– note II, used to continue the answer when the space reserved on the map is insufficient for the full answer transcription or for some additional information specified by the interviewee related to the answers given to the questions in the questionnaire;

– comments;

– audio recording – there is the possibility to add a set of audio / video recordings of the answers in the database. In the case of interviews conducted with media recording, the audio / video sequence can be used both to facilitate the transcription editing and to create a multimedia linguistic atlas as is the case of ALAB. It must be said that the database does not contain the recording but the path of the file in which the recording is stored. Because the stored path is relative to the dictionary file path, it is recommended to store the recordings in subfolders of project folder. This also applies to images associated to the title words.

In the previous version of the phonetic transcription editing the basic symbols were selected using a large number of key combinations. They are still available and are presented as a list in a control bar on the application interface. There is also the possibility to copy / paste them from this list to the edited transcription.

3.2.2. Groups

In order to automatically generate uncartographed material, linguistic plates as well as synthetic maps, it is necessary to group the inquiry points according to some common linguistic features (phonetic, morphological, lexical, and semantic) of the answers to the question asked for the documented notion. The fully redesigned interface allows to define groups using a single editing window (as opposed the to previous version of the application), as shown in the figure below. The following facilities are available:

– legend title editing for the case of synthetic maps;

– creating, removing and modifying the order of edited forms of the title word;

– editing of form name as well as of the graphic features used for synthetic map generation (colour and style of hatches and borders as well as the typology of symbols);

– adding and removing inquiry points to / from the list associated with a form. Unlike the previous version, the inquiry points can be associated to several forms – a technically essential aspect, given that several forms or variants (phonetic, semantic, lexical) can appear in the same inquiry point for the same documented notion;

– specifying a comment for each inquiry point associated to a form.

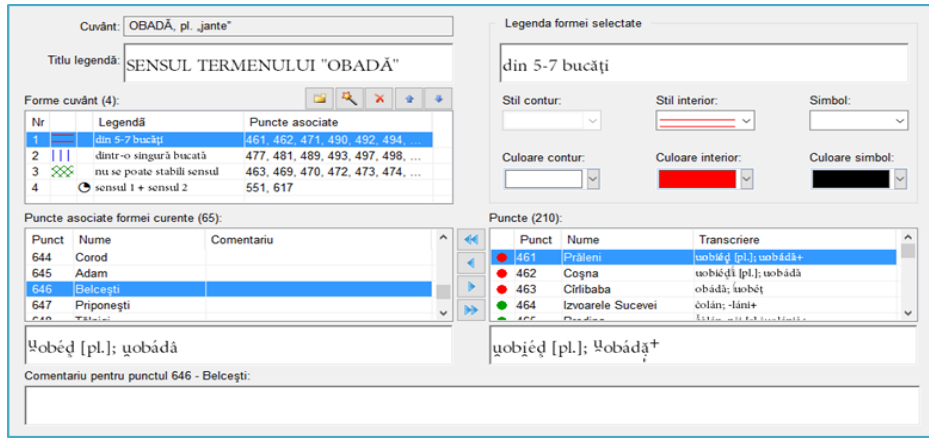

Figure 5. *Interface for grouping modes definition*

4. Geographical Database

Given the fact that the geographical position is an important property of information in geolinguistics and it is processed, analyzed or viewed in graphic form as linguistic maps, the solution for the geographic database design was the use of a Geographic Information System (GIS) application. An advantage of the GIS system is that the geographic database may include the regions of Romania as well as other regions. If the linguistic database is complete in the sense that it includes the linguistic information for a particular notion in all regional atlases, the analytical maps corresponding to these fully documented concepts, as well as the synthetic maps, can be generated automatically using the AlrMaps application. Also, linguistic maps for other historical regions such as Bukovina (cf. ALAB) can be automatically generated once the graphic coordinates of the region concerned have been introduced into the application.

On the other hand, the final linguistic plates published in the linguistic atlases include other specific elements (layout) that are prepared independently from the linguistic contents of the maps. They are designed separately in the same GIS application, and the actual map of the studied area is selected and transferred into the layout (Fig. 6). To this end, all elements of the map are organized in layers according to their significance (Bejinariu et alii 2016). In its current form, the geographic database contains 31 layers, divided into 3 categories, depending on the elements they define on the final sheet: page layout, linguistic map, synthetic map. The case of NALR. *Moldova and Bukovina* linguistic plates is presented in Fig. 6.

The first category includes layers that define the size and position for: page title, borders, title word, word-associated notes, page number and sheet, and image associated with the title word. In the second category are included the layers that define: the internal and external borders for the studied area, the watercourses and their names, the big cities and their names, the inquiry points, the spaces reserved for the phonetic transcripts and their associated notes. The last category indicates the positioning of the synthetic map and it includes layers that define the position of the legend and a number of layers similar to those in the second category: borders, watercourses, inquiry points, etc.

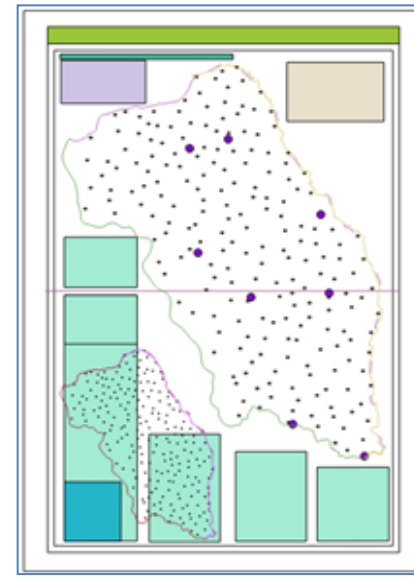

Figure 6. *Layout of NALR-MB linguistic plates*

All objects placed on a linguistic plate are characterized by two types of features: (a) fixed features, such as position, horizontal and vertical alignment for text objects and phonetic transcriptions, and (b) variable features, such as colour, font size, drawing mode (opaque or transparent). Fixed features are set at the layout level in the GIS representation of the map, and cannot be

modified by the user in the AlrMaps system, while in the case of variable features, new values can be selected either for a particular element or for all elements on a particular layer. Fixed feature values are encoded in the database fields associated with each layer in the GIS representation. Because the format used by the GIS application is comparatively complex, when the linguistic linguistic plate is automatically generated, it is converted and saved in an AlrMaps specific format, and all subsequent changes will be saved in this new format.

5. Linguistic Plates

The main goal of the linguistic atlas publishing system is to produce linguistic plates (maps and uncartographed material). For the beginning, these are automatically generated, and then the user can modify the linguistic plates to meet aesthetic criteria or to correct any errors. Because in the previous version of the system the images of phonetic transcriptions were stored in the linguistic plate description, any changes in the dictionary files involved re-generating the linguistic plates, obviously with the loss of all previously made changes. In the new version of the system, only the positioning and drawing properties (colour, thickness and line type, font) of the elements defined in the linguistic database (title, transcriptions, notes, inquiry points numbers, page number) are stored in the drawing document. The values stored in the dictionary are retrieved from it at each redraw, which allows any changes to the content to be immediately reflected in the linguistic plate.

The new version of the system includes both NALR. *Moldova and Bukovina* specific features and those that meet the requirements of other atlases such as ALAR. Thus, the linguistic plates that AlrMaps can automatically generate are: linguistic maps, cartographed and uncartographed material, list of title words presented as cartographed and uncartographed material. The new system introduces new facilities for the user in terms of the linguistic maps preparation, but the final result is relatively similar excepting the graphical quality of the phonetic transcriptions, which is higher. The generation of the other types of linguistic plates is completely different.

5.1. Linguistic maps

After the automatic generation, the linguistic map includes 4 types of basic objects: (a) Graphical elements: borders, internal and external borders, watercourses; (b) text objects: title of the atlas, title word, name of major cities and codes of the inquiry points; (c) phonetic transcriptions and (d) symbols and images. If a synthetic map is also generated then two grouped objects, corresponding to the synthetic map itself and its legend, are created automatically. These contain simpler objects of the types mentioned above. Objects in the linguistic map are also organized on layers, which generally correspond to the layers in the geographic database.

The linguistic map editing is illustrated in Fig. 7. In this example, the properties of the phonetic transcription of title word "*obadă*", corresponding to inquiry point 538 of the NALR. *Moldova and Bukovina* network, are presented. The properties are available in the control panel throughout the edit period and any changes are immediately reflected on the map image. Regardless of the type of active layer (in this example: phonetic transcriptions), there are 4 groups of settings.

The first category of settings controls the display of graphical elements and associated text – if applicable. The second category is informative in the sense that the settings cannot be changed by the user and specifies the permissions for the editing actions (selection, editing, movement, resizing, copy/paste). The third category of properties allows the actual editing of the properties of the selected object, in this case, for the phonetic transcription: size of the used font, colour and drawing mode. If the phonetic transcription editing is initiated, the corresponding window is activated (Fig. 4). Any changes to the phonetic transcription will be immediately visible on the linguistic map. The list of properties in this category depends on the active layer and selected object type. In the last category of properties, the values associated with the element selected in the geographic database are presented in an informative manner (the inquiry point code and the alignment modes in both directions, vertical and horizontal – Fig. 7). After changing the properties of an object, they can be automatically transferred to all objects on the same layer in the current map, or they can become default for all the maps generated from the current project. Regarding how the areas with common phonetic properties from the synthetic maps are drawn, these settings are defined when the groups have been edited and can be modified using this interface only (Fig. 5).

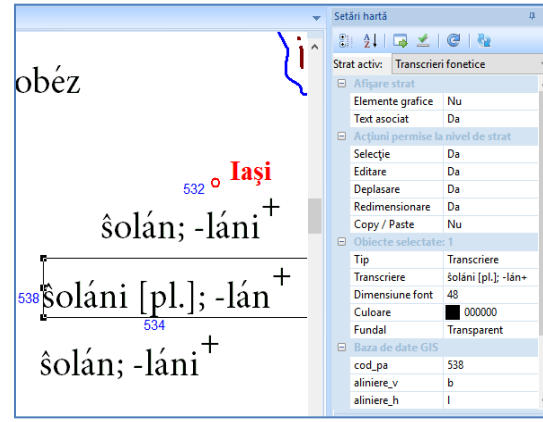

Figure 7. *Editing window for the objects in the linguistic map*

Compared to the original version of the system, the drawing mode has been improved, especially for phonetic transcriptions, which are legible even when in small font sizes.

5.2. Linguistic plates of cartographed and uncartographed material

First of all, it must be noticed that in some atlases, such as ALAR, there are also printed linguistic plates that contain the cartographed material in tabular form, while the notes corresponding to the cartographed and uncartographed material are published in tabular form on separate sheets. A new automatic generation method was designed for the types of linguistic plates that do not contain uncartographed material, as well as for lists of these materials; this method is currently being implemented. Linguistic plates will be generated in Microsoft Word format using a set of layouts specific to each type of linguistic plates. Using the Application Programming Interface of the Microsoft Office package, normal text will be inserted in the classical way using fonts, while the phonetic transcriptions will be inserted as images after the rows have been previously distributed where appropriate. Since all font-related calculations are made with the system's default printer, for an improved print quality it is desirable to generate the linguistic plates of this type for the printer that will be finally used.

6. Conclusions

In the current context of large-scale digitizing of linguistic resources required in philology, the development of increasingly performing applications facilitating this innovative approach becomes a major desideratum for the specialists. In addition to editing dictionaries or critical editions in electronic format, the digitization of linguistic atlases which are the fundamental primary resources for the study of any language, was responsibly assumed by the Romanian researchers, and there are already significant achievements in all the three important academic centers for Romanian geolinguistics: Cluj-Napoca, Bucharest and Iasi. New perspectives are opened with the new IT solutions designed by the computer scientists from the Iasi academic center, in collaboration with dialectologists from Iasi and Bucharest, for the digitization of the Romanian geolinguistic works, by creating efficient and easily accessible instruments for the variational study of the Romanian language, and last but not least, in synchronizing Romanian research with that in Europe.

References

Bejinariu et alii 2009: Silviu-Ioan Bejinariu, Vasile Apopei, Stelian Dumistrăcel, Horia-Nicolai Teodorescu, *Overview of the Integrated system for dialectal text editing and Romanian Linguistic Atlas publishing – 2009*, The 13th International Conference „INVENTICA 2009", Iaşi, Performantica Publishing House, 2009, p. 564–572.

Bejinariu et alii 2016: Silviu-Ioan Bejinariu, Ramona Luca, Florin-Teodor Olariu, *A GIS Based Approach for Information Management in Geolinguistics*, "Memoirs of the Scientific Sections of the Romanian Academy", Tome XXXIX, 2016, Computer Science, 2016, p. 37–45.

Botoşineanu et alii 2013: Lumini Botoşineanu, Florin-Teodor Olariu, Silviu-Ioan Bejinariu, *Un projet d'informatisation dans la cartographie linguistique roumaine: "Noul Atlas lingvistic român, pe regiuni. Moldova şi Bucovina" en format électronique (e-NALR) – réalisations et perspectives*, în Casanova Herrero, Cesareo Calvo Rigual (eds.), *Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas, 6-11 septiembre 2010, Valencia*, Berlin, Walter de Gruyter, 2013, vol. VI, p. 2921–2930.

Botoşineanu et alii 2016: Luminiţa Botoşineanu, Elena Muscă, Florin-Teodor Olariu, Ioan Păvăloi, *Aspects relatifs à la transcription phonétique interactive du signal audio*, în Luminiţa Botoşineanu, Ofelia Ichim Florin-Teodor Olariu (eds.), *Linguistic and Cultural Contacts in the Romanian Space – Romanian Linguistic and Cultural Contacts in the European Space*, Roma, Aracne Editrice, Italia, 2016, 500 p. 33–48.

Olariu et alii 2016: Florin-Teodor Olariu, Veronica Olariu, Marius-Radu Clim, Ramona Luca, *La cartographie linguistique roumaine face à l'informatisation : quelques projets et résultats*, în Jean-Paul Chauveau, Marcello Barbato, Inés Fernández-Ordóñez (eds.), *Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 8 : Linguistique variationnelle, dialectologie et sociolinguistique*, Nancy, ATILF, 2016, p. 193–207.

Olariu 2017: Florin-Teodor Olariu, *Variaţie şi varietăţi în limba română. Studii de dialectologie şi sociolingvistică*, Iaşi, Editura Institutul European.

Uriţescu 2014: Dorin Uriţescu, *Graiul din Ţara Oaşului în perspectivă informatică*, în Gheorghe Chivu, Oana Uţă Bărbulescu (eds.), *Ioan Coteanu – in memoriam*, Bucureşti, Editura Universităţii din Bucureşti, p. 369–380.

Abstract

This paper focuses on the new designed data model of AlrMaps and new possibilities of linguistic maps generation and editing. The first version of the Linguistic Atlases Publishing System (LAPS) was developed to enable the publishing of the *New Romanian Linguistic Atlas by Regions. Moldova and Bukovina* (NALR-MB). The third and the fourth volumes were prepared using LAPS. Over time, the interest of dialectologists for computer-aided preparation of the linguistic atlases has increased and new requests have been received to adapt the system for the *Audio-Visual Linguistic Atlas of Bukovina* (ALAB) and the *Aromanian Linguistic Atlas* (ALAR) publishing.

Considering the previously accumulated experience, a new version of LAPS entitled AlrMaps was developed in order to eliminate the shortcomings identified, to make the system independent of the map and region for which it will be used and last but not least to create a more friendly interface using the new tools available as a result of software and hardware progress. One of the main innovations of the new LAPS is the integration of GIS functions into the archiving and processing of linguistic data. In the future this aspect of the new application will facilitate the development of national geolinguistic projects for the Romanian dialectologists, by also using multimedia technologies in the configuration of complex, multidimensional linguistic databases.