DOI: 10.30842/ielcp230690152723

N. G. Gorlov
Institute for Linguistic Studies of the Russian Academy of Sciences. SPb.
nikita@iling.spb.ru
M. S. Morozova
Institute for Linguistic Studies of the Russian Academy of Sciences. SPb.
morozovamaria86@gmail.com
A. N. Sobolev
Institute for Linguistic Studies of the Russian Academy of Sciences. SPb.
sobolev@staff.uni-marburg.de

ETHNOLINGUISTIC GROUPS OF SOUTHEASTERN EUROPE: WAYS OF PRESENTATION

The experimental article discusses the problem of uniform analysis, digitalization and proportional presentation (visualization) of information available in scientific publications about modern ethnolinguistic groups in Southeastern Europe (Balkan and Carpathian-Danube areas). Based on the results of a systematic analysis and digitization of data sources, 200 ethnolinguistic groups and 50 linguistic varieties (languages) were found in the region. In a specially developed web application, digital methods for presenting these groups and languages in the form of interactive electronic graphs and geographical maps were selected and practically applied. The problems of providing comprehensive lists of ethnolinguistic groups and languages are discussed. The prospects for the development of quantitatively substantiated approaches to the compilation of representative *linguistic samples* for the region are outlined.

Keywords: ethnic groups of Southeastern Europe, languages of Southeastern Europe, Balkan Sprachbund, linguistic samples, visualization

Acknowledgements: This work has been carried out within the framework of the "Atlas of the Balkan Linguistic Area" project supported by the French National Research Agency (ANR-21-CE27-0020-ABLA) granted to Evangelia Adamou and the Russian Science Foundation (22-48-09003) granted to Andrey Sobolev.

The authors wish to express their gratitude to Professor Dr. Biljana Sikimić of the Balkanology Institute at SANU in Belgrade, Serbia, for providing valuable critical feedback on the initial version of this article.

Disclaimer: The authors disclaim any responsibility for errors or omissions in the political maps' content. The information presented in the maps regarding political borders, including partially recognized borders, is accurate as of 1.1.2014.

Н. Г. Горлов
ИЛИ РАН, СПб, Россия. nikita@iling.spb.ru
М. С. Морозова
ИЛИ РАН, СПб, Россия. morozovamaria86@gmail.com
А. Н. Соболев
ИЛИ РАН, СПб, Россия. sobolev@staff.uni-marburg.de

Этнолингвистические группы Юго-Восточной Европы: способы презентации

В экспериментальной статье обсуждается проблема единообразного анализа, цифровизации и пропорциональной презентации (визуализации) имеющихся в научных изданиях сведений о современных этнолингвистических группах Юго-Восточной Европы (Балканского и Карпатско-Дунайского ареалов). По итогам системного анализа и перевода источников материала в цифровой формат в регионе установлено проживание 200 этнолингвистических групп и бытование 50 идиомов (языков). В специально разработанном веб-приложении произведен отбор и практически применены цифровые методы презентации этих групп и языков в виде интерактивных электронных графиков и географических карт. Обсуждены проблемы инвентаризации этнолингвистических групп и языков. Намечены перспективы разработки количественно обоснованных подходов к составлению репрезентативных *лингвистических выборок* по региону.

Ключевые слова: языки Юго-Восточной Европы, этнические группы Юго-Восточной Европы, балканский языковой союз, лингвистические выборки, визуализация

Introduction

This article examines potential methodologies and preliminary findings for the uniform analysis, digitalization, and proportional presentation (i.e. visualization) of data relating to modern ethnolinguistic groups in Southeastern Europe, specifically the Balkan and Danube-Carpathian areas. The data, which are available in authoritative scientific reference publications, currently exist only in analog formats, such as reference texts and numerous geographical maps, and are quantitative, spatial and qualitative in nature. In addition to fulfilling the methodological, practical, and informational objectives of the study, this experimental work also outlines initial approaches for developing quantitatively substantiated approaches of compiling representative *linguistic samples*, *specifically areal-linguistic (linguogeographic) samples*, for the region under investigation. Such samples are essential and commonly used, either explicitly or by default, in the creation of

linguistic atlases. It should be noted that this paper's scope is limited to exploring only a few approaches and techniques, without offering a comprehensive and fully systematic and structured presentation of information. Furthermore, the primary sources were not cross-checked during this study, nor were any additional sources used by the authors.

The article's first section addresses the problem of presenting and visualizing information concerning the ethnolinguistic groups of Southeastern Europe in the digital era and sets the task required to resolve this issue. The second section discusses the sources of information available for presenting and digitizing this data, while the third section concentrates on the selection and practical application of digital presentation and visualization techniques, implemented through a specially designed web application. The conclusion of the article summarizes key findings and identifies future research opportunities, including the discussion of potential problems and perspectives.

The Problem of Presenting Information on Ethnolinguistic Groups

The linguistic samples forming the basis of the lists of locations for cross-border linguistic atlases covering the region under consideration and not limited to the data about one language family, such as Atlas Linguarum Europae (Alinei et al. 1975), Atlante linguistico mediterraneo (Deanović 1964), the General Carpathian dialectological atlas (Bernshtejn et al. 1987), the Small Dialectological Atlas of the Balkan Languages (Sobolev 2003), the World Atlas of Linguistic Structures (Dryer, Haspelmath 2020), and the less-known latest Mouton Atlas of Languages and Cultures (Carling 2019), do not reflect the complete local linguistic and ethnolinguistic diversity of the region. Moreover, they disproportionately represent the linguistic varieties (languages, dialects, and idioms) in terms of the number of speakers, not to mention the varying degrees of linguistic differentiation. Regarding the latter, we confine ourselves to paying attention to WALS (Gil 2020), which allows synchronous mapping at the same taxonomic level of data for several dialects of the German language, designated as German (Mansfeldisch), German (Ripuarian), etc., for the German standard language (linguonym German) and for several particular idioethnic languages of Eastern Europe (with linguonyms Albanian, Bulgarian, Romanian, Russian, etc.). Reference cartographic editions, such as Asher and Moseley's (2007), present arbitrary depictions of linguistic reality. For

example, Modern Greek is placed in the Western European area (Lachlan 2007), while the other Balkan languages and idioms are attributed to the languages of North Asia and Eastern Europe (Comrie 2007). The Atlas Linguarum Europae, as known, maps the territories of nation-states as a monolingual space of the language of the titular ethnic groups, ignoring the numerous idioms of minorities (see, for example, the set of data points in the Hellenic Republic).

The inadequate and biased representation of ethnolinguistic groups, languages, and dialects in Southeastern Europe cannot solely be attributed to linguistic atlases. Rather, it is reflective of the general state of affairs in the field of European history and ethnography, Balkan areal-typological linguistics, and world linguistics in general. Recent well-documented work (Demeter, Bottlik 2021) convincingly demonstrates that historically, the mapping of ethnic groups in Southeastern Europe has always reflected not the objective state of affairs in specific regions, but rather the political intentions of nationally, if not nationalistically oriented geographers, including staunch national socialists (Krallert 1941). Although nowadays the replacement of archaic national biases with politically correct, minority-oriented ones (Magoczi 2018; Kamusella 2021) has added important details to the overall picture, it does not solve the general problem. The main issue lies in the technical limitations of analogue printing, which prevent ethnographic and generalizing linguistic maps, including those that are politically neutral (Straka 1979), from including and objectively, visually, and proportionally presenting all relevant information. Thus, such maps are incomplete by definition.

Despite the fact that achieving ideal linguistic sampling, which includes compiling representative lists of locations for linguistic atlases, was unattainable in the past and is unlikely to be feasible in the near future, given the current state of affairs in Balkan linguistics, the goal of achieving a sufficient, maximally comprehensive and proportional, as well as reliable representation of the diversity of linguistic (and relevant ethnographic) facts, taking into account the different degrees of dialect differentiation on the ground, raises questions about the methods of compilation and digitalization, as well as ways to visualize the results in the new international project, the Atlas of the Balkan Linguistic Area (Adamou, Sobolev (Eds.) 2023)[1]. Is it possible to find ways to correlate the available unstable politico-geographical (Dami 1976),

[1] See also the developments of the Institute of Linguistics of the Russian Academy of Sciences: https://minlang.iling-ran.ru.

often estimative quantitative ethnographic, very fragmented demographic and ethnographic, and only partially available sociolinguistic information (such as "on the territory of states X, Y there is a number of N representatives of the ethnic group Z") with the qualitative linguistic one (like "on the territory of Ω there are linguistic varieties α, β, γ..."), to visualize it on linguistic and ethnolinguistic maps, and use it to create representative linguistic samples for further in-depth research? Until these tasks are resolved, many far-reaching generalizations can be considered premature, particularly the quest for regular relationships between ethnographic, sociolinguistic, and proper linguistic information (Scherbakova et al. 2023), at least for Southeastern Europe, which still retains some of the characteristics of terra incognita in the 21st century.

The first move towards achieving this ambitious goal can be made by addressing two challenges: compiling a comprehensive inventory of available reliable information and visualizing it in digital format. To address the first task, we draw on the material from two academic sources (Jordan 2006; Kahl 2014), while for the second task, we compile diagrams and graphs and use mapping tools.

It is important to note that in this article, we are not attempting to solve the ontological problem of enumerating and inventorying all the ethnolinguistic groups of Southeastern Europe (we leave this task to geography, demography, ethnography, descriptive linguistics, and sociolinguistics). Instead, we aim to address the methodological and applied technological problem of creating a convenient and interactive visualization of the results of such an enumeration and listing, with the aim of further creation of representative linguistic samples for the region.

Compiling and Digitizing Information on Ethnolinguistic Groups

From a methodological perspective, the following steps are distinguished as important for solving the task. Firstly, it is essential to ensure that the sample is comprehensive in terms of including all the ethnolinguistic groups of Southeastern Europe that are known to science. If we compile even a representative sample of only those groups currently of interest to a certain circle of scholars[2], we may

[2] This approach is often used in linguistic atlases and databases. Cf. the WALS sample, which includes a total of 2,662 languages and dialects, but at the same time represents a selected set of language types that reflect the behavior of individual interesting parameters of interlingual variation.

find that the observed distributions are far from fully corresponding to reality. The region under study poses a particular challenge in presenting certain groups, such as Aromanians or Roma, where even the approximate size of the group is difficult to determine. Secondly, an approach to quantitatively assess the number of ethnolinguistic groups must be developed, which adequately considers both relatively precise data and rough estimates.

The authoritative "Atlas of Eastern and Southeastern Europe" (Jordan 2007) was the primary source of data for our study. This atlas contains information about the ethnic composition of the population in countries belonging to these regions, approximately at the end of the 20th and beginning of the 21st century. The compilers of the Atlas used first of all the official population censuses conducted in 2001–2004 as the primary sources of information. Occasionally, the authors of sections on individual countries also provided other quantitative data from literature on ethnography and anthropogeography. For Kosovo, information on the ethnic groups was based on estimates from UNHCR, OSCE, and KFOR as of 31.8.1999. One of the challenges we faced while working with the Atlas was to compile a comprehensive list of ethnolinguistic groups in Bosnia and Herzegovina. We had to rely on incomplete estimates for the territory of the Federation of Bosnia and Herzegovina as of 31.12.2003. In the future, we aim to consider the ethnolinguistic composition of Republika Srpska, as well as the number of speakers of Romani, Sephardic, Italian, German, Ukrainian, Czech and other languages throughout the country once reliable information becomes available.

As the Atlas does not include data for the Turkish and Greek Republics, and does not always provide information on the smallest ethnic groups, which are often classified as "Other" in the results of population censuses, we used an additional source — the publication by Thede Kahl, one of the compilers of the Atlas (Kahl 2014). This publication attempts to provide a comprehensive list of the idioethnic languages and corresponding populations of Southeastern Europe. Quantitative estimates are often approximate (e.g., the number of Italian speakers is "several tens of thousands") or absent (e.g., Arabic-speaking Muslims in Cyprus, Turkey, Greece, and Romania).

It is important to note that using an additional source did not allow us to compile an exhaustive list of ethnolinguistic groups residing in Southeastern Europe. Our sources, as well as censuses and other publications on population composition that served as

primary sources for them, do not take into account a range of groups that are officially recognized or unrecognized as ethnic (national) minorities. For instance, Croats (Letnica, Janjevo) and Bosniaks of Kosovo (not mentioned in our data), Bunjevci in Hungary (not yet recognized as a minority), Bulgarians in North Macedonia (usually from Eastern Serbia, where they are recognized as a minority), Bulgarians in Hungary ("Bashtovans", a recognized minority), [Balkan] Egyptians in southern Serbia and Montenegro, and many others. Therefore, it will be necessary to search and select a larger number of sources in the future to achieve the goal of comprehensively and reliably reflecting ethnic and linguistic diversity.

The data extracted from both sources were collated in an Excel spreadsheet. In accordance with the methodology adopted by the sources, ethnolinguistic groups were categorized based on the political-administrative principle, i.e. by the countries where they were registered or mentioned. For their linguistic varieties, language families and language groups were also indicated. However, this approach did not allow us to consider some of the quantitative estimates given in our sources without distribution by country, such as the total number of Aromanian speakers, which is estimated at 400,000 in (Kahl 2014: 97).

The German ethnonyms and linguonyms found in the sources, as well as the self-denominations of groups and their varieties (if mentioned), were recorded in separate columns of the table. English translations were identified for the German terms, and were used for captions on graphs and maps during subsequent data visualization experiments.

The resulting table contains information on 200 ethnolinguistic groups and 50 linguistic varieties (idioethnic languages) in Southeastern Europe. At this stage of work, in case of any discrepancies between the data of our sources and information from others, we relied on the estimates and terminology of the former. For instance, separate positions in the list of groups are given to "Romanians (Timok Vlachs)" in Serbia and Bulgaria, "Romanians" of Serbian Banat, "Aromanians", etc. — which corresponds to Germ. "Rumänen" ("Timok-Vlachen"), "Rumänen", "Aromunen" mentioned in our two sources.

As separate "languages", the speakers of which constitute a specific ethnic group, we distinguished, for example: the linguistic varieties spoken by the two historical Albanian diasporas — the Arvanites of Greece and the Arbëresh of Italy (along with the Albanian language); Cretan and Cypriot dialects of the Greek

language, Greek varieties spoken by Pontic Greeks, Romaniotes and Istanbul Karaites, Karakachans, Valahads, Tsakonian Greeks (along with the Greek language); Gorani language and Pomak variety; Italian and Istro-Romance (Istriot); (Daco)-Romanian (together with Moldavian), Aromanian, Megleno-Romanian and Istro-Romanian. Conversely, the umbrella term "BCMS" (Bosnian, Croatian, Montenegrin, Serbian) is used for the "language" spoken by multiple groups, including Serbs, Croats, Bosniaks, Montenegrins, "Muslims"[3], Šokci, Bunjevci, and Burgenland Croats, as identified by our sources and their primary sources. While it would be preferable to clearly distinguish and evaluate the number of speakers of specific languages or language variants by country, this is often impossible due to the lack of reliable information in sources about the linguistic affiliation of groups like "Muslims", Serbs, and Montenegrins (as in Montenegro).

In the table, quantitative data extracted from sources is presented as follows. In general, missing data is marked as NA (not available), while census data and other specific estimates of varying degrees of accuracy are reproduced unchanged. For approximate estimates, we use forms of numerical expression like "more/less/about [1,000/10,000]" = "[1,000/10,000]" and "several tens/hundreds/ thousands/tens of thousands" = "50/500/5,000/50,000." For example, the approximate number of Cretan Greek-speaking Muslims in Turkey is 50,000 people ("several tens of thousands," according to Kahl 2014), while the number of Greek Cypriots is 0.5 million ("more than 500,000 thousand," according to the same source). For some groups, the size remains unknown due to insufficient data from two sources, such as the number of Istriot speakers in Croatia, Aromanian speakers in Greece, Gorani language speakers (who did not appear as a separate ethnic group in censuses until the 2010s) and Turkish speakers in Turkey. The latter is due to the absence of data on the Republic of Turkey in Jordan et al. 2007 and the decision not to provide data on the number of Turkish speakers in Kahl 2014[4].

[3] Denomination of Slavic-speaking Muslims in the former Yugoslavia; currently used by some groups as a self-denomination. This term hereinafter appears in the text of this article in quotation marks.

[4] "Die Zahl der Türkischsprecher in der Türkei ist aufgrund der hohen Zahl von Zweitsprechern anderer etnischen Herkunft (v.a. Kurden) schwer zu ermitteln" (Kahl 2014: 114).

We were required to adopt a special approach for certain groups due to the specific nature of their ethno-linguistic consciousness and its presentation by official sources. These groups include the Albanian-speaking Ashkali and [Balkan] Egyptians residing in Albania, Kosovo, and North Macedonia; Serbo-Croatian-speaking minorities (Bunjevci, Šokci), as well as "Muslims" in the former Yugoslavia member states and Pomaks of Bulgaria. The Ashkali and Egyptians are not represented as separate ethnic minorities in the late 20th and early 21st-century population censuses. Before appearing in the censuses, both groups tended to identify themselves as Albanians. Additionally, Ashkali occasionally self-identify as Roma, while Egyptians never do so, but in Yugoslavia they could unconditionally be counted as Roma. Hence, we classified the number of Ashkali and Egyptians in their respective countries of residence as unknown (NA), and official estimates of the number of Albanians in these countries, derived from sources, were interpreted in the table — and subsequently in the captions on graphs and maps — as follows: Albanians_with_Egyptians (Albania, 2,763,959 people; North Macedonia, 509,083 people), Albanians_with_ Egyptians Ashkali (Kosovo, 1,564,000 people).

Similarly, we accounted for the total number of "Muslims" in countries such as Montenegro, Serbia, and North Macedonia, based on available census data, but differentiated separate groups with an unknown number, such as Gorani and Macedonian-speaking Muslims in the Reka and Župa regions of North Macedonia (in the 2002 Macedonian census, "Muslims" who spoke Serbo-Croatian and Macedonian were combined into a single "ethnic group"). Bunjevci and Šokci in the official censuses of Serbia (2002) and Hungary (2001) were categorized as a Croatian-speaking group with Croats (Croats_with_Bunjevci_Shokci, 76,312 and 25,730 people, respectively). However, for our calculations, they were considered as distinct groups with an unknown number, as in the two previous cases. Lastly, estimating the number of Pomaks in Bulgaria was complicated because Slavic-speaking Pomaks can classify themselves or be classified as Bulgarians, whereas the Turkic-speaking ones can identify as Turks. Therefore, the data on the ethnic composition of the population from the 2001 Bulgarian census used in our study were interpreted as follows: Bulgarians_with_Pomaks — 6,655,210 people, Turks_with_Pomaks — 746,664 people, Pomaks (Bulgarian-speaking) — NA, Pomaks (Turkish-speaking) — NA.

In terms of linguistic differentiation among identified groups and the further visualization of linguistic diversity, one particular challenge was to present the number of Romani speakers residing in Southeastern European countries. A comparison of census data, which reflect the ethnic and linguistic makeup of the population, reveals that not all individuals who identify as Roma necessarily speak the Romani language fluently or at all. For instance, in Romania, the 2002 census recorded 535,140 Roma, but only around 238,000 individuals reported Romani as their mother tongue. Consequently, we had to distinguish Hungarian-speaking (in Hungary), Romanian-speaking (in Romania), Turkic-speaking (in Turkey and Bulgaria), and other Roma groups, without indicating their size, as we were unable to obtain such data from our sources.

Digital Methods for Presenting Information on Ethnolinguistic Groups

The use of modern digital tools has made it possible to conveniently represent data on the number and distribution of studied ethnolinguistic groups through graphical means, such as point representations of the distribution territories. This includes also the choice of character size for the developed scale, taking into account the difference between the minimum and maximum values of the group size. Using the tools described below, we conducted a series of experiments to visualize the political-geographical, estimated ethnographic, and sociolinguistic information that we collected.

These experiments were based on an Excel table containing data on 200 ethnolinguistic groups and 50 language varieties in Southeastern Europe. We utilized the programming language R (R Core Team 2023) as a toolkit, which provides a wide range of possibilities for statistical processing, visualization of data, as well as exporting and publishing the resulting visualizations due to the language's basic functionality and add-on packages that extend it.

The source table was imported into the R environment using the readxl package (Wickham, Bryan 2023). We then made several changes to the resulting data frame (data table) to ensure adequate and convenient visualization of data, including redefining the formats of its columns (text and numeric), and conventionally marking the size of all ethnic groups whose numbers were originally listed as unknown (NA) with the number 1,000.

The data visualizations, generated using the functionality of various R packages, are HTML widgets that can be exported as separate files or integrated into web pages and applications. To

provide free access to the results of our data visualization experiments, we utilized the functionality of the R shiny package (Chang et al. 2022) to create an interactive web application. This application generates the required visualizations and displays them for users to interact with. The web application is hosted on the server of the Institute for Linguistic Studies, Russian Academy of Sciences (Gorlov et al. 2023).

We utilized the plotly package (Sievert 2020) to create an interactive bar chart as the first way to visualize our data. This package allows for the creation of various interactive charts, including bar charts, line charts, histograms, correlation charts, etc. The interactivity of charts generated with this package includes features such as

- scaling, both general (using corresponding interface buttons or scrolling the mouse wheel) and local (using the mouse to select a specific area on the chart),
- visual movement through the chart,
- selection of individual elements using “square” or “lasso” tools,
- exclusion and inclusion of individual elements in the overall visualization by clicking on elements of the chart legend,
- viewing detailed information by hovering over columns with the mouse pointer after selecting either “Show closest data on hover” or “Compare data on hover” in the interface of the chart, and
- saving the chart in its current configuration as a static image.

We have incorporated the plots generated using the plotly package into the first tab of our application, named “Barplots”. The displayed graphs have been divided into two groups, which can be toggled by the user in the control panel located in the upper left corner of the tab.

The first group comprises bar charts that present information on the ethnolinguistic composition of countries and the size of their respective ethnolinguistic groups, using data extracted from the source spreadsheet. To visually differentiate the groups and column segments, we have used a partially modified ready-made palette from the pals package (Wright 2021). The control panel located to the left of the generated graphs has toggle switches that enable users to define the following parameters not regulated by the interactive graph interfaces themselves:

- Y-axis type (linear or logarithmic);
- the order of data output on the X-axis (alphabetical or descending column heights);
- the type of data on the X-axis (countries or ethnic groups).

The second group includes charts that display information about the ethnolinguistic diversity of the countries presented in the source Excel spreadsheet, i.e., the number of ethnic groups and languages found in each of the countries. In this case, toggle switches have been added to the control panel, making it possible to define the following chart parameters:

- Y-axis type (linear or logarithmic);
- the order of data output on the X-axis (alphabetical or descending column heights);
- the type of data on the Y-axis (the number of languages or the number of ethnic groups found in countries).

The next method of data visualization that we explored was digital cartography. The functionality of the R language and the downloadable leaflet package (Cheng et al., 2023) enabled us to create interactive digital maps that display various graphic elements, such as markers, polygons, and legends, and have basic features like map scaling, movement, and interaction with displayed graphic elements. Such maps, along with plotly graphs, can be embedded in a shiny web application. Using information from the source spreadsheet, we have displayed two maps on the second tab of our application, titled "Maps". Switching between them is possible through the control panel located in the upper left corner of the tab.

The first map visualizes data on the ethnic and linguistic diversity of countries, with country boundary polygons being used as the format for the visual representation of these countries. We obtained publicly available cartographic information in the GeoJSON format for these polygons from the geoBoundaries database (Runfola et al., 2020), and imported the data into the R environment using the geojsonio package (Chamberlain et al., 2023). The resulting polygons were assigned manually specified colour tints according to the number of ethnic groups (ranging from 3 to 26) and the number of languages (ranging from 1 to 23) found in a particular country. In addition to the polygons, we added the following functions to the map:

- an explanatory legend located in the upper right corner of the map;
- a toggle switch located in the lower right corner of the map for displaying either ethnic or linguistic diversity;
- pop-up fields which show the exact number of ethnic groups or languages found in a particular country when hovering the mouse over its territory, along with a full list of groups or languages on click.

In our second map, we opted to display information derived from the data in the source Excel spreadsheet concerning both the ethnic diversity of countries and their precise ethnic composition. To accomplish this, we used pie charts positioned over each country according to the coordinates of their approximate geographic centers or "centroids", obtained from the downloadable CoordinateCleaner package (Zizka et al. 2019). The charts were created using the leaflet.minicharts package's features (Bachelier et al. 2021) and coloured using the previously mentioned pals package's palette. Each pie chart illustrates the percentage of the total population of the country belonging to each ethnic group. We chose to display information about the ethnic diversity of countries using varying diameters of these pie charts. Additionally, we added the option to view the precise information regarding the size of ethnic groups in a pop-up field that appears after clicking on a particular chart.

Finally, on the third tab of our application, "Table", we presented our initial data in an unaltered tabular format (specifically, unknown values for the number of ethnic groups are not represented as 1,000, as was the case for visualizations in graphs and maps). We accomplished this using the R reactable package (Lin 2023). The resulting interactive table enables users to conduct full-table searches, search by individual columns, and sort the entire table by values in a specific column.

Conclusions

The primary data from both sources have been extracted and presented in digital form, allowing for convenient visualization of the information contained within. It is obvious that in the digital era, the gradual splitting of large and medium ethnolinguistic groups, up to the indication and visualization of infinitely small ones, is not a technical problem as any groups can be combined into any groups according to any relevant feature and visualized in any way.

In the digital age, there is no need to establish a quantitative limit and exclude even the smallest groups from consideration, such as the Istro-Romanians (with less than 1,000 people) or the Turks (with less than 400 people) of Croatia. However, it would be incorrect to mechanically attribute the entire number of representatives of any group to the speakers of the corresponding language, in specific cases such as Istro-Romanian and Turkish, and even more so to the speakers of one of the dialects of these languages. For instance, among the Turks of Croatia, there may be both Muslims

and non-Muslims, speakers of various Turkic varieties, and speakers of South Slavic varieties.

The digitalization and visualization tools chosen in this work are ergonomic, accurate, and convenient. The results of the synthesis of primary data in graphs and on geographical maps are clear, and the interface is user-friendly. Continued work with the R software package, which has a number of advantages over similar tools such as Wordpress and Drupal, can be considered a promising direction.

In contrast to the analogue publications of the Vienna Atlas, our visualization result is not only for the first time made using digital technologies and presented in digital form, but it also fully includes the Hellenic Republic, which is rarely considered, perceived and presented as a country distinguished by ethnolinguistic and linguistic diversity in Western science. Additionally, we introduce an index of ethnolinguistic diversity in our work. Although we have to apply it to individual states at the present stage, in the future, it can be introduced for various regions of Southeastern Europe, regardless of political boundaries. Despite the unavoidable incompleteness of the primary data at present and the inability to accurately localize a significant part of them in geographic space, a certain general opposition can be distinguished between the ethnolinguistically and linguistically more uniform western part of the Balkan Peninsula with its dispersed, often very small ethnic groups, and the northeast, east, and south of the entire Southeast European area with compact subareas of residence of large ethnolinguistic groups numbering hundreds of thousands of people. The geographical distribution of quantitatively different groups may indicate different ways of forming the ethnolinguistic landscape, including migrations and various models of colonization, such as Hellenic, Roman, Slavic, Ottoman, Habsburg, and others.

The mechanical enumeration of the number of points for a hypothetical Linguistic Atlas of Southeastern Europe, aimed at reflecting the ethnolinguistic diversity of the territory as fully as possible, results in either a “politically correct” outcome of 200 ethnolinguistic groups, or a “typologically oriented and quantitatively disproportionate” outcome of 50 purely linguistic entities, or, if we consider an ethnolinguistic group of 5,000 as the minimum unit, an idealistic outcome of several thousands. Our analysis reveals that, in reality, ethnic and linguistic groups of people are treated separately, as different entities, in modern reference scientific literature, despite declarations of an integrated approach to them. Meanwhile, linguists, who typically deal with the latter, have to rely

on incomplete, imperfect, and often unreliable information about the former.

We can formulate a problem that arises at the intersection of different sciences and is relevant to linguistics: how to quantify knowledge about linguistic entities (different forms of the existence of languages that are not national standards) and visualize this knowledge in lists, tables, graphs, diagrams, and maps? Currently, on the basis of the quantitative and geographical data available in world science, we cannot draw generalizations about a number of ethnolinguistic groups (including quite large ones like the Aromanians); the quantitative aspect of the actual linguistic diversity of ethnolinguistic groups (for example, how many Croats currently speak Chakavian varieties or spoke them in the mid-20th century, how many Bulgarians speak Mysian, how many Albanians speak the Labëri subdialect, etc.); or how to account for and visualize the quantitative aspect of dialectal and sociolinguistic variation in the languages of these groups. It is also difficult to determine how these data vary across sub-regions of Southeastern Europe and how they can be mapped regionally.

Therefore, we face, in particular, the problem that the analogue dialectological maps of the idioethnic languages of Southeastern Europe are currently not correlated with quantitative data on the corresponding ethnolinguistic groups. Consequently, we cannot *quantify* the linguistic differentiation within an ethnic group and calculate the relationship between linguistically differentiated ethnic groups. At the same time, the depth of dialect differentiation and fragmentation of the dialect division of different idioethnic languages and different parts of the dialectal landscape of one language can vary significantly, and for some cases, such as the Greek language, this is still unknown. It is also impossible to distinguish between ethnolinguistic groups where this distinction is not explicitly provided in our sources (for example, this concerns the separate representation of Ashkali and Egyptians in Kosovo, Albanian-speaking Egyptians and Ashkali in Montenegro, Romanian-speaking and other Banyashi in Serbia, Hungarian- and Serbian-speaking Jews in Serbia, “Yugoslavs”, and other speakers of the Serbo-Croatian language, etc.). A particular problem arises due to the fact that linguistically (primarily contactologically) interesting peripheral idioms, including diasporas, have a very small number of speakers.

It is highly possible that the exhaustive identification, compilation, and explication of all the factors affecting the “number

of speakers" in L1, L2, and so forth, in historical dynamics, social stratification, and uneven geographical distribution is a task that may not be feasible even for "long- and well-studied" regions such as the bilingual cantons of Switzerland in Western Europe. Complicating the matter further is the overlapping of dialects by standard languages, regiolects, and koines, which is already widespread in Southeastern Europe since the latter half of the 20th century. For instance, the number of speakers of the basic Rhodopean or basic Tran Bulgarian dialect in Bulgaria in the 21st century tends to zero, while a century ago, the entire population of the corresponding region could be counted as such speakers (bearing in mind that linguistic and administrative borders usually do not coincide). Additionally, it is equally difficult to consider the changing parameters of bilingualism during significant political transformations. For example, should all Albanians of Kosovo be considered speakers of Serbo-Croatian as L2 from 1913 to 1999, and Albanians of present-day North Macedonia as speakers of Serbian and then Macedonian as well? Finally, to what extent can both of these ethnolinguistic groups be considered speakers of the standard Albanian language as of 1972?

Digital tools that we use can offer the possibility of producing both generalizing graphs of the entire region without taking into account its political fragmentation and mapping the regional, cross-border distribution of ethnolinguistic groups. Both approaches can be useful in visualizing information on groups like Aromanian speakers. To do this, the list of data sources must be expanded, constantly structured, and verified, clarified, updated, and supplemented according to the most reliable, depoliticized academic publications (Sorescu-Marinković et al., 2020). We are confident that the direct replenishment of primary data using information from national censuses and international and local public organizations is possible, provided they are confirmed by the independent *scientific*, *academic* community in authoritative peer-reviewed scientific publications.

It is evident that more research is required to achieve the aim of presenting and visualizing the diversity of linguistic and ethno-linguistic facts comprehensively, reliably, and proportionally, taking into account the varying depths of dialect differentiation on the ground in Southeastern Europe. In the future, compiling genuinely representative and inclusive *linguistic and areal-linguistic (linguogeographic) samples* for the region is desirable. While

automating this process seems to be a matter of the distant future, expert evaluation remains the most critical tool.

Bibliography

Adamou, Evangelia (Ed.). 2008: *Le patrimoine plurilingue de la Grèce*. Leuven: Peeters.

Adamou, Evangelia. Sobolev, Andrey N. (Eds). 2023: *Atlas of the Balkan Linguistic Area Online*. https://abla.cnrs.fr/dataset/serbian-tuchep/ (Last accessed 2023-04-26)

Alinei, Mario A. et al. (Eds.). 1975: *Atlas Linguarum Europae (ALE): Introduction*. Assen: Van Gorcum.

Asher, Robert E. Moseley, Christopher. (General Editors). 2007: *Atlas of the World's Languages*. Second Edition. London and New York: Routledge.

Bachelier, Veronique et al. 2021: *leaflet.minicharts: Mini Charts for Interactive Maps*. R package version 0.6.2. https://CRAN.R-project.org/package=leaflet.minicharts. (Last accessed 2023-04-26)

Bernštejn, Samuil B. et al. (Eds.). 1987: *Atlas dialectologique des Carpathes. Tome introductif.* Skopje: MANU.

Carling, Gerd (Ed.). 2019: *Mouton Atlas of Languages and Cultures*. Berlin; New York: de Gruyter Mouton.

Chamberlain, Scott. 2023: *geojsonio: Convert Data from and to 'GeoJSON' or 'TopoJSON'*. R package version 0.11.0. https://CRAN.R-project.org/package=geojsonio. (Last accessed 2023-04-26)

Chang, Winston et al. 2022: *shiny: Web Application Framework for R*. R package version 1.7.4. https://shiny.rstudio.com/ (Last accessed 2023-04-26)

Cheng, Joe et al. 2023: *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*. R package version 2.1.2. https:// CRAN.R-project.org/package=leaflet. (Last accessed 2023-04-26)

Comrie, Bernard. 2007: Northern Asia and Eastern Europe. In Asher, Robert E. Moseley, Christopher. (General Editors). *Atlas of the World's Languages*. Second Edition. London and New York: Routledge, 231–240.

Dami, Aldo. 1976: *Les frontières européennes de 1900 à 1975 : histoire territoriale de l'Europe*. Genéve: Médecine et hygiène.

Deanović, Mirko. 1964: Deux atlas plurilingues et la slavistique. In *Revue des Études Slaves*. 40, 55–60.

Demeter, Gábor. Bottlik, Zsolt. 2021: *Maps in the service of the nation : the role of ethnic mapping in nation-building and its influence on political decision-making across the Balkan Peninsula (1840–1914)*. Berlin: Frank & Timme.

Dryer, Matthew S. Haspelmath, Martin. (Eds.). 2020: *WALS Online* (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/54, Accessed on 2023-04-03.)

Gil, David. 2013: Distributive Numerals. In Dryer, Matthew S. Haspelmath, Martin. (Eds.). *WALS Online* (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/54, Accessed on 2023-04-03.)

Gorlov, Nikita G. Morozova, Maria S. Sobolev, Andrey N. 2023: Web Only Supplemental Material (Open Access). Ethnolinguistic groups of Southeastern Europe: Ways of presentation. In *Indoevropejskoe âzykoznanie i klassičeskaâ filologiâ* [Indo-European Linguistics and Classical Philology Yearbook]. https://shiny.iling.spb.ru/Gorlov_et_al_2023/ (Last accessed 2023-04-26)

Jordan, Peter et al. (Eds.). 1995: *Atlas Ost- und Südosteuropa. Ethnische Struktur Südosteuropas um 1992*. Wien: Österreichisches Ost- und Südosteuropa-Institut Wien.

Jordan, Peter et al. (Eds.). 2006: *Atlas Ost- und Südosteuropa. Ethnisches Bewusstsein in Mittel- und Südosteuropa um 2000*. Wien: Österreichisches Ost- und Südosteuropa-Institut Wien.

Kahl, Thede. 2014: Ethnische, sprachliche und konfessionelle Struktur der Balkanhalbinsel. In Himstedt-Vaid, Petra et al. (Eds.). *Handbuch Balkan*. Wiesbaden: Harrassowitz, 87–134.

Kamusella, Tomasz. 2021: *Words in Space and Time : A Historical Atlas of Language Politics in Modern Central Europe*. Budapest; Vienna; New York: Central European University Press.

Krallert, Wilhelm. 1941: *Volkstumskarte von Jugoslawien*. Wien: o. Verl.

Lachlan, Mackenzie J. 2007: Western Europe. In Asher, Robert E. Moseley, Christopher. (General Editors). *Atlas of the World's Languages*. Second Edition. London and New York: Routledge, 259–264.

Lin, Greg 2023: *reactable: Interactive Data Tables for R*. R package version 0.4.4. https://CRAN.R-project.org/package=reactable. (Last accessed 2023-04-26)

Magoczi, Paul Robert. 2018: *Historical Atlas of Central Europe*: Third revised and expanded Edition. Toronto: University of Toronto Press.

R Core Team. 2023: *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/. (Last accessed 2023-04-26)

Runfola, Daniel et al. 2020: geoBoundaries: A global database of political administrative boundaries. In *PLoS ONE*. 15(4): e0231866.

Scherbakova, Olena et al. 2023: Societies of strangers do not speak grammatically simpler languages. In *Sciences Advances.* Submitted manuscript. DOI:10.31235/osf.io/svfdx

Sievert, Carson. 2020: *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Florida: Chapman and Hall/CRC. https://plotly-r.com. (Last accessed 2023-04-26)

Sobolev, Andrey N. (Ed.). 2003: Malyi dialektologicheskii atlas balkanskikh iazykov (MDABYa) [Minor dialectological Atlas of the Balkan languages (MDABL)]. München: Verlag Otto Sagner; St. Petersburg: Nauka.

Sorescu-Marinković, Annemarie. Mirić, Mirjana. Ćirković, Svetlana. 2020: Assessing Linguistic Vulnerability and Endangerment in Serbia. A Critical Survey of Methodologies and Outcomes. In *Balcanica*. LI, 65–104.

Straka, Manfred. 1979: *Völker und Sprachen Europas unter besonderer Berücksichtigung der Volksgruppen*. Graz: Akademische Druck- und Verlagsanstalt.

Wickham, Hadley. Bryan, Jennifer. 2023: *readxl: Read Excel Files*. R package version 1.4.2. https://readxl.tidyverse.org, https://github.com/tidyverse/readxl. (Last accessed 2023-04-26)

Wright, Kevin. 2021: *pals: Color Palettes, Colormaps, and Tools to Evaluate Them*. R package version 1.7. https://CRAN.R-project.org/package=pals. (Last accessed 2023-04-26)

Zizka Alexander et al. 2019: CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. In *Methods in Ecology and Evolution*. 10(5), 744–751. DOI:10.1111/2041-210X.13152. https://github.com/ropensci/CoordinateCleaner (Last accessed 2023-04-26)

Interactive digital appendix to this article: Gorlov, Nikita G. Morozova, Maria S. Sobolev, Andrey N. 2023. Web Only Supplemental Material (Open Access). Ethnolinguistic groups of Southeastern Europe: Ways of presentation. In *Indoevropejskoe âzykoznanie i klassičeskaâ filologiâ* [Indo-European Linguistics and Classical Philology Yearbook] is located on the link https://shiny.iling.spb.ru/Gorlov_et_al_2023/ (Last accessed 2023-05-11).