ISSN: 0975-766X
CODEN: IJPTFI
Available Online through
Research Article
www.ijptonline.com

ELECTRONIC ATLAS OF TATAR SUB-DIALECTS

Farid Salimov, Rustem Salimov
Kazan Federal University, Russia.
farid.salimov@kpfu.ru

Received on 14-08-2016 *Accepted on 20-09-2016*

Abstract

Experience in creation of electronic atlas of Tatar national sub-dialects is discussed. History of atlas creation is described, major principles that were laid in basis of program creation are discussed. Major issues of requests to database of electronic atlas are considered. Here is shown that information represented in atlas could be interpreted as subset of points of three-dimensional space. Axis of this space are, respectively, a set of inhabited communities in which were conducted sub-dialects researches, a set of meaning of language phenomenon, and set of atlas maps, each of the reflecting a certain language phenomenon. Any request realizes a selection of corresponding sub-collection of points of this space. Atlas is placed in Internet. URL address is http://atlas.antat.ru.

Keywords: Electronic atlas, Tatar dialects, Tatar sub-dialects, maps visualization, dialectometry.

Introduction

One of the most important aspects of language dialects research is a survey of dependence of language phenomena on their territorial positioning. Common geographic medium of residence forms certain stable connections between languages of nations inhabiting the given geographic space, often explains similarity or diversity of language phenomena that are related to different language groups, allows to understand a nature and rate of diverse processes that are going on in languages and dialects. Collection of information for fixation of corresponding language phenomena is usually conducted by preliminary formed unified program of research in frames of set geographic territory within a long time period. On basis of analysis of collected materials dialectological atlases of diverse languages are published. These atlases comprise multiplicity of maps that demonstrate expansion of certain language phenomena in limits of set territories. Also these atlases represent peculiar databases that are reflecting dependency of language phenomena on base inhabited communities in which the questioning was conducted. Similar atlases were published in XX century in the majority of European countries [1-3], in Russia the atlas of dialects and sub-dialects

of Russian language was published in three volumes [4-7]. First expeditions on material collection for dialects of Tatar language are dated by the beginning of 60s of XX century. Geographically these expeditions were covering a vast territory of Middle and Lower Volga, Ural and West Siberia that are representing major regions of Tatar population accommodation. Collection of information was conducted by special program composed in 1959 and comprising more than 200 questions on vocabulary, phonetics, morphology and syntax of Tatar language. Institute of language, literature and arts of Academy of Sciences of Republic of Tatarstan was chosen as a base organization for studies conduction. In expeditions on material collection for atlas was involved a large collective of dialectologists under the supervision of N. B. Burganova and L.T. Machmutova. At collection of information was performed a thorough analysis of material collected, its interpretation, registration. Distances between inhabited communities covered by research were not exceeding 10 – 15 km. A printed version of Tatar sub-dialects atlas was published in two volumes in 1989 [8-9]. In 2016 a second edition of atlas was issues, complemented by materials of Lower Volga and Siberia [10]. Similar atlases in different periods were published in Russia for Udmurt language [11], Bashkir language [12].

In structural aspect the atlas of Tatar sub-dialects consists of 215 maps of language phenomena, auxiliary maps and integral maps of isoglosses. Major signs of dialect difference, patterns and nature of expansion of dialect phenomena are reflected in maps. Every map in the atlas is accompanied by commentaries. Integral maps of isoglosses are represented by 17 maps in Atlas. Here, via isoglosses, are distinguished and generalized some language phenomena that have similar outlines.

Program description

Development of information technologies sets the task of creation of electronic databases with extended abilities of addressing to information sources and additional abilities of visualization and of distribution of properties of languages and dialects. Electronic database of atlas of sub-dialects of any language should comprise both cartographic part in which information about inhabited communities is preserved, and attributive part that comprises information about distribution of language signs by selected inhabited communities. At creation of electronic version of atlas developers were guided by principle of succession that was expressed in the fact that a language of linguistic facts description used in printed version of atlas was preserved to maximum extent. This concerns both methods of information presentation on maps and additional information comprised in comments to atlas. Particularly, in book version local numeration of inhabited communities by volumes was applied. At this numbers of different inhabited

communities in regions relating to different volumes matched. In course of creation of electronic atlas, developers did not introduce global numbering due to the fact that in published works on dialectology created on basis of atlas exist references to numbering accepted in book version. In order to facilitate the work with such literature, numbers of book version were complemented by prefix of volume number. As atlas was primarily designed as Internet-resource, convenient means of information presentation allowing performing click-throughs by hyperlinks from one objects comprised in atlas to other were developed. Furthermore, the ability of synchronous work on atlas maps and comments to them, ability of simultaneous viewing of several maps, a set of means of quick search of objects in maps with their further focusing were designed in. At this the user has the ability to conducts filtration of objects on atlas maps, selecting those among them that are needed for current work only. In course of atlas creation a particular attention was paid to request language: this language should be understandable for end user; library of requests should include major types of requests used by linguists at solution of certain tasks with using a book version of atlas. Composition of atlas includes 215 maps divided into four parts (vocabulary, phonetics, morphology and syntax) dependent on nature of language phenomena reflected by them. User is able to perform navigation on maps of atlas, received additional information from comments to maps and also information on every one of 1021 inhabited communities fixed in atlas base. At this the user can determine composition of objects for presentation in display himself. Figure 1 shows the variant of displaying of map No. 4 from chapter of phonetics with mode of showing of comment on. The comments can he hidden, if necessary. By user's desire can be shown a list of questions this maps answers to. Maps, questions and comments are connected in database by "many-to-many" connections. This allows conducting navigation through atlas, both through separate maps and maps answering same questions from different sides. The necessary map can by selected by its number in corresponding window. In original editions of atlas all inhabited communities are divided in volumes depending on their geographic position, in electronic version this division is supported by coloration.

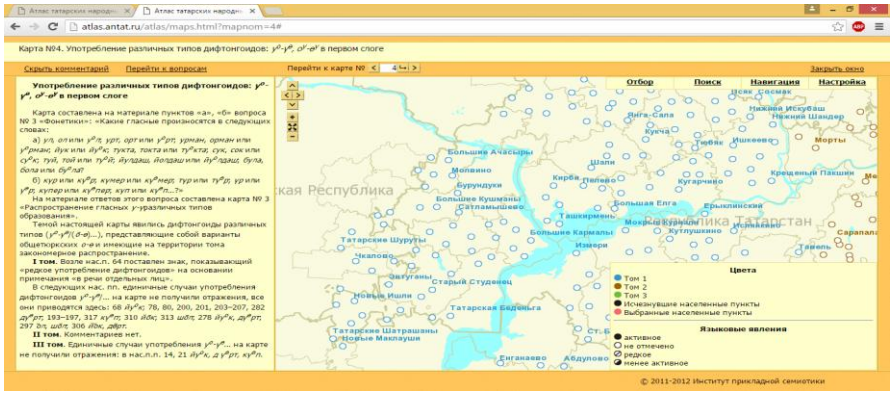

Fig. 1 Atlas page sample.

Every map of atlas is constructed of several layers, every one of which bears a separately applied water surfaces, administrative borders and their names. Layer diversity of information allows conducting filtration of objects and showing maps in form, convenient for end user. Each one inhabited community can have one of the following names:

- at present moment (in Russian or Tatar language),
- at moment of questionnaire (in Russian language, in Tatar language)
- identification number (matching the book version),

value of which on maps is selected by user.

Cartography information of atlas provides inclusion of inhabited communities geographic coordinates in database. Due to certain changes happened since time of expeditions' conduction in names and status of inhabited communities, some inhabited communities ceased to exist or changed their names, the work on reconstruction of inhabited communities list was conducted. Due to the fact that developers were not able to find the unified map with statement of coordinates of all inhabited communities included in composition of atlas and information on values of inhabited communities coordinates, they had to collect this information from different sources (mostly regional). Every inhabited community underwent identity check by existing databases; furthermore, additional information was collected in parallel via Internet use. One need to click on inhabited community of interest on atlas map to receive information about it. In displayed window is reflected information on name of inhabited community, its administrative identity, dialect observed in this inhabited community and also links to information in Internet.

In course of creation of attributive base for each one of 215 maps the information on distribution of language phenomena by base communities where researches were conducted had to be restored. Unfortunately, after release of book version of atlas in 1989, the set of records with this information was lost. In order to restore this information every one map in printed copy of atlas had to be analyzed, information on connection of studied language phenomena to inhabited communities had to be restored and added to the database. At this, due to density of labels placement on maps of atlas, at times occurred an ambiguous interpretation that was solved by expert board of linguists who participated in creation of printed atlas version.

Library of requests

Data storage in form of electronic database is especially useful for tasks of statistic analysis which is rather difficult to conduct by book version. Electronic atlas can be considered as an object defined in three-dimensional space, where every map being a plane bears information on one language phenomena in itself. Set of cards creates a volumetric

image, and requests by selected set allow calculating and comparing distributions of corresponding language phenomena, which could be useful at solution of tasks on clustering of sub-dialects. Such approach, creation of database of dialects, estimation of degree of similarity and difference between them, application of clustering algorithms for distinguishing of language areas, is popular in world dialectology in last years. Researches of this type were conducted for dialects of Bulgarian [13], Dutch [14] and other languages. The result of these works conduction was creation of a new scientific discipline, dialectometry [15].

Abilities of information presentation described above, being substantially requests to atlas database, serve for diverse options of data visualization. On the other side, while considering information put in database as basis, would be useful to receive additional analytical information by requests, to build new objects of base of existing ones. Library of requests that could be selected by user by display forms is built in program. Such requests are coming down to extraction of information from cartographic and attributive databases and presentation of this information in projections that are interesting for end user. All requests to database are divided into two large classes:

1. Requests that are executed on one map;

2. Requests that are executed on set of maps selected by user.

To the first class are mostly related requests of statistic nature executed in frames of one map. As all maps are constructed in a similar way, such requests could be applied to any one of 215 maps composing the content of atlas. Atlas database comprises information about administrative division of Russia with localization of inhabited communities on the level of subjects and districts. Using screen forms the user can set the necessary parameters and in that way conduct calculations in frames of territorial devision selected by him. A list of some requests related to the first class is shown below:

- Distribution of number of language phenomena variations in frames of one map in limits of selected territorial division (set volume, subject of RF or district, in limits of region around a certain inhabited community with a set radius);
- Comparative analysis of summary number of surveyed inhabited communities distribution (in limits of map, volume, subjects, districts);
- List of inhabited communities with a set value of language phenomenon in frames of selected territorial division;

- For selected language phenomenon and fixed inhabited community where this phenomenon is observed, the selection of minimal (maximal, average) distance to the nearest inhabited community with a set (different from a specified) language phenomenon;
- Calculation of maximal (average) distance in set of inhabited communities in which the same language phenomenon is observed;
- Calculation of maximal (average) distance in set of inhabited communities for selected pair of language phenomena;

Requests listed have mainly a statistic nature.

As can be seen from list shown, every request is parametrized and actually comprises a group of similar requests. It is anticipated that user will have the ability to independently conduct the division of a map into non-crossing regions and collect statistic in frames of selected division. For example, user can select (fix) a certain inhabited community and, by fixing selected coordinates as a coordinate origin, to build a rectangular net with a set step, and further to execute a request with grouping by building net. Requests of such type give into user hands the ability to research distribution of language phenomena in frames of division template selected by him.

A little aside stands the task of clustering (division of map into sites with determined properties, for example, comprising an inhabited community with a set language phenomenon) and building of isoglosses that presents a rather serious problem from the field of computational geometry by itself [16]. Methods of cluster analysis [17] and also algorithms of building of convex hulls could be used for solution of these tasks. However configuration of dispersion of language phenomena by map sites could be so complicated and fancy that it is anticipated to automatize only the first step in building of such division, and allow the user to adapt it in manual mode. At this the user can use a separate layer for application of additional signs on a map, which being layered on major map will give a necessary result. Experimental samples of integral maps where via isoglosses are distinguished and generalized language phenomena with similar outlining are included in atlas. It should be noted that manual composition of integral maps is a very complicated and labor-costing process. Automatization of the process of comparing maps that are describing extension of different language phenomena would allow simplifying the labor of linguists and accelerating the process of integral maps construction.

The second class of requests is composed of requests by set of maps:

- Construction of linguistic passport for predetermined inhabited community (set of inhabited communities) by all set of maps (by selected sub-collection of maps).
- Construction of new maps, in which distribution of certain aggregation of language phenomena by inhabited communities, is visualized, from basic maps presented in atlas.
- Comparison on language phenomena distribution for diverse maps.

Under linguistic passport of inhabited community (IC) we understand a set of all language phenomena attributed to this settlement. Construction of linguistic passport for IC with further comparison of these passports for different IC (IC groups) represents an interest for dialectologists. At this information could be grouped by vocabulary, phonetics, morphology and syntax. Request of this type are practically unrealizable in manual mode and presence of such mechanisms of information analysis represents a rather powerful instrument for diverse researches. Requests of the second type determine the reverse reflection of selected multiplicity of language phenomena on inhabited communities. At this a set of language phenomenon for which the research is conducted could be determined in form of logical formula that significantly expands analysis abilities. Third type of requests represents a comparative analysis of distributions of diverse language phenomena on the same territory. This is important for language phenomena from different positions describing similar parameters of dialects in language (for example, comparing of distribution of certain phenomena of phonetics and vocabulary). These requests can serve as basics for checking of diverse statistic hypotheses on basis of atlas materials.

Results

In frames of this project D.B. Ramazanova had conducted a work on construction of dialect segmentation of Tatar language that was absent in published book version of atlas. All information on dialects and sub-dialects was added to atlas database and can be received by user in form of report for corresponding request. However a large volume of initial data puts substantial limitations on accuracy of task solution: in course of manual analysis the dialectologist is able to cover only a limited volume of linguistic information and, respectively, results have an approximate nature. Mathematically, a task of dialect segmentation represents a task of clustering in linguistic space, points of which are sets of vectors characterizing language phenomena distribution. For points of this space occurs a task of selection of metric characterizing the degree of their similarity or diversity, with further application of clustering algorithms. A similar task for Russian sub-dialects has been solving in [18]. For Tatar atlas as conducted a preliminary analysis of clustering algorithms for limited territory (Republic of Tatarstan) with application of a certain set of metrics [19].

Comparative analysis with expert variant of division showed a difference in area limits description at sufficiently high per cent of matching of centers of such regions.

Conclusion

Atlas represents an interactive instrument for conduction of researches on dialects of Tatar language. Atlas is placed in Internet at address http://atlas.antat.ru. Programs of such type could serve as open platform for joint researches of Tatar dialect by a large number of scientists. Particularly, materials placed in atlas could be used, on one side, for teaching of Tatar language, on the other side could be applied for studying of language using methods of mathematic statistics.

Acknowledgements

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

1. Kirwin, William J. and G. M. Story. (1987). Linguistic atlas of Newfoundland: dialect questionnaire. Typescript, second ed. St. John's, NL: Memorial University of Newfoundland, Dept. of English Language and Literature.
2. Labov, W., Ash, S., and Boberg, C. (2006). Atlas of North American English: Phonetics, Phonology, and Sound Change. New York: Mouton de Gruyter.
3. Story, G. M. and William J. Kirwin. (1963). Linguistic atlas of Newfoundland dialect questionnaire. Ms, Memorial University of Newfoundland. [Update of original 5- page 1958 questionnaire, compiled by G.M. Story and P.D. Drysdale. Copy in the English Language Research Centre, Memorial University of Newfoundland (cf. Kirwin and Story 1987).]
4. Dialectological atlas of Russian language. Center of the European part of USSR. Issue I: Phonetics / Edited by R. I. Avanesov and S.V. Bromley. — M.: Science, 1986.
5. Dialectological atlas of Russian language. Center of the European part of USSR. Issue II: Morphology / Edited by S.V. Bromley. — M.: Science, 1989.
6. Dialectological atlas of Russian language. Center of the European part of Russia. Issue III: Maps (part 1). Vocabulary. — M.: Science, 1997.
7. Dialectological atlas of Russian language. Center of the European part of Russia. Issue III: Maps (part 2). Syntax. Vocabulary. — M.: Science, 2005.

8. Atlas of Tatar national sub-dialects of Middle Povolzhie and Cisurals. / Edited by N. B. Burganova, L.T. Machmutova, F. S. Bayazitova, D. B. Ramazanova, Z. R. Sadykova, T.H. Hayrutdinova: in 2 vol. - Kazan, Tatprokattechprobor, 1989 - 240 p.
9. Comments to Atlas of Tatar national sub-dialects of Middle Povolzhie and Cisurals. Kazan, Tatprokattechprobor, 1989 - 300 p.
10. Atlas of Tatar national sub-dialects. / resp. editor D. B. Ramazanova, T.H. Hairutdinova. 2nd edition, revised and reworked. – Kazan. ILLA, 2015. – 632 p.
11. Nasibullin R. Sh. Dialectological atlas of Udmurt language. Maps and comments. [Text] / R.Sh. Nasibullin, S.A. Maximov, V.G. Semenov, G.V. Otstavnova; Federal agency of education of RF; SEI HPE "Udmurt state university" etc. Izhevsk: RDE Regular and chaotic dynamic, 2009. Issue 1. 260 p.
12. Dialectological atlas of Bashkir language. — Ufa: Gilem, 2005. 234 p.
13. Prokić, J., J. Nerbonne, V. Zhobov, P. Osenova, K. Simov, T. Zastrow, and E. Hinrichs. 2009. "The Computational Analysis of Bulgarian Dialect Pronunciation". Serdica Journal of Computing 3 (3): 269-298.
14. Manni, F., Heeringa, W., and Nerbonne, J. (2006). To what extent are surnames words? Comparing geographic patterns of surname and dialect variation in the Netherlands. Literary and Linguistic Computing, 21(4): 507-27.
15. John Nerbonne and William Kretzschmar, Jr. Dialectometry HYPERLINK "http://urd.let.rug.nl/nerbonne/papers/Nerbonne-Kretzschmar-2012-Dec-1.pdf"++. *LLC: Journal of Digital Scholarship in the Humanities* 28(1), 2013, pp.2-12.
16. F.Preparata, M. Shamos, Computational Geometry. Springer-Verlag, New York, 1985
17. Everitt, B.S., Landau, S. and Leese, M. (2001), Cluster Analysis, Fourth edition, Arnold.
18. Pshenichnova N.N. Typology of Russian sub-dialects. M: Science. 1996. 208 p.
19. Polyakov V.N., Soloviev V.D. Computer models and methods intypology and comparativistics, Kazan: KFU, 2006. 208 p.