Memoirs of the Scientific Sections of the RomanianAcademy
Tome XXXIX, 2016

COMPUTER SCIENCE

A GIS-BASED APPROACH FOR INFORMATION MANAGEMENT IN GEOLINGUISTICS

SILVIU-IOAN BEJINARIU[1], RAMONA LUCA[1] and FLORIN-TEODOR OLARIU[2]

[1]*Institute of Computer Science, RomanianAcademyIasi Branch, Iasi, Romania*
[2]*"A. Philippide" Institute of Romanian Philology, RomanianAcademyIasi Branch, Iasi, Romania*
Corresponding author: silviu.bejinariu@iit.academiaromana-is.ro

The geographic position is an important property of linguistic information in geolinguistics. This paper proposes a Geographic Information System (GIS) based framework for linguistic information management, consulting and analysis. Dialectal information is presented and consulted mainly in graphical format, as maps in linguistic atlases. These atlases are periodically published for each historical region of Romania, but each of them is prepared using different data structures and codification methods. Maps are drawn as graphics without a real localization based on geographic coordinates, which makes difficult further spatial analysis of the linguistic information. In the proposed framework, linguistic information is an attribute of a geographic location benefiting from the advantages of a GIS-based system. In this paper, the framework is briefly described and the data models for geographical and linguistic information are detailed. The GIS-based framework allows future developments as spatial analysis and statistics of linguistic data.

Keywords: geolinguistics, Geographic Information System, linguistic maps, database.

1. INTRODUCTION

In geolinguistics, the geographic position is an important property of linguistic information, which makes possible not only its graphical representation as maps in linguistic atlases [2] but also spatial analysis. The linguistic atlases are periodically published by the Romanian Academy for seven historical regions of Romania: *Muntenia şi Dobrogea*, *Oltenia*, *Banat*, *Crişana*, *Maramureş*, *Transilvania* and *Moldova şi Bucovina*. Besides these regional linguistic atlases, the Romanian Academy publishes also a *Sinteza* volume in which the spelling variations are presented for the entire Romania area. Each of these volumes is prepared using different data structures and codification methods. The main issue is represented by the large number of symbols used in the phonetic transcription of the Romanian language. Our knowledge is that, until now, only volumes III and IV of the *Moldova şi Bucovina* atlas, volume III of the *Crişana* atlas and partially volumes IV and V of the *Muntenia*

şi Dobrogea atlas are prepared using specialized software. It must be noticed that for the *Sinteza* volume preparation, the maps are redesigned even if the same could be automatically obtained by unifying the corresponding maps already included in the regional atlases. Unfortunately, this is not possible, due to the different system used for each regional atlas preparation. Sometimes, in the context of other research projects, different maps have to be generated for specific dialectal studies [9]. The same symbols are used for the phonetic transcriptions included in the *Basarabia, Nordul Bucovinei şi Transnistria* atlas published by the Academy of Sciences of Moldova. Also, it is only partially prepared using a specialized software tool.

Excepting the volumes of the *MoldovaşiBucovina* atlas, which are prepared using a dedicated application able to generate the graphical symbols for all phonetic phenomena combinations [1, 2], all the other atlases are prepared using specific and limited fonts.

The paper proposes a new GIS (Geographic Information System) based framework for the publishing system developed about ten years ago by researchers from Institute of Computer Science and "A. Philippide" Institute of Romanian Philology. Using the previous system, two printed volumes of *Noul Atlas Lingvistic Român pe Regiuni - Moldova şi Bucovina (The New Romanian Linguistic Atlas by Regions - Moldavia and Bukovina,* abrev. *NALR-MB*) were published (Fig. 1). The most important issues of the previous work were linked to:

– information encoding – given the large number of variants, the Unicode coding of characters was used prior to its becoming the implicit coding method used by Windows operating systems;

– definition of data structures – 3 different dictionaries were used, locally stored in files, without the possibility of concurrent operations;

– maps generation, editing and printing – a hard-coded map with predefined location of included elements.

Some analysis instruments were also developed, including the synthetic maps that present in a graphical manner the linguistic particularities in different geographical areas [1–3, 7, 8]. The software system was particularized for the case of the *NALR-MB*. An attempt was made at adapting the system for the *Crişana* region. A large number of system components were redesigned, including the linguistic map, the most important component.

All these led to the idea of a GIS-based system for geographical data manipulation and data storage in a relational database. The GIS-based system allows increasing of the map design possibilities with higher interactivity, while giving the user the possibility to choose the studied geographic area [12]. The most common situations are those in which the linguistic map of multiple regions, the entire country or reduced areas of a region must be generated. All these facilities are obtained by exploiting the advantages of a GIS system.

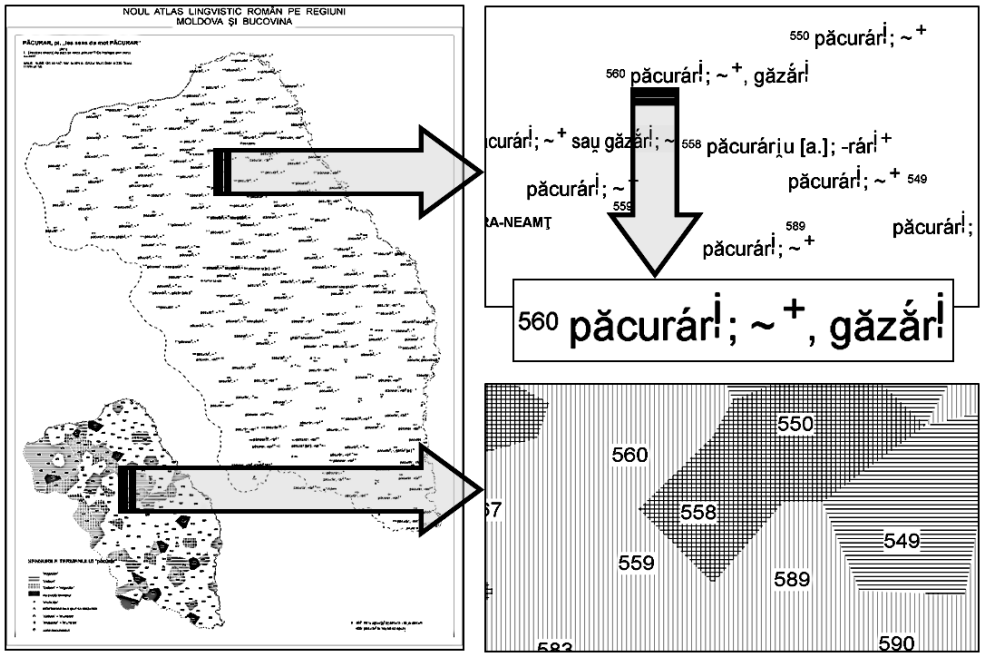

Fig. 1. Details from a linguistic map included in *NALR-MB*, vol. III. Source: component images are captured from ALR-IIT application, designed by authors [1,2].

2. GEOGRAPHIC INFORMATION SYSTEMS IN DIGITAL LINGUISTIC CARTOGRAPHY

The most important attribute of a GIS system is that it uses the geographical position as a common factor, allowing interconnection of data and joint analysis of information from different domains, including here geolinguistics and dialectology.

2.1. GEOGRAPHIC INFORMATION SYSTEMS

GIS is defined as an information system which allows the acquisition, storage, validation, integration, manipulation, analysis and visualization of data related to various points on earth's surface (geospatial data) [12]. A more complete definition of GIS, starting from the fact that it integrates all the resources used for geospatial data processing is: GIS is a coherent assembly of equipment, applications, people, rules, information and methods used for the management of geographical-related descriptive data, by correlating them with the topographical maps of the area [12]. A geospatial database integrates information from different sources and the association to the geographical position allows reducing of redundancy, databases inter-connection and – the most important advantage in our case – maps visualization with different levels of details.

The geospatial database includes: (1) the digital map – geometric description (shape, position and size) of spatial objects in the studied area. Depending on their type, spatial objects are organized in layers; (2) internal database – descriptive data that specify the attributes of the objects included in the digital map. Each layer has an associated table and each element in the map has an associated record in the corresponding table; (3) external databases – other application-specific information which does not correspond to a specific element in the map.

2.2. GIS IN GEOLINGUISTICS

One of the first GIS-based approaches in linguistic data management [6] describes the application of quantitative spatial analysis and GIS functions to the analysis of language data. As input data, the Linguistic Atlas of the Middle and South Atlantic States was used in the analysis. The authors focused on the frequency of the use of a word among a particular set of speakers. In [5] it is presented a review of recent GIS applications in linguistics, including linguistic atlases, lexical and phonological surveys, and a sociolinguistic analysis. The authors conclude that integration of GIS and linguistics requires the expertise of linguists and data processing and visualization skills of both disciplines. A GIS-based infrastructure for linguistics information management is presented in [10]. It is one of the first operational Web-based platforms that supports transparent and interoperable integration of all possible linguistic and related data. It is also a platform dedicated for collaborative research in linguistics. The portal is based on ArcGIS – one of the most used GIS infrastructures.

3. PROPOSED GIS-BASED APPROACH

In the next paragraphs, some issues related to the proposed GIS-based linguistic information management system are presented. The geospatial database contains mainly the map of the studied area and some identification and link information in the internal database. The external database includes linguistic information and a full sketch customizable description. For instance, only the descriptions of the *Moldova şi Bucovina* atlas and of the *Bucovina* multimedia atlas are included. In future, all maps descriptions will be added in the database using the configuration facilities of the system.

3.1. GEOSPATIAL DATABASE

The connection between geographic and linguistic information is most obvious in the linguistic maps, which represent the main content of linguistic atlases. In our previous implementation, used to prepare two volumes of *NALR-MB*, the geospatial database and its links to linguistic information were hard-coded, which rendered difficult to implement any changes or re-design demands. Analysis of the already implemented systems (devoted to the three atlases: *Moldova şi Bucovina*, *Crişana* and *Bucovina*) showed that the graphical content of these maps may be classified as (Table I):

– *page layout* (Page) – which includes the size and location of the page title, borders, page number, notes, etc.; the layout may vary, as it depends on the final linguistic atlas requirements, consequently it should be re-designed for each particular situation;

– *linguistic map* (LM) – including the studied area borders, location of main cities, rivers, inquiry points and, the most important of all, location of the phonetic transcriptions;

– *synthetic maps* (SM) – a reduced version of the linguistic map in which common dialectal features of the spoken language are highlighted using colors and/ or hatches.

Table 1 contains all layers defined in previous implementations, their type and, in the last three columns, their effective usage in different atlases.

Table 1
Layers used in linguistic maps preparation

	Layers	Type	Moldova_Bucovina	Crișana	Bucovina
Page Layout	Page_Limits	polygon	*	*	*
	Page_Title	polygon	*	*	*
	Page_Frame	polygon	*	*	
	Page_Title_Word	polygon	*	*	*
	Page_Notes_Word	polygon	*		
	Page_Sketch_Number	polygon		*	
	Page_Notes	polygon	*	*	
	Page_Image	polygon	*	*	
	Page_Divider	polyline	*		
Linguistic map	LM_Border_Ext	polyline	*	*	*
	LM_Border_Int	polyline	*	*	*
	LM_River	polyline	*		*
	LM_City_Name	polygon	*		*
	LM_City	point	*		*
	LM_Transcription	polygon	*	*	*
	LM_Inquiry_Point	point	*	*	*
Synthetic map	SM_Frame	polygon	*	*	
	SM_Profile	polygon		*	
	SM_Legend	polygon	*	*	
	SM_Border_Ext	polyline	*	*	
	SM_Border_Int	polyline	*	*	
	SM_Inquiry_Point	point	*	*	
	SM_River	polyline		*	
	SM_River_Name	polygon		*	
	SM_City	point		*	
	SM_City_Name	polygon		*	

In the proposed approach, all layers included in the linguistic and synthetic maps would be prepared using a GIS platform, using the standard shape *format*

(Fig. 2). For easily handling, the page layout layers are prepared in GIS style, even if they have not the geographic location as property. One should also notice that information in the linguistic and synthetic maps is redundant and will be unified in a single description. Different scaling and rendering options will allow obtaining of the specific appearance of the two maps.

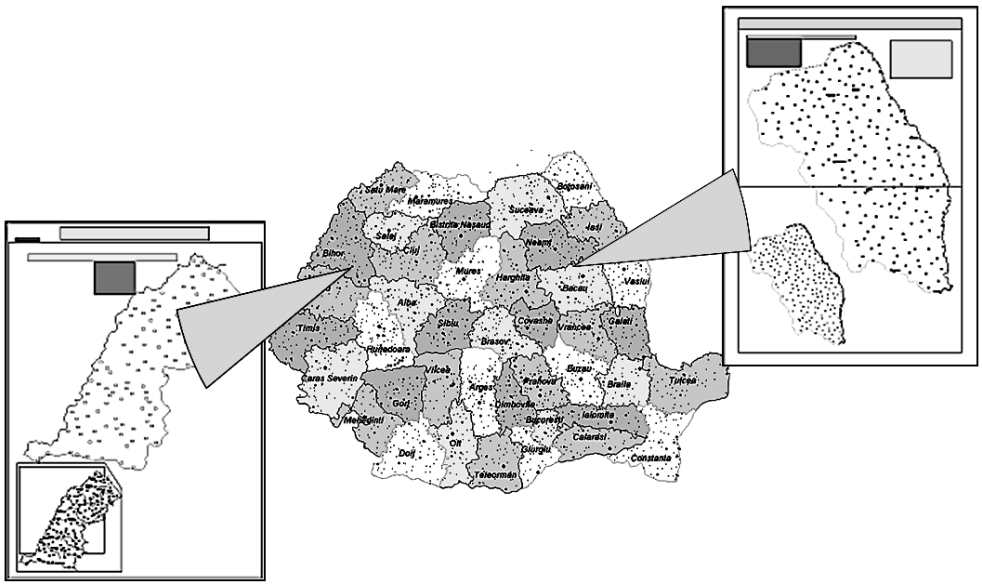

Fig. 2. GIS-based approach of linguistic maps. The *Crişana* and the *Moldova şi Bucovina* maps are parts of the map of Romania. Source: component images are designed by authors and taken over from NetSET Map application [11].

To allow a unified processing method of all Romanian linguistic atlases, we propose to use a single map that includes all the required geographical elements for all historical regions as a geospatial database. In contrast to a topographical map on which all villages are present, in our map only the inquiry points are included. To avoid the phonetic transcription overlapping in case of close inquiry points, in the previous version of the system their positions were established in the map design step. Even so, for aesthetical reasons, these positions were changed in the final editing step. In the proposed system, the phonetic transcription positions are specified relatively to the corresponding inquiry point position.

In the associated internal databases, specific information, such as: region membership, name of the geographic feature and links to the information in the external databases would be included.

3.2. LINGUISTIC DATA STORAGE

In the previous version of the system, linguistic information is stored in three dictionaries:

– *title words* dictionary – contains studied words description;

– *inquiry points* dictionary – contains inquiry points description;

– *phonetic transcription* dictionary – contains a record for each pair < *title word, inquiry point* > including the phonetic transcription.

All information about the known meanings of the title word is encoded in the corresponding dictionary, and inquiry point grouping mode based on similar pronunciations particularities are encoded in the phonetic transcription dictionary. Other linguistic map settings are hard-coded in the application.

In the proposed system, the linguistic information is stored in an external database based on the MySQL database management system. The preliminary version of the relational database "*nalr*" would contain the following main tables:

– *Words* table – description of all title words. The structure is similar to the initially designed dictionary, except the images and sounds which are stored now in separate tables.

– *Word_meanings* table – a record for each "meaning" associated to a title word.

– *Multimedia* table – all the audio/video recordings related to the title words of the linguistic atlas.

– *Images* table – specifies the images associated to a title word.

– *Inquiry_points* table – description of all the inquiry points.

– *Dictionary* table – the main table of the database, containing all phonetical transcriptions, foot notes and comments for each title word and for each inquiry point.

– *Dict_meanings* table – information about the "*meanings*" associated to a title word, inquiry points where the "*meaning*" is used.

– *Dict_groups* table – used to define the inquiry points grouping criteria and groups for each title word.

– *Atlas_description* – settings used to define the layout of maps included in each atlas.

– *Map_description* – used to define the drawing settings for each map.

It must be noticed that this structure may be subject of further small changes, depending on each regional atlas specific requirements.

3.3. OTHER ISSUES

The main components of the proposed system are: the database management module and the map generation module.

The *Database Management* module is used for linguistic information editing. The previous version of the system was developed concomitantly with linguistic data collection, which is why the database is not well organized. The new approach uses the existing data which will be exported in the *nalr* database. This work is in progress as a stand-alone Windows application. A version of the database management

module as Web application with client-server facilities is also considered. The first option is still preferable as, during editing, phonetical transcription is displayed as image, not as a string of characters, and a Windows control is anyway required for this. As mentioned in [1] and [2], the phonetic transcription of the Romanian language requires about 25,000 symbols, that are impossible to be designed as font typefaces, so that a graphical generator is used to generate the image of each character.

The *Maps Generator* module will be used to generate linguistic maps in an editable file format. The possibility to edit the phonetic transcriptions in database directly on the map is also considered. This facility may help linguists to analyze data end easily discover the geographic-related particularities of the language.

4. CONCLUSIONS AND FUTURE WORK

This paper describes a work in progress. The geodatabase was partially prepared using the NetSET Map platform [11]. For instance only the *Moldova şi Bucovina*, and the *Crişana* regions are detailed on the full map. The linguistic information included in the previous version of the system is analyzed, in order to more precisely define the structure of the external linguistic database.

The novelty of the proposed approach consists in using GIS facilities in linguistic maps generation. The final goal is to create a standard model for all Romanian linguistic atlases, allowing future quantitative analysis of geographic variables which, depending on the inquiry conditions, may be correlated to *diastratic* and *diaphasic* variables [4].

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

Authors contributions: All three authors had equal contribution in writing this paper. The first author had the idea of the GIS-based approach, organized information, designed and implemented the geospatial database, implemented the application; the second author is responsible for the MySql external database design and implementation, web based application implementation; the third author organized the information from the linguistic point of view.

A preliminary version of this research, entitled "*O abordare bazată pe Sisteme Informatice Geografice în cartografia lingvistică*" was presented at the *Anniversary Workshop: Written and Spoken Romanian Language in the Context of New Information Technologies. Achievements and Prospects*, Iaşi, Romania, http://iit.academiaromana-is.ro/RoNLP/program_RoNLP.pdf, March 24th, 2016.

REFERENCES

1. BEJINARIU S., APOPEI V., ROMAN M., *Mediu pentru editarea transcrierilor fonetice în Limba Română. Realizarea Atlasului Lingvistic Român pe Regiuni*, in TUFIŞ D., FILIP F. GH. (Eds.), *Limba Română în Societatea Informaţională – Societatea Cunoaşterii*, Expert Publishing House, 2003, 265–276.
2. BEJINARIU S., APOPEI V., DUMISTRĂCEL S., TEODORESCU H.-N., *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, 564–572.
3. BOTOSINEANU L., OLARIU F., BEJINARIU S., *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*, XXVIe Congrès International de Linguistique et de Philologie Romanes, 26, CILFR València, 2010, 456–457.
4. DUMISTRĂCEL S., HREAPCĂ D., *La variation linguistique sous la perspective de l'informatisation des atlas régionaux roumains*, RRL. 2008, **LII**(*1–2*), 87–103.
5. HOCH S., HAYES J.J., *Geolinguistics: The Incorporation of Geographic Information Systems and Science*, The Geographical Bulletin, 2010, **51**, 23–36.
6. LEE J., KRETZSCHMAR W., *Spatial Analysis of Linguistic Data with GIS Functions*, Int. J. Geographical Information Systems, 1993, **7** (6), 541–560.
7. LUCA R., BEJINARIU S., APOPEI V., *Electronic Linguistic Atlases. Client-Server Version*, Proceedings 9th International Symposium on Automatic Control and Computer Science, SACCS 2007, POLITEHNIUM Publishing House, 2007, 262–266.
8. OLARIU F.-T., OLARIU V., BEJINARIU S., APOPEI V., *Los atlas lingüísticos rumanos: entre manuscrito y formato electrónico*, Revista española de lingüística, Año nº 37, Fasc. 1, 2007, 215–246.
9. OLARIU F.-T., OLARIU V., CLIM M.-R., LUCA R., *La cartographie linguistique roumaine face à l'informatisation: quelques projets et résultats*, in Actes du XXVIIe Congrès international de linguistique et de philologie romanes, Nancy, 2013, 285–286.
10. XIE Y., ARISTAR-DRY H., ARISTAR A., LOCKWOOD H., THOMPSON J., PARKER D., COOL B., *Language and Location: Map Annotation Project – A GIS-Based Infrastructure for Linguistics Information Management*, Proceedings of the International Multiconference on Computer Science and Information Technology, 2009, **4**, 305–311.
11. ⁂, Datainvest, Manual de utilizare al pachetului de programe NetSET, 2008.
12. ⁂, *Open Geospatial Consortium, "Glossary"*, http://www.opengeospatial.org/ ogc/ glossary/g, last accessed on March 5th, 2015.

Received March 3, 2016