ISSN 1019-3316, Herald of the Russian Academy of Sciences, 2016, Vol. 86, No. 6, pp. 448–453. © Pleiades Publishing, Ltd., 2016.
Original Russian Text © I.I. Isaev, V.D. Solov'ev, F.I. Salimov, A.G. Pilyugin, V.R. Bairasheva, 2016, published in Vestnik Rossiiskoi Akademii Nauk, 2016, Vol. 86, No. 11, pp. 972–977.

Organization of Research

Creating a Database of Russian Dialects and Prospects for Dialectometric Studies

I. I. Isaev[a], V. D. Solov'ev[b], F. I. Salimov[b], A. G. Pilyugin[b], and V. R. Bairasheva[b]*

[a] *Vinogradov Russian Language Institute, Russian Academy of Sciences, Moscow, Russia*

[b] *Kazan Federal University, Kazan, Russia*

e-mail: ignatis@mail.ru; maki.solovyev@mail.ru; Farid.Salimov@kpfu.ru; pag@kch.ru; vbayrasheva@gmail.com

Received January 15, 2016

Abstract—The necessity to transfer from the paper-based *Dialectological Atlas of the Russian Language,* published 30 years ago, to an electronic database is substantiated. This would make the information contained in the atlas available to a large number of users and ensure on-the-fly information retrieval, isogloss plotting, and finding dialects with a specified set of properties. It would also be possible to use state-of-the-art analytical tools for dialectological data, including mathematical methods, particularly cluster algorithms and multidimensional scaling, widely used in dialectometric studies in the West.

Keywords: Russian dialects and subdialects, database of Russian subdialects, Dialectological Atlas of the Russian Language, dialectometry.

DOI: 10.1134/S1019331616060022

Attempts to classify Russian dialects were made by the Imperial Academy of Sciences as early as the end of the 19th century. Systematic collection of materials according to a single program allowed the Moscow Dialectological Commission to publish *An Attempt at a Dialectological Map of the Russian Language with the Addition of an Essay on Russian Dialectology* in 1915 [1]. In the opinion of linguists of that time, the Russian language was represented by three main dialects: Great Russian, Little Russian, and Belorussian, which, in turn, subdivided into smaller groups of subdialects. This concept of the structure of the Russian language was corrected, and the idea of a Russian dialectological atlas appeared in the 1930s. The data on Russian subdialects was collected under the supervision of R.I. Avanesov.

The Dialectological Atlas of the Russian Language (*DARL*) [2–5] is a product of field and analytical work of dialectologists of the center of the European part of Russia (the Soviet Union), which started in 1938 and was renewed after WWII according to The Program of Collecting Data for the Compilation of a Dialectological Atlas of the Russian Language [6]. The program contained 294 questions that helped acquire information on the phonetics, morphology, and, to a lesser degree, lexicon and syntax of the subdialect under study (we use the terms *subdialect* and *dialect* as synonyms). The expeditionary task force collected material in more than 4200 settlements on the surveyed territory. Data collection and linguistic analysis took 50 years.

The territory mapped in *DARL* (Fig. 1) is limited to the center of the European part of Russia, because there the Russian language proper started to form after the disintegration of the Old Russian language union by the 14th–15th centuries. On the rest of the territory of Russia, Russian language dialects genealogically related to the center of the European part prevailed.

The work on the atlas maps has led to the formation of the Moscow school of linguistic geography, the principle concept of which is presented in the monograph *Problems of the Theory of Linguistic Geography* [7].

The *DARL* maps (Fig. 2) contain fundamental lingual–geographical information. More accurate and voluminous data on the geographical representation of Russian dialects of the center of European Russia do not exist.

THE GOALS OF CREATING A DATABASE OF RUSSIAN LANGUAGE DIALECTS

Kazan (Volga) Federal University and the Vinogradov Russian Language Institute, RAS, have imple-

* Igor' Igorevich Isaev, Cand. Sci. (Philol.), is a Senior Researcher at the Vinogradov Russian Language Institute (RLI), RAS. Valerii Dmitrievich Solov'ev, Dr. Sci. (Phys.–Math.), is a Leading Researcher at Kazan Federal University (KFU). Farid Ibragimovich Salimov, Cand. Sci. (Phys.–Math.), is an Associate Professor at KFU. Aleksandr Gennad'evich Pilyugin is an Assistant Professor at KFU. Venera Rustamovna Bairasheva, Cand. Sci. (Phys.–Math.), is an Associate Professor at KFU.)

Fig. 1. The territory of the subdialects described in *DARL* [8].

mented the project of a computer database of Russian language dialects, information for which is retrieved from the *DARL* paper maps. The database has a relational structure; i.e., the data are represented as a table: a settlement and a set of values of linguistic features for this settlement. Previously, a similar database was created for dialects of the Tatar language.

Analysis of the digital maps of individual linguistic phenomena permits solving several problems. One of them is the formation of a dialectal area by map layering. This program function is necessary to specify the boundaries of a dialectal array for working out the route of an expedition of dialectologists.

The drawback of this method is the complexity of the map derived. Aligned with internal lacunae, areas are readable only at the two-layer map level; three or more layers make it unreadable. To overcome the complexity of the dialectal picture, an isogloss could be used, i.e., a line on a geographical map that limits the territory of an individual linguistic phenomenon. However, an isogloss is the result of a cartographic dialectologist's analytical work. The linguistic goal of the project is to present a precise picture of the dialectal landscape of the center of the European part of Russia for lingual–geographical and field studies of Russian dialects.

The second objective is the choice of a settlement from the *DARL* list for expeditionary (re)survey. The technical assignment of a dialectologist who sets out for an expedition is to collect information from the atlas to characterize a group of neighboring settlements for further study of the language structure. In the final variant, the data about a dialect should look like a list of dialectal features (structural properties) marked on the atlas maps, i.e., as a minimal linguistic description of a dialect.

The most important objective of the project is to grant access to the electronic maps of the dialectological atlas during the educational process. The university course Russian Dialectology cannot do without map illustrations, and the paper atlas format prohibits illustrating the geography of a phenomenon under study using multimedia teaching aids.

Popularization of scientific information about the organization of the national Russian language in the context of territorial dialects and exposure to the history of linguistic phenomena, primarily, in the fields of phonetics and morphology, are impossible without the *DARL* maps. They can at last be included into articles on Russian dialectology, history, and ethnography.

DIALECTOMETRY

No less important is the fact that a database of feature values helps in study of the degree of resemblance of dialects (and languages) using rigorous mathematical methods. In the first place, it is possible to define the distance between dialects. If whole words belong to feature values, the most popular is the Levenshtein edit distance [9], which takes into account the phonetic similarity of words. If feature values are more abstract, as, for example, in the World Atlas of Language Structures (WALS) [10] and Languages of the World [11] typological databases, the distance between two languages is measured by the number of features that take different values in these languages. Usually, the number is divided into the total number of features (normalized). This is the so-called city metrics. The formula for distance calculation can be more complex to take into account the degree of similarity of feature values [12].

After the distance has been determined, a matrix of distances is built between any two dialects from the database. Then, various mathematical algorithms can be applied to the matrix—cluster analysis [13], multidimensional scaling [14], etc.—which help break the entire multitude of dialects studied into similar groups. Hierarchical clustering algorithms additionally order the groups of dialects hierarchically. This approach is used increasingly often in historical linguistics to study the evolution of languages. Monograph [15] gives a detailed overview of the methods used and the results obtained in this field.

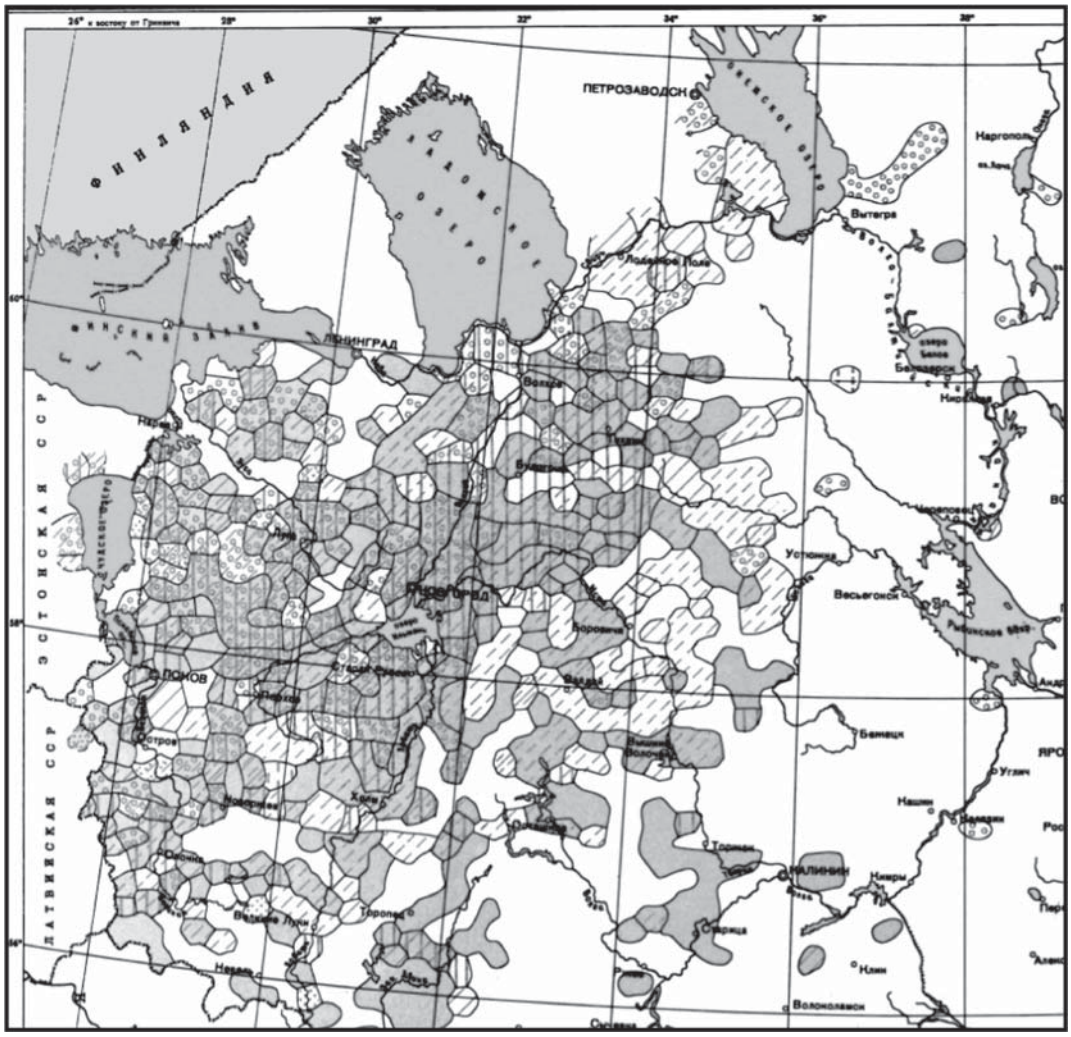

Fig. 2. A fragment of map 66, *DARL*, vol. 1 [2].

Mathematical methods help not only build dialect clusters but also solve many other problems. N.N. Pshenichnova mentions the following problems that are topical today: the identification of features that are most significant in classification for each cluster of dialects; the clarification of notions such as *group of dialects, transitional dialects, a mixed body of heterogeneous dialects,* and *a dialectal type*; the description of *typicality patterns* in each cluster of dialects; and the description of the structure of the Russian dialectal area in terms of fuzzy logic [12]. Pshenichnova [27] suggested solutions to these problems; however, since the computational power of computers at that time was insufficient for the use of exact clustering algorithms, approximate algorithms were used instead. Significant progress has been made in the development of clustering algorithms since then, in particular those aimed at work with large data arrays, making it possible to count on the clarification of previously obtained results.

The creation of a database of dialects, the calculation of distances between them, and the use of clustering algorithms constitute an approach, popular in foreign dialectology in recent years, that has been termed *dialectometry.* Dialectometric studies have been done for the dialects of the Bulgarian [16], Dutch [17], and other languages. An overview of the current status of dialectometry was given in [18].

A database described in [16] contained 197 dialects of the Bulgarian language by 156 features. Concepts serve as features, and words that verbalize these concepts in specific phonological realizations act as feature values. The database is based on the Archive of the Ideographic Dialect Dictionary of Bulgarian, which was created at the University of Sofia in the 1950s under the supervision of S. Stojkov. The Levenshtein metrics is used as distance.

The clustering algorithm of the weighted pair group method using arithmetic averages (WPGMA) is used in the Bulgarian database [19]. The algorithm divides

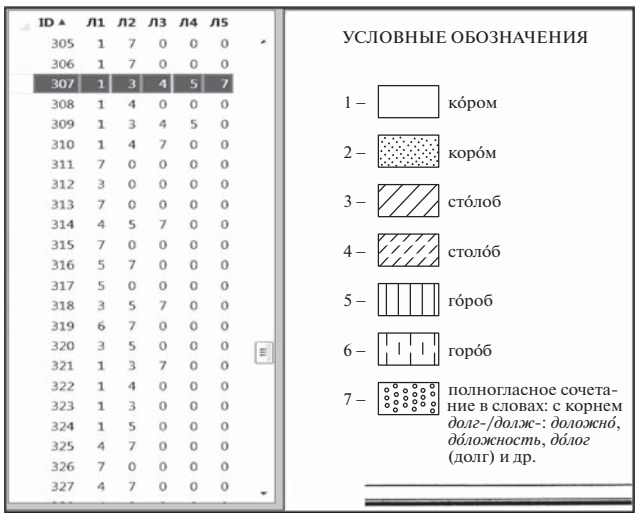

Fig. 3. A fragment of the attributive database.
The ID column contains the number of the settlement; the next columns contain the numbers of attributes that correspond to this settlement. On the right is the legend of the map.

all dialects into two clusters at the upper clustering level. The traditional classification is given in [20]. Both classifications lead to a similar division of dialects, but the difference is still very substantial. This yields additional material for the study of intermediate dialects and for the clarification of the boundary.

What are the reasons for such differences between the traditional and quantitative classifications? First of all, a difference is possible in the initial data. The database of Bulgarian dialects included, for example, only lexical–phonetic data, while the traditional classification can also use morphological and syntactic information. Adequate comparison is valid only when the initial data coincide. *DARL* gives a classification of dialects built by the traditional method based on features represented in the atlas. Therefore, having built a database according to *DARL* and having used the methods of dialectometry, it is possible to compare adequately the results of the traditional and quantitative classifications. Moreover, humans are unable to keep in mind and take into account simultaneously multiple data: the values of 156 features for Bulgarian and 4400 features for Russian. Eventually, the classification is built on a small number of features that are viewed as the most important. Thus, the division of Bulgarian dialects into western and eastern was based on a single feature: the variants of realization of the
2 Bulgarian phoneme *ě (jat). Unfortunately, the importance of features is usually not revealed in a clear numerical form. The methods of cluster analysis make it possible to calculate automatically the coefficients of feature importance. Further on, they can become the subject of analysis and clarification for dialectologists. Thus, the use of traditional and quantitative methods contributes to the clarification of classifications and a clearer understanding of the structure of the dialectal and feature areas.

A CARTOGRAPHIC DATABASE

As was noted above, the main objective of linguistic geography is to study the dependence of linguistic phenomena on their territorial distribution. An electronic database of the dialectological atlas of any language is natural to build so that it includes both the cartographic part that stores information about settlements and the attributive one that contains data about the distribution of linguistic features by selected settlements.

The cartographic part includes data about reference settlements where information is collected, their geographical coordinates, administrative subordination, and additional information on history and ethnography, as well as on the national and quantitative composition of the inhabitants.

The cartographic part of the electronic database requires the exact location of each settlement on the map. Unfortunately, appendices to an atlas usually contain only the names of surveyed settlements and information on their administrative subordination without their geographical coordinates. The problem is that the maps of dialectological atlases published in the 1950s and 1980s represented schematic maps on which each settlement was plotted approximately (for the atlases of the Tatar [21, 22], Bashkir [23], and Udmurt languages [24]) or was not plotted at all (for *DARL* [2–5]). Hardline censorship on publishing such information existed in the country at that time; in addition, the technical capabilities of publishing houses, as a rule, left much to be desired.

The conversion of the atlas into an electronic form requires reconstruction of the list of settlements with an indication of their geographical coordinates. The list cannot be compiled by simple revision of current maps. Some settlements ceased to exist; others changed their names or administrative subordination. Several settlements have the same name even within one administrative district, for example, Volkovo is encountered 117 times in the Bank of Cities database (http://www.bankgorodov.ru/): ten times in Tver oblast and nine times in Moscow oblast. It is clear that the exact location of a settlement is a fairly difficult task. Moreover, projects implemented on Russian territory are characterized by very voluminous lists of reference settlements. For example, the list for *DARL* has 4206 names, and that for the atlas of Tatar popular dialects, 1031 dialects, which is explained by the vast territories of dialect distribution and significant migratory processes that occurred there.

Electronic databases, compared with their book analogs, have greater opportunities to represent information on settlements in which questionnaires are conducted. Such databases may be seen as distributed if links are provided to information stored on various Internet resources (e.g., http://www.bankgorodov.ru/, https://ru.wikipedia.org/, http://wikimapia.org/). The links may indicate historical, ethnographical, and linguistic materials in various databases and be of interest for researchers in certain dialects and languages. A very good resource on the history and ethnography of Bryansk oblast is located on the web site http://www.kray32.ru; some interesting facts about settlements of Arkhangel'sk oblast can be found at http://www.russia29.ru/; and those of Vladimir oblast, at http://vladimirskaya-rus.ru/. It is clear that the creation of such resources is a troublesome business, associated with large expenses, but dialectological atlases can largely justify these processes.

AN ATTRIBUTIVE DATABASE

The main stage of creating a database is the correlation of feature values represented in *DARL* with each settlement. In most cases, feature values in dialectological maps are linked to settlements; however, this is different in *DARL*: it only has areas drawn with the same feature.

We have developed and implemented a transformation of graphic paper maps into a database format. Each *DARL* map is processed separately; at the first stage, it is scanned; at the second stage, it is linked to a contemporary digital map of Russia. In addition, a problem arises for the reason that projections used to create the *DARL* maps are not described anywhere and are different for different maps. A special computer method, based on the identification of a *DARL* map and maps of several contemporary settlements (reference points), transforms *DARL* maps. Then, a standard algorithm of projection conversion is used, and later on it is possible to impose a *DARL* map on a contemporary map without distortions. Then, the boundaries of oblasts with the same feature values are transferred in a semiautomatic mode using the EasyTrace software package (http://www.easytrace.com). A set of code values that corresponds to the feature values from the map legend is assigned to each oblast, after which an attributive database of settlements is formed, containing information about all features assigned to this settlement in all maps.

A fragment of an attributive database is given in Fig. 3. At present, the creation of a database for the first *DARL* issue (phonetics) has been completed [2]. Later on, it is planned to convert the remaining *DARL* issues into the database format; however, since phonetic features are described only in the first issue, the database can already be used for phonetic studies of Russian dialects. The databases created and other materials are published on the project site, the implementation of which will create a technological basis for deeper study of Russian dialects, as well as their classification and evolution.

ACKNOWLEDGMENTS

This work was supported by the Russian Science Foundation for the Humanities, grant no. 15-04-12008 v.

REFERENCES

1. *An Attempt at a Dialectological Map of the Russian Language with the Addition of an Essay on Russian Dialectology* (Moscow, 1915) [in Russian].
2. *The Dialectological Atlas of the Russian Language. The Center of the European Part of the USSR,* Vol. 1: *Phonetics* (Nauka, Moscow, 1986) [in Russian].
3. *The Dialectological Atlas of the Russian Language. The Center of the European Part of the USSR,* Vol. 2: *Morphology* (Nauka, Moscow, 1989) [in Russian].
4. *The Dialectological Atlas of the Russian Language. The Center of the European Part of the USSR,* Vol. 3: *Maps* (part 1). *Lexicon* (Nauka, Moscow, 1997) [in Russian].
5. *The Dialectological Atlas of the Russian Language. The Center of the European Part of the USSR,* Vol. 3: *Maps* (part 2). *Syntax. Lexicon* (Nauka, Moscow, 2005) [in Russian].
6. *The Program of Collecting Data for the Compilation of the Dialectological Atlas of the Russian Language* (Moscow, 1947) [in Russian].
7. *Problems of the Theory of Linguistic Geography* (Izd. Akad. Nauk SSSR, Moscow, 1962) [in Russian].
8. http://www.gramota.ru/book/village/index.html.
9. D. Gusfield, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology,* 1st ed. (Cambridge Univ. Press, 1997).
10. *The World Atlas of Language Structures Online* Ed. by D. Matthew and M. Haspelmath (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013). http://wals.info. Cited February 2, 2015.
11. V. N. Polyakov and V. D. Solov'ev, *Computer Models and Methods in Typology and Comparativistics* (Kazan' Gos. Univ., Kazan, 2006) [in Russian].
12. N. N. Pshenichnova, *Typology of Russian Subdiaects* (Nauka, Moscow, 1996) [in Russian].
13. M. S. Aldenderfer and R. K. Blashfield, *Cluster Analysis* (Sage, 1984).
14. Yu. N. Tolstova, *Basics of Multidimensional Scaling* (KDU, Moscow, 2006) [in Russian].
15. J. Nichols and T. Warnow, "Tutorial on computational linguistic phylogeny," Linguistics Language Compass **2** (5), 760–820 (2008).
16. P. Houtzagers, J. Nerbonne, and J. Prokić, "Quantitative and traditional classifications of Bulgarian dialects compared," Scando-Slavica **56** (2), 163–188 (2010).

17. J. Nerbonne and P. Kleiweg, "Lexical distance in LAMSAS," Computers Humanities **37** (3), 339–357 (2003).

18. J. Nerbonne and W. Kretzschmar, "Dialectometry++," J. Digital Scholarship Humanities **28** (1), 2–12 (2013).

19. P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy: The Principles and Practice of Numerical Classification* (Freeman, San Francisco, 1973).

20. S. Stojkov, *Bălgarska dialektologija,* 2nd ed. (Nauka i Izkustvo, Sofia, 1968).

21. *The Atlas of Tatar Popular Subdialects of the Middle Volga and Cis-Ural Regions,* in 2 vols., Ed. by N. B. Burganova, L. T. Makhmutova, F. S. Bayazitova, D. B. Ramazanova, Z. R. Sadykova, and T. Kh. Khairutdinova (Tatprokattekhpribor, Kazan, 1989) [in Russian].

22. *Commentaries on the Atlas of Tatar Popular Subdialects of the Middle Volga and Cis-Ural Regions* (Tatprokattekhpribor, Kazan, 1989) [in Russian].

23. *A Dialectological Atlas of the Bashkir Language,* Ed. by F. G. Khisametdinova (Ufa, 2005) [in Russian].

24. R. Sh. Nasibullin, S. A. Maksimov, V. G. Semenov, and G. V. Otstavnova, *The Dialectological Atlas of the Udmurt Language: Maps and Commentaries,* in 2 vols. (NITs Regulyarnaya i Khaotichnaya Dinamika, Izhevsk, 2009) [in Russian].

Translated by B. Alekseev

SPELL: 1. isogloss, 2. jat; ПЕРЕВОД РИСУНКОВ