The website «Minority languages of Russia»: visualizing language data

Konstantin Polivanov[a], Olga Kazakevich[a,b], Natalia Serdobolskaya[a], Zaira Khalilova[a], Elena Budyanskaya[a], Anastasia Evstigneeva[a, c], Karina Mishchenkova[a, d, e], Daria Mordashova[a, c], Sofie Pokrovskaya[a] and Evgeniya Renkovskaya[a, f]

a. *Institute of Linguistics, Russian Academy of Sciences, 1 bld. 1, Bolshoy Kislovsky Lane, Moscow, 125009, Russian Federation*

b. *Institute of Linguistics, Russian State University for the Humanities; GSP-3, 6, Miusskaya Pl., Moscow, 125993, Russian Federation*

c. *Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation*

d. *School of Linguistics HSE University, 21/4, Staraya Basmannaya str., Moscow, 105066, Russian Federation*

e. *Ivannikov Institute for System Programming, 25, Aleksandra Solzhenitsyna str., Moscow, 109004, Russian Federation*

f. *Institute of Oriental Studies, Russian Academy of Sciences, 12, Rozhdestvenka str., Moscow, 107031, Russian Federation*

Abstract

The paper deals with the issue of language data visualization on the website "Minority languages of Russia". This site is created as an open information resource containing materials on the functioning and structure of minority languages of Russia and their local varieties. Methods of data visualization in five areas are considered: genealogy, areal distribution, domains of language usage, dynamics of language usage and phonetic data representation. For each area, theoretical problems, related tasks and technical implementation are indicated, and prospects for further work are discussed.

Keywords

Minority languages of Russia, visualization, genealogy, areal distribution, domains of language usage, dynamics of language usage, phonetic data representation, JavaScript, HTML, CSS.

1. Introduction

The website "Minority languages of Russia" is created on the basis of Laboratory for Research and Preservation of Minority Languages, Institute of Linguistics, Russian Academy of Sciences, as an open information internet resource containing materials on the functioning and structure of minority languages of Russia and their local varieties. The pilot of the website is available at: http://minlang.iling-ran.ru.

The main task of this resource is to systematize the available information on languages. This includes field data obtained from language experts, as well as various published materials. The website contains both modern data, which we intend to update regularly, and those related to the previous century.

Proceedings of the Linguistic Forum 2020: Language and Artificial Intelligence, November 12-14, 2020, Moscow, Russia EMAIL: polivanov.studio@gmail.com (Konstantin Polivanov); kazakevich.olga@gmail.com (Olga Kazakevich); serdobolskaya@gmail.com (Natalia Serdobolskaya); zaira.khalilova@gmail.com (Zaira Khalilova); budyanskaya.lena@gmail.com (Elena Budyanskaya); evstigap@gmail.com (Anastasia Evstigneeva); karinam6@mail.ru (Karina Mishchenkova); mordashova.d@yandex.ru (Daria Mordashova); sofie.v.pokrovskaya@gmail.com (Sofie Pokrovskaya); jennyrenk@gmail.com (Evgeniya Renkovskaya)

ORCID: 0000-0003-4571-9476 (Konstantin Polivanov); 0000-0003-0597-2979 (Olga Kazakevich); 0000-0003-2417-5537 (Natalia Serdobolskaya); 0000-0003-1604-4510 (Zaira Khalilova); 0000-0002-6306-6280 (Elena Budyanskaya); 0000-0002-3413-2058 (Anastasia Evstigneeva); 0000-0001-9175-987X (Karina Mishchenkova); 0000-0002-8330-520X (Daria Mordashova); 0000-0002-4895-7069 (Sofie Pokrovskaya); 0000-0003-1944-0746 (Evgeniya Renkovskaya)

Further on the paper is structured as follows: §2 provides an overview of related work, §3 describes data and methods - the arrangement of the language profile in terms of content and technical implementation. §4 is the central part of the paper, it is devoted to the results of our work on language data visualization in five areas of concern. §5 discusses the future prospects of the work, and §6 summarizes the results. [1]

2. Related work

In the development process we take into consideration the existing tradition of sociolinguistic descriptions of the world's languages: in particular, the results of the long-term project "The written languages of the world", created through joint efforts of International Centre for Research on Language Planning, Laval University (Canada, Quebec) and research groups from all over the world [1, 2, 3, 4, 5]. The two-volume edition "Written languages of the world: the Russian Federation" is also a part of this project [6, 7]. Furthermore, we take into account such large-scale monographs on the sociolinguistic situation in Russia as "Languages of the peoples of Russia: The Red Book" and "Language and Society" [8, 9].

However, the online format of our resource provides more opportunities than the format of a printed publication. For instance, the information can be regularly updated to remain relevant, and special types of data can be presented (see §3 for further details). Before we started working on the website, there had already been several resources dedicated to certain regions and languages. Let's take a closer look at two examples.

The website "Minority languages of Siberia as our cultural heritage" [2] is a long-term research project on three languages: Ket, Selkup and Evenki [10]. We use the experience of this resource to a large extent as far as technical implementation concerns. In particular, content management framework Drupal was chosen as the site's system core, which is well suited for working with a large amount of complex data [11]. Besides, we follow the way of presenting certain types of information, such as texts.

Another online project is the Atlas of Multilingualism in Dagestan[3] [12]. It is concerned with a single area of sociolinguistics (namely, multilingualism) on the basis of a limited number of languages (namely, the languages of Dagestan, Russia). The website contains an up-to-date database on Dagestanian multilingualism, equipped with a search interface. The implementation of some information blocks is similar to our website: a map with an indication of the languages and the settlements where these languages are spoken, and visualization of census data.

In general, the projects on multilingualism in Dagestan and the minority languages of Siberia can be characterized as narrowly focused, while our website is aimed at a more encyclopedic approach that enables the users to obtain a wide range of information on a large number of languages accommodated in the same place. In the future we intend to cover all the languages of Russia.

3. Methods and data

The main section of the site is the webpage "Languages" that contains a list of the minority languages of Russia compiled by the Institute of Linguistics, Russian Academy of Sciences[4]. At the same time, the languages are not just sorted alphabetically, but are grouped in language families (including isolated and mixed languages) and listed alphabetically inside language families and/or groups. Regarding the language families' structure, we relied on the latest elaborations (for more details see §4.1). There is an access to the detailed information on every language from this list.

[1] We are currently working on the Russian version of the site, an English version is planned for the future. In the paper we use illustrations translated into English for the convenience of the readers.

[2] This resource is available at http://siberian-lang.srcc.msu.ru.

[3] This resource is available at https://multidagestan.com . As of November 2020, it contained the information on multilingualism of the residents of 60 villages.

[4] The list of the languages of Russia is presented on the IL RAS website at https://iling-ran.ru/web/ru/jazykirf . This section was prepared by Yu. B. Koryakov in collaboration with T. B. Agranat, M. A. Goryacheva, A. V. Dybo, O. A. Kazakevich, A. A. Kibrik, O. V. Khanina and A. B. Shluinsky.

One of the main sources of materials is field data collected during the recent field trips (including illustrative materials: photographs, current cartographic and sociolinguistic data, video and audio recordings of the texts with transcripts). To obtain these materials, we communicate directly with language researchers (both in Russia and abroad). The experts are invited to fill out a questionnaire developed by the Lab, that enables to unify the presentation of languages on the site.

The website also contains the results of the Russian Federation and the Soviet Union population censuses (1926, 1989, 2002, 2010). In addition, we draw on early accounts of the functioning of languages, mostly compiled in the last century. It allows to track the language situation development for each particular language.

As for the technical implementation, page templates are built in an optimal way, semantic elements (such as header, footer, article) are used in the construction in order to clearly define the content. Unnecessary wrapper elements are excluded, thus reducing nesting and simplifying the structure of the DOM tree (the DOM object model of the document). The names of identifiers and classes of markup elements are called by understandable words in order to facilitate the reading and machine parsing. A set of SEO optimization tasks is being performed. All this provides a fast and obvious path to the data, as well as efficient indexing by search engines.

The user interface is developed by means of modern data structuring and markup tools - HTML5, as well as cascading CSS3 style sheets. We select the most convenient ways of presenting the information.

JavaScript and jQuery libraries are used to represent interactive elements. In particular, the leaflet library is used (with additional extensions such as "Leaflet Markerclusterer" for clustering points on the map) for programming interactive maps [13]. For developing the interactive dynamic data visualizations we use the libraries JavaScript D3 and chart.js [14, 15].

Some visualizations are built with the help of "datawrapper" - a service based on JavaScript and HTML5 that can represent the data in the form of graphs and maps [16].

The structure of the language page and the features of its technical implementation are presented in Appendix 1.

4. Results

An important task in the development of the language page was to avoid the presentation of data in the form of a continuous text. It was necessary to organize a comfortable navigation through the sections of the page, as well as to enliven the presentation of the data. An interactive table of contents and various rendering methods were used to make the page more aesthetically appealing and user-friendly.

During the development of the language page, we faced a number of difficulties in terms of language data visualization. The paper examines five of them: the representation of genealogy, areal distribution, domains of language usage, dynamics of language usage and phonetic data. Each subsection is structured according to the following scheme: theoretical problems, related tasks of visual representation and the technical implementation of this representation.

4.1. Representation of genealogy

Genealogical relationship is basic information that is usually given in the beginning of a language description in the form of a short note about a related language group and a family; it does not imply further enumeration of related languages belonging to the same group or other groups of the same family. Still, since our website targets a wide range of users, we find it crucial to provide a complete picture of the family tree for each language.

As already mentioned in §2, we follow the theoretical paradigm developed in the Institute of Linguistics regarding the list of languages and genealogical relationships. For example, we consider that separate groups of the Altaic languages are genetically related and not just form a Sprachbund. Moreover, we keep track of the latest studies in this field and present a new division within the Uralic language family [20] on our site.

It is important for us to achieve the following goals: easy usage; adequate data display; visualization of the full genealogical picture (from dialect to family) with a possibility to narrow it down to specific segments.

As a solution, we've chosen two interactive schemes that differ in the way of data visualization. The first scheme is a common representation of a language family as a genetic language tree (see Figure 1). Most trees have a lot of branches, so we provided an opportunity to collapse tree nodes to easily focus on a specific segment of the scheme. The collapsed nodes are coloured grey. To make the diagram clearer we coloured the nodes of the languages, groups of dialects and dialects in contrasting colours that distinguish them from the higher level nodes. The values of the colours are given in the diagram legend.

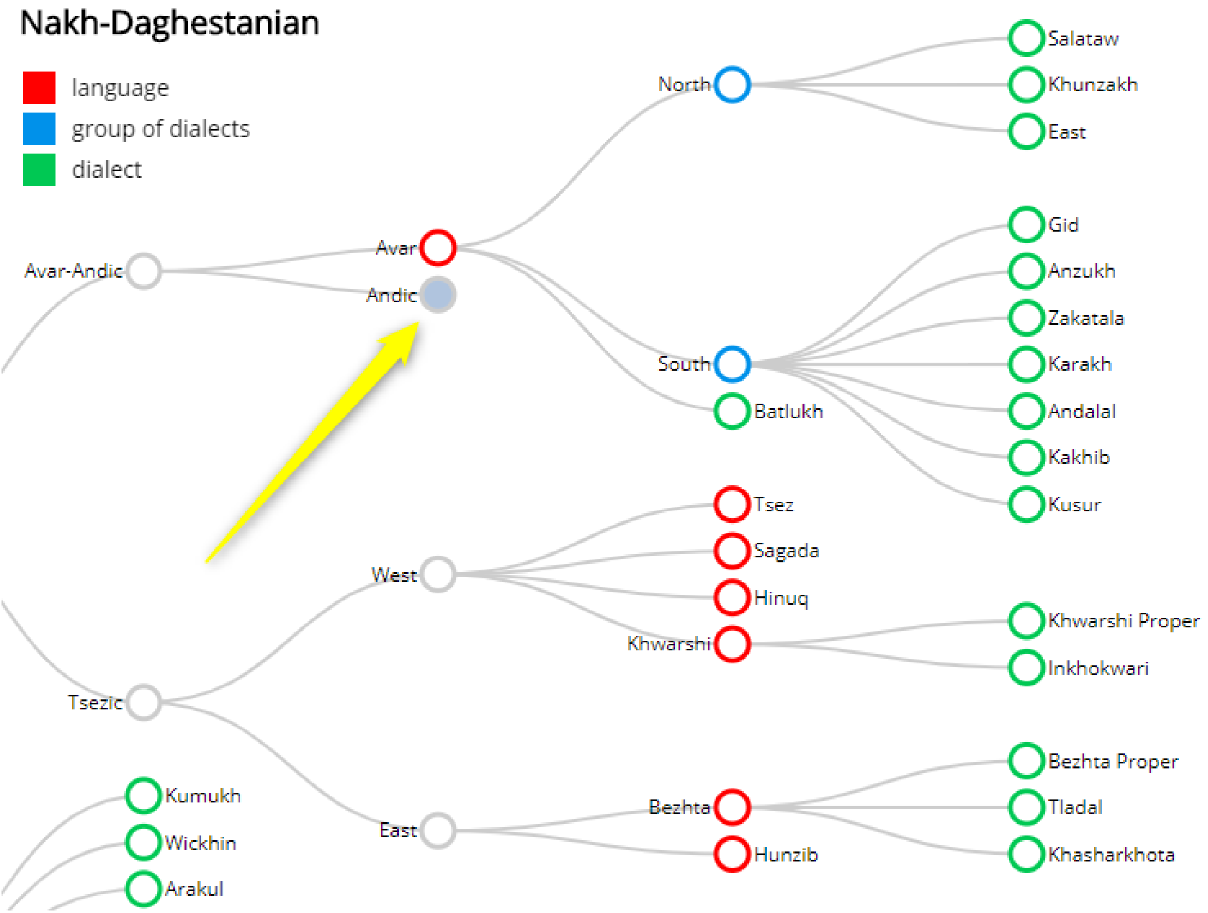
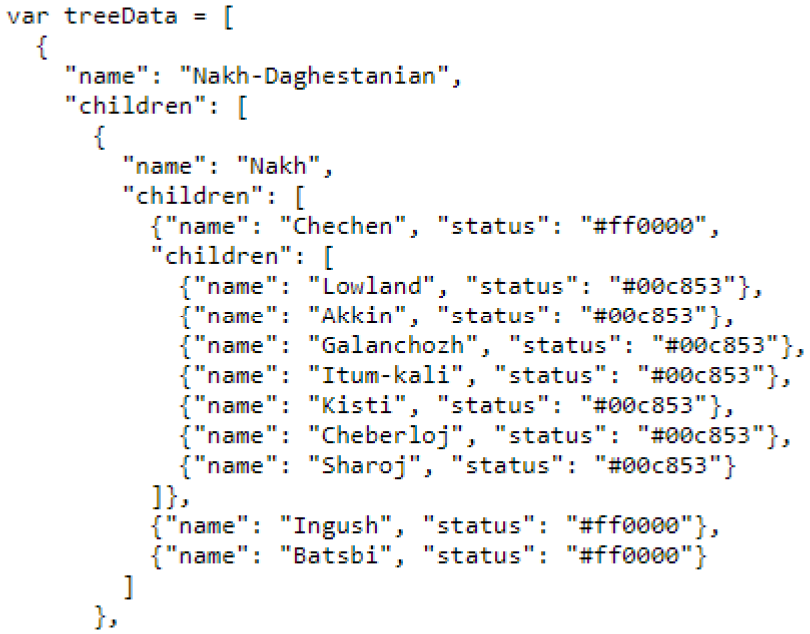


Figure 1: A fragment of the Nakh-Daghestanian genealogical tree. The arrow points at the collapsed node that can be unfolded.

The visualization is based on the D3 tree diagram. The data presented in the tree are pre-packaged in a json file.

The default structure of the json file from D3 tree did not meet our goals, so we needed to expand the set of properties and values. Our json file specifies the following properties for the elements: child / parent elements and colors indicating that the elements belong to a language, a group of dialects, or a dialect. A snippet of such a file is given in Figure 2.



```
var treeData = [
  {
    "name": "Nakh-Daghestanian",
    "children": [
      {
        "name": "Nakh",
        "children": [
          {"name": "Chechen", "status": "#ff0000",
          "children": [
            {"name": "Lowland", "status": "#00c853"},
            {"name": "Akkin", "status": "#00c853"},
            {"name": "Galanchozh", "status": "#00c853"},
            {"name": "Itum-kali", "status": "#00c853"},
            {"name": "Kisti", "status": "#00c853"},
            {"name": "Cheberloj", "status": "#00c853"},
            {"name": "Sharoj", "status": "#00c853"}
          ]},
          {"name": "Ingush", "status": "#ff0000"},
          {"name": "Batsbi", "status": "#ff0000"}
        ]
      },
```

Figure 2: A snippet of the json-file

Figure 2 demonstrates the code of the json-file. In line 9, the node with the name "nakh" (in technical terminology, in this case, the attribute "name" with the value "nakh") has the parent node "nakh-daghestanian" in line 6 and child nodes with the names "Ingush" and "Batsbi" in lines 21 and 22, respectively. The "status" attributes contain RGB HTML colour codes in their values. For example, "name":"Chechen","status":"#ff0000" means that the node "Chechen" is coloured red. The code for parsing the json file and displaying it in the diagram was rewritten.

The second scheme represents a language family as circles packed in one another (see Figure 3). The nested circles correspond to the levels of genealogical classification, and the intensity of the colour depends on the nesting depth. The lowest levels are marked in white.

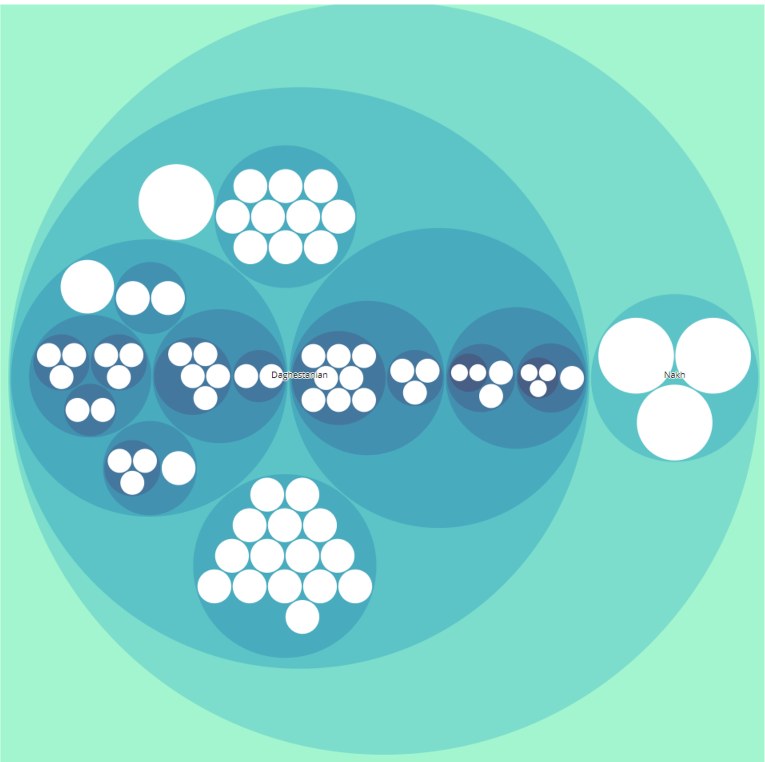


Figure 3: The Nakh-Daghestanian language family

Technically, this scheme is implemented by using the js-library D3 Zoomable Circle Packing. The data are packed in a json file with the appropriate structure. Then, js parses the json file and builds a diagram.

The advantage of this scheme is that it allows to see the whole family in one picture, to get the idea of its size and complexity. Nevertheless, for the moment the choice was made in favour of the tree model, because the default characteristics of the second scheme do not meet some of our goals. In particular, it is not possible to see the names of all languages at once. And what is even more important we cannot mark dialects, groups of dialects and languages with different colours (colours in this scheme are assigned automatically according to the nesting depth of a circle), and a suitable solution for this problem has not been found yet. The modification of this diagram in accordance with our tasks is the subject of further work.

4.2. Representation of areal distribution

An important part of language description is the representation of the areal distribution. There are two ways of displaying this area on the map: with polygons (we paint over the territory where the language is spoken) or with markers (we mark the settlements where the language is spoken). In the future, we intend to use both options: markers - for languages with a small number of local variants, polygons - for languages with a large number of local variants. Since we started with languages with a small number of local variants, at the moment the second option with some modifications is represented on the language page.

We saw our task here in the following: visualization of distribution of the local variants in the settlements; easy usage; adequate, but at the same time compact display of the data.

To achieve these goals we developed an interactive map where the settlements are marked according to the local variants of the language that are spoken there (see Figure 4). Each dialect gets its colour (see two blue markers at the top of the map for Bezhta Proper dialect in Figure 4). The values of the colours are given in the legend under the map. For settlements where more than one dialect is spoken a special multicoloured marker was introduced (see the marker on the right side of Figure 4). When a marker is clicked, a list of dialects appears in a pop-up window (see Figure 5). When zoomed out, several points on the map are combined into one cluster marked with a number that indicates the amount of settlements within the radius of this cluster (see the green marker in the center of Figure 4). This makes the presentation of the data more compact.

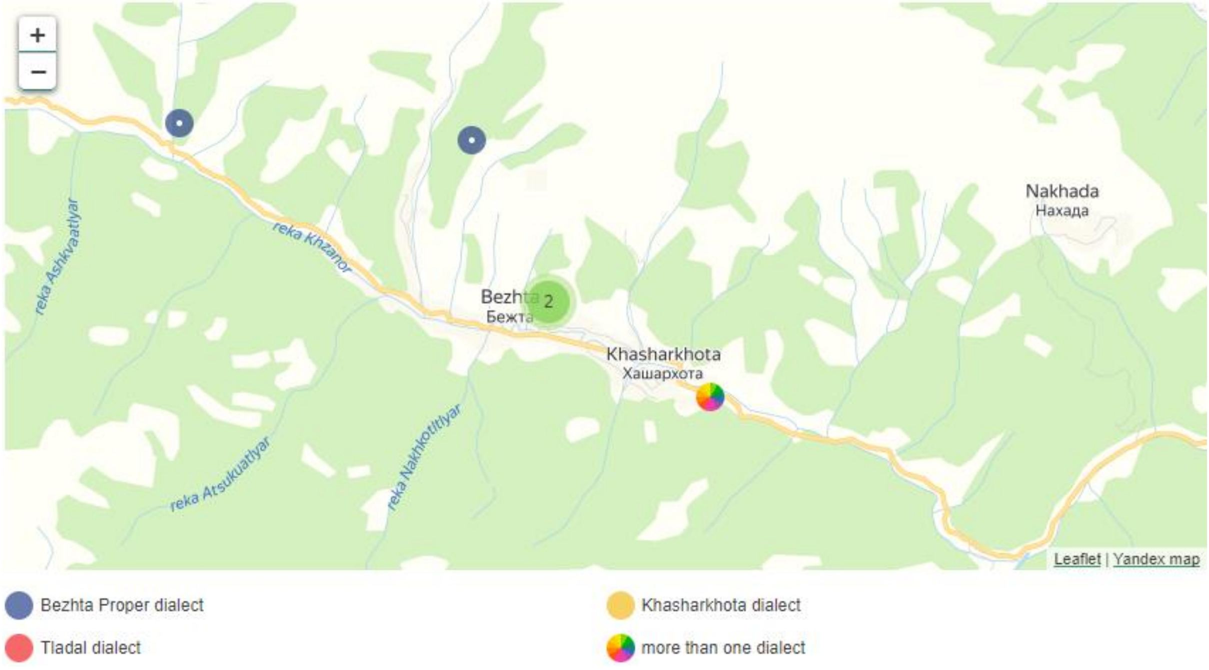


Figure 4: A fragment of the territory where Bezhta is spoken

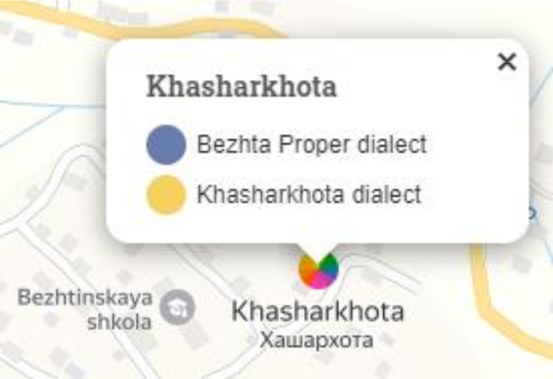


Figure 5: A window with a list of dialects that pops up when you click on a special multicoloured marker of a settlement

The data are represented on the map using the leaflet JavaScript library [13]. Leaflet is a library that has proved its capability in a large number of resources. It provides great opportunities for working with geodata. Leaflet enables one to operate with any type of data (markers, polygons, lines), to use any substrates (OpenStreetMap, Yandex, GMap, etc.), and to customize the appearance of the markers and polygons displayed on the map according to one's needs. Leaflet can be extended with plugins, for example, for clustering data or for attaching json files with a database. Unlike Mapbox (a paid platform for working with GIS), there are no traffic restrictions.

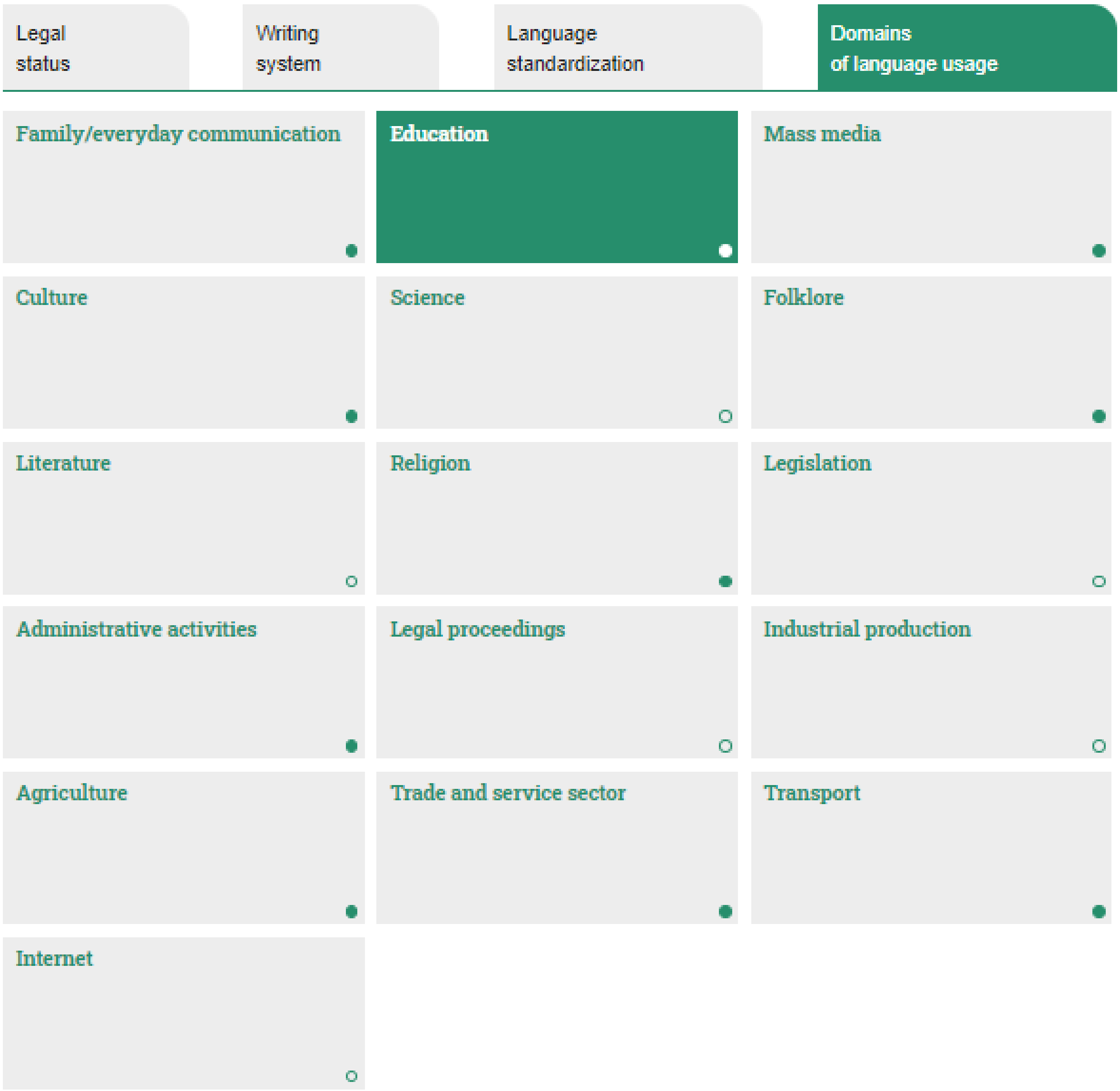
The geographical coordinates of the settlements are pre-packed in an array with the following properties: settlement name, dialect, and dialect code. Then the array is output to the map with marker clustering. The array elements correspond to the following model: [{latitude Coordinate}, {longitude Coordinate}, {pop-up window content}, {dialect number (more than one dialect is indicated by the letter m)}].

4.3. Representation of the domains of language usage

An important component of the sociolinguistic description of a language is a list of domains in which it functions. Thus, the volume and nature of the language use in these domains allow one to judge the vitality of the language. A list of 16 core domains was compiled, starting with family communication and education and up to the language use on the Internet.

When choosing a method for representing the domains, the following tasks were primarily solved: visualization of a complete picture of the domains within a single user screen; convenience and speed of obtaining the information.

This has been achieved by drawing up an interactive table of sixteen cells (see Figure 6), corresponding to the domains of the language usage. For ease of viewing and perception, the table is structured as follows: the cells of the table contain the titles of the domains, and it is noted whether the language is used in this domain or not (see the “empty or filled circle” marker in the lower right corner of each cell in Figure 6). When clicking on a cell, a pop-up window opens with expanded information about the peculiarities of the language use in this domain.



Legal status | Writing system | Language standardization | Domains of language usage

Family/everyday communication ●	Education ●	Mass media ●
Culture ●	Science ○	Folklore ●
Literature ○	Religion ●	Legislation ○
Administrative activities ●	Legal proceedings ○	Industrial production ○
Agriculture ●	Trade and service sector ●	Transport ●
Internet ○		

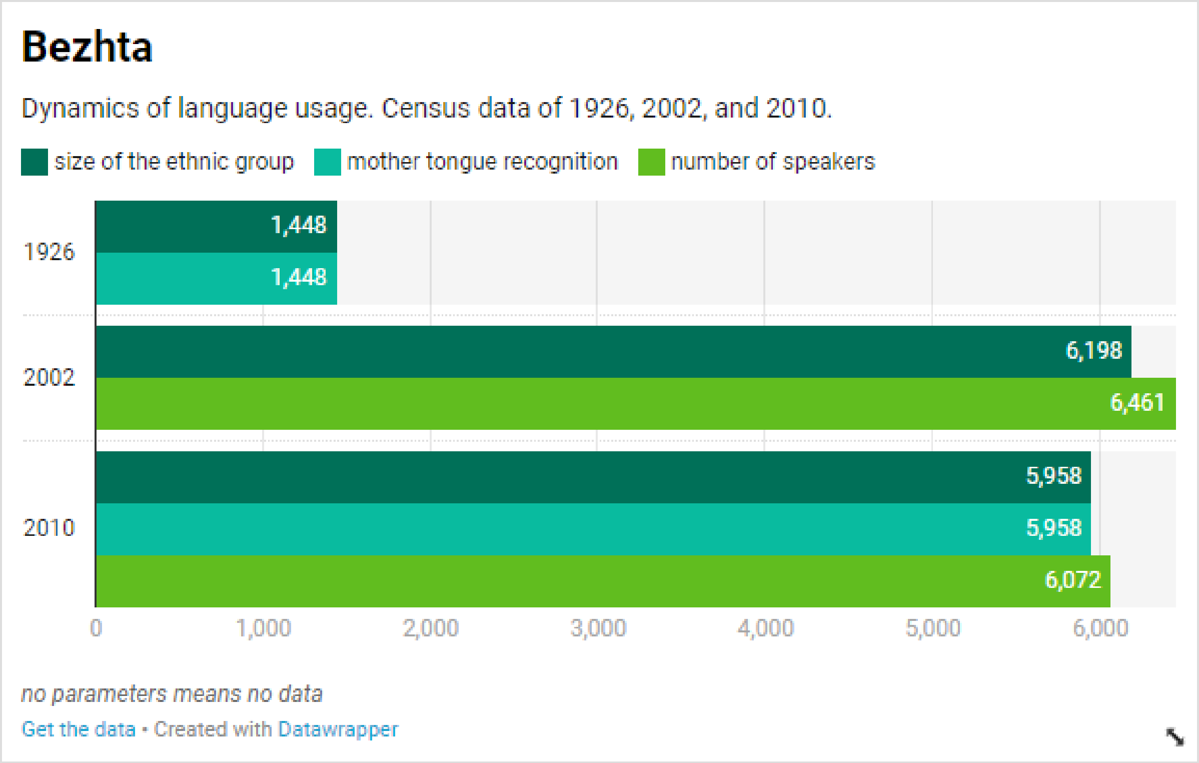
Figure 6: Domains of language usage in Bezhta

4.4. Representation of the dynamics of language usage

For our internet resource dedicated to minority languages, it turns out to be critical to track the change in the size of the ethnic group and the number of speakers over time in the case of each particular language. This section of the language page currently accumulates the data from the population censuses of various years, and in the future, we hope to represent the data from local administrations and researchers' estimates.

Thus, the main task was to visualize the general picture of dynamics, which allows tracking the change of various quantitative indicators in time.

To solve this problem, an interactive diagram was developed. It displays the change in the size of the ethnic group, the number of group members who consider the ethnic language of the group to be their mother tongue, and the number of those who reported proficiency in this language (see Figure 7). A simple charting tool - datawrapper.de - was used [16].



Bezhta

Dynamics of language usage. Census data of 1926, 2002, and 2010.

no parameters means no data

Get the data • Created with Datawrapper

Figure 7: Dynamics of the Bezhta language usage

4.5. Representation of phonetic data

‘Phonetic data’ usually refers to an array of systems, describing both acoustic/articulatory phenomena and phonology.

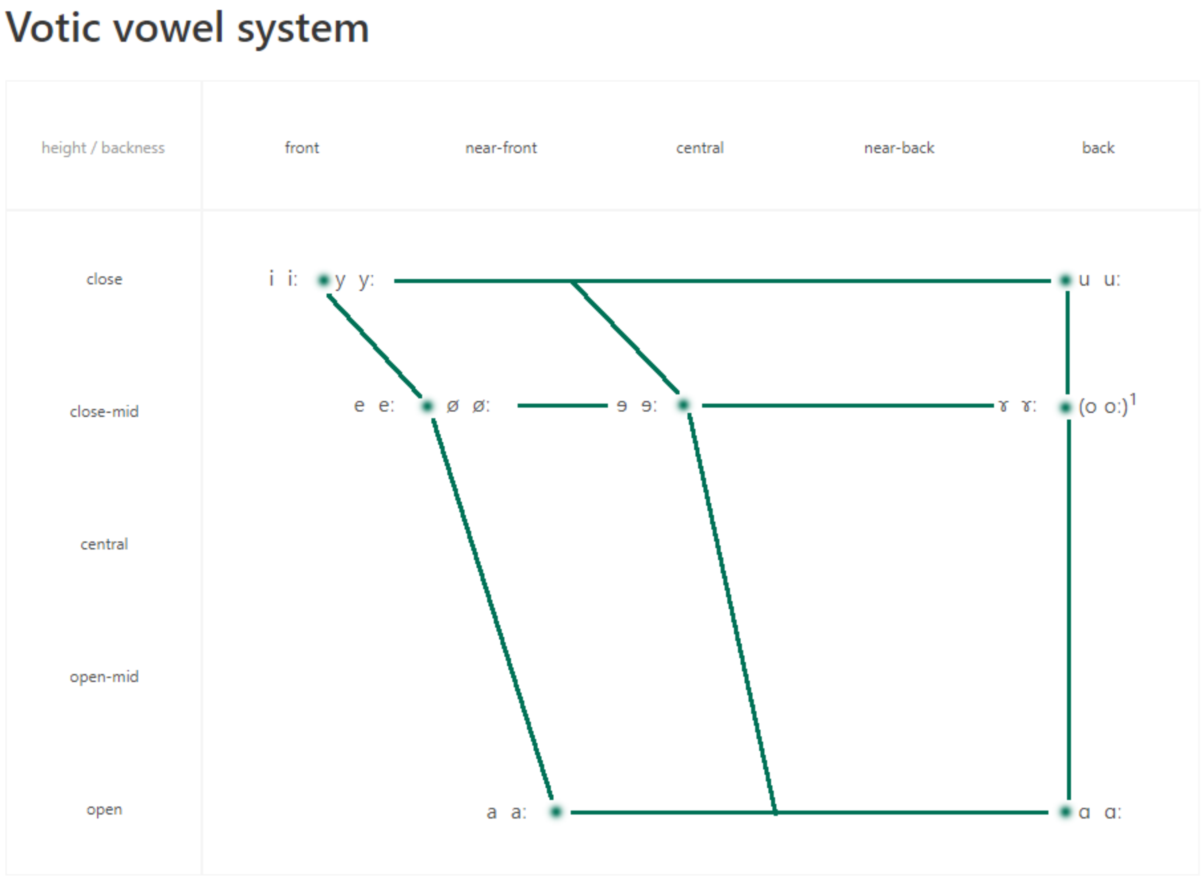
A coordinate plane, which is determined by the first and the second formants, or a trapezoid vowel diagram are often used for representation of a sound system of vowels. However, for phonological vowel systems a table structure is more relevant, as it enables to describe several differential features apart from height and backness. As for consonant sound systems, they are traditionally provided in tables: it is possible to draw a line between different types of features without data loss in this case.

Language description implies a phonology description and a presentation of its sound manifestations. Our goal is to combine the acoustic and the articulatory features with the phonological representation in order to make our description compact and easy to analyze without special work with audio data.

We primarily aimed at a united representation of phonetics and phonology, alongside with user-friendliness, objectivity and descriptive representativeness.

As a solution, we have constructed graphic representations for vocalism (see Figure 8) and consonantism (see Figure 9) that reflect a phonological system with its correct sound manifestations. Each sound is transcribed according to the IPA[5], and the whole classification is synchronized with Russian terms. An optimal representation of the structure within universal consonant and vowel tables has been elaborated, in order to facilitate the comparison of languages.

[5]This resource is available at: https://www.internationalphoneticassociation.org/content/full-ipa-chart.



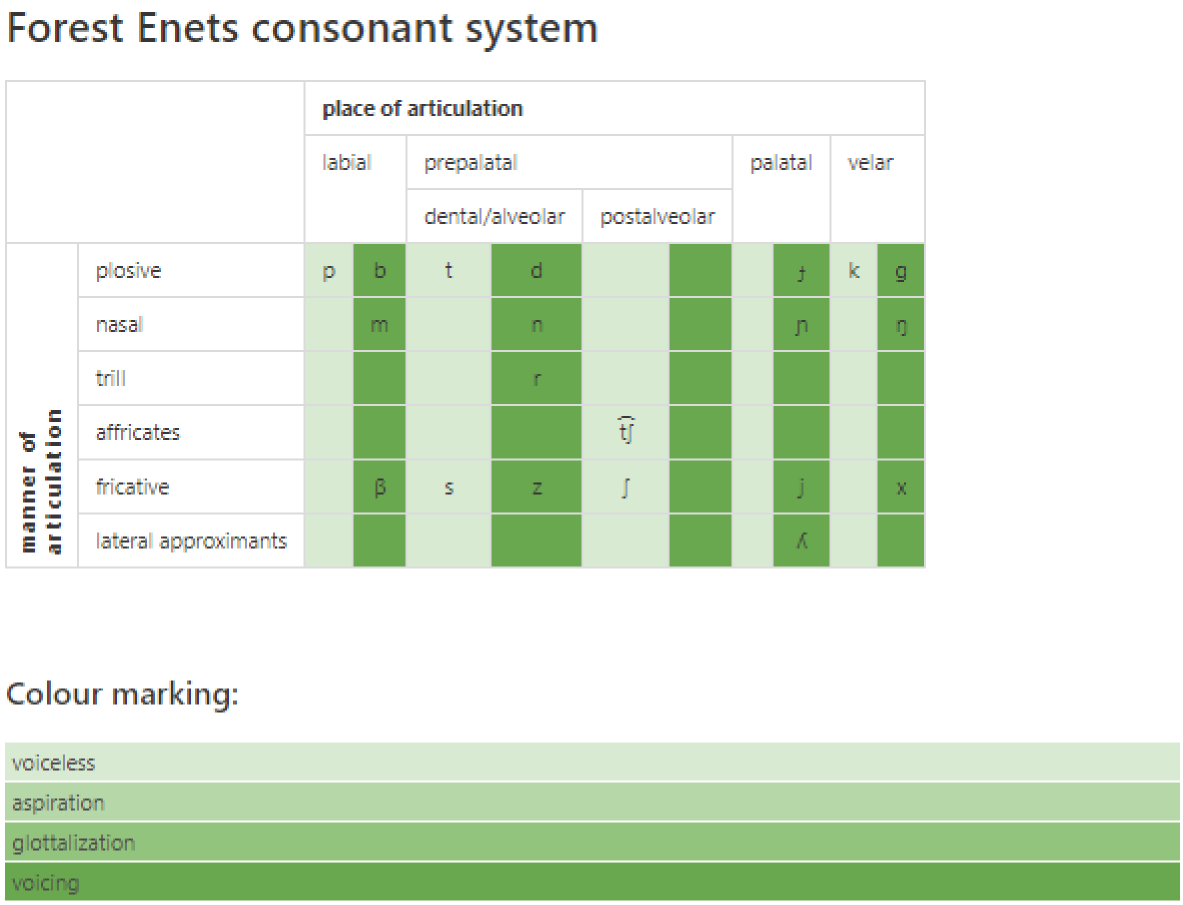
Markus, Rozhanskij [2017]

Figure 8: Votic vowel system

The representation of vocalism is based on filling out an initial matrix with height values in strings and backness values in columns. We provided about two or three slots for each traditional element of classification (height/backness), as the same phoneme in the classification may display different spectrum in various languages. This solution is crucial both for language comparison and for the reflection of phonological processes.

In addition to height and backness, there are secondary articulations significant for phonology: vowel length, labialization, nasalization and pharyngealization. For each characteristic there exists an additional string or a column, which tentatively accounts for the articulation manner and its impact on the spectrum. Thus, pharyngealized vowels are placed in columns to the right of the unpharyngealized ones, and nasal vowels are placed in strings below the non-nasal ones. The length of the vowel is marked when it is relevant: long vowels take place closer to the cardinal locus than their short pairs. The recited secondary articulations are marked by diacritics, while labial vowels have their own symbols in the IPA. Traditionally labials are placed to the right of their illabial pairs.

If the language's vowel system lacks nasalization/pharyngealization/length, the corresponding table cells are removed from the initial matrix. Primary strings and columns (based on height and backness) are preserved, empty cells are visualized as an empty space. It is necessary to distinguish the phonologically similar vowels with different acoustic features in different languages. Then a trapezoid is superimposed on the table: phonemes with the same height value adjoin to its horizontal contours, while phonemes with the same backness value adjoin to its vertical contours, so that each line corresponds to phonologically meaningful height/backness.



Forest Enets consonant system

		place of articulation									
		labial		prepalatal				palatal		velar	
				dental/alveolar		postalveolar					
manner of articulation	plosive	p	b	t	d				ɟ	k	g
	nasal		m		n				ɲ		ŋ
	trill				r						
	affricates					t͡ʃ					
	fricative		β	s	z	ʃ			j		x
	lateral approximants								ʎ		

Colour marking:

voiceless

aspiration

glottalization

voicing

Figure 9: Forest Enets consonant system

The universal table of consonants is defined by the place and manner of articulation as columns and strings respectively. Secondary articulations refer to the place of articulation. In particular, labialization or pharyngealization splits each column into additional subcolumns. Labialized consonants are placed to the left of the neutral column, while pharingalized ones are placed to the right of it. Palatalization is described as follows: subcolumns are usually added to the right of the initial column in case of labial and coronal consonants, while in case of dorsal consonants they are added to the left. Abruptives are put in substrings below each string for pulmonic consonants. The last splitting of columns is done according to the state of the vocal fold (4 subcolumns: voicelessness, aspiration, glottalization and voicing). This level of organization has colour marking.

After the initial universal table is filled out, strings and columns with empty cells are removed from it. Therefore, we end up with a compact table describing the consonant sound and phonological system of a given language.

5. Discussion

The visualization solutions we have adopted for the current version of the website can be further refined. We have already provided some development options for certain sections of the language page. For instance, the visualization of genealogy in the shape of nested circles can be proposed as an alternative representation, its advantages were formulated in §4.1. We intend to search for the ways of modifying this representation so that it could meet our requirements. In particular, it is essential to display the language names at several specified levels, as well as to indicate languages and dialects by means of different colours. For now, these colours are marked up in this scheme by default according to certain nesting characteristics.

In §4.2, we have mentioned several options for representing the areal distribution of the language. At present, we implement the strategy of indicating the settlements with markers on the language page.

In the future, we aim to represent the minor languages of Russia on a separate map located on the main page of the website. A polygonal representation enabling the users to sort and display the languages according to their choice seems to be the most suitable solution for this task. In this case, it will be necessary to provide a possibility of switching from the map to the pages of the corresponding languages.

As regards the dynamics of language usage (see §4.4), there are also prospects for further work in terms of visualization. In particular, it is planned to expand the available data through the researchers' estimations (for those languages that linguists work in the field with), as well as by using the information that can be received from local administrations. These data would significantly specify the information provided in censuses. In addition, we intend to take into account the results of the Russian Census 2021.

As for the visualization of phonetic data (see §4.5), the accumulation of material on a larger number of languages will possibly allow us to further adjust the strategy for representing the systems of vocalism and consonantism. Moreover, it is planned to transform the phonetic tables into interactive ones: clicking on the phoneme symbol, the corresponding audio file will be played.

6. Conclusion

In this paper we considered the methods of language data visualization on the website "Minority languages of Russia" in five areas of concern: genealogy, areal distribution, domains of language usage, dynamics of language usage and phonetic data representation.

Visualization mechanisms were implemented by means of JavaScript tools, as well as available JavaScript libraries. However, the nature of the data dictated the need to modify the standard functionality of JavaScript libraries and the layout embedded in them, so that the representation in each case could adequately meet our tasks. The layout of visual representations is developed with the help of HTML and CSS, namely HTML5 and CSS3. More detailed guidelines on a visual representation of our resource can be found at: https://minlang.site/about#tech-materials.

7. References

[1] H. Kloss , G. D. McConnell (Eds.), The written languages of the world: A survey of the degree and modes of use, volume 1: The Americas, Laval University Press, Quebec, 1978.

[2] P. Padmanabha, B. P. Mahapatra, V. S. Verma, G. D. McConnell (Eds.), The written languages of the world: A survey of the degree and modes of use, volume II: India (2 books), Laval University Press, Quebec, 1989.

[3] H. Kloss, A. Verdoodt, G. D. McConnell (Eds.), The written languages of the world: A survey of the degree and modes of use, volume III: Western Europe, Laval University Press, Quebec, 1989&

[4] G. D. McConnell, Tan Ke Rang (Eds.), The written languages of the world: A survey of the degree and modes of use, volume IV: China (2 books), Laval University Press, Quebec, 1995.

[5] G. D. McConnell, Les Langues Écrites du Monde: Afrique Occidentale, Les Presses de l'Université Laval, Québec, 1998.

[6] G. D. McConnell, V. M. Solntsev, V. Yu. Mikhal'chenko (Eds.), Pis'mennye yazyki mira: Rossiĭskaya Federatsiya. Sotsiolingvisticheskaya entsiklopediya {The written languages of the world: Russian Federation. A sociolinguistic encyclopedia}, volume 1, Academia, Moscow, 2000.

[7] V. Yu. Mikhal'chenko (Ed.), Pis'mennye yazyki mira: Yazyki Rossijskoj Federatsii.{The written languages of the world: Russian Federation.}, volume 2, Academia, Moscow, 2003.

[8] V. P. Neroznak (Ed.), Yazyki narodov Rossii: Krasnaya kniga. Entsiklopedicheskij slovar'-spravochnik {Languages of the peoples of Russia: The Red Book. An encyclopedic dictionary.}, 2nd ed., Academia, Moscow, 2002.

[9] V. Yu. Mikhal'chenko (Ed.), Yazyk i obshchestvo. Entsiklopediya {Language and society. An encyclopedia.}, «Azbukovnik», Moscow, 2016.

[10] O. A. Kazakevich, M. I. Vorontsova, E. L. Kliachko, K. K. Polivanov, Multi-Functional Web-Site "Minority Languages Of Siberia As Our Cultural Heritage", in: Materials accepted for publication on the website of the 19th International Scientific Conference on Computational Linguistics

"Dialogue 2013", Moscow, 2013 (electronic publication). URL: http://www.dialog-21.ru/digests/dialog2013/materials/pdf/KazakevichOA.pdf.

[11] Drupal: an open source content management software written in PHP and distributed under the GNU General Public License. URL: https://www.drupal.org/.

[12] N. Dobrushina, D. Staferova, A. Belokon (Eds.), Atlas of Multilingualism in Dagestan Online, Linguistic Convergence Laboratory, HSE, 2017. URL: https://multidagestan.com, accessed on 2020-11-19.

[13] Leaflet: an open source library written in JavaScript designed for displaying maps on websites. URL: https://leafletjs.com/.

[14] D3: a JavaScript library for creating dynamic interactive data visualizations in web browsers. URL: https://d3js.org/.

[15] Chart: an open source JavaScript library for data visualization that supports 8 chart types. URL: https://www.chartjs.org/.

[16] Datawrapper: a tool for creating interactive maps and diagrams. URL: https://www.datawrapper.de/.

[17] M. P. Lewis, G. Simons, Assessing Endangerment: Expanding Fishman’s GIDS, in: Revue Roumaine de Linguistique/Romanian Review of Linguistics, volume 2, 2010.

[18] MediaElement: a powerful JS/HTML5 audio and video library that creates a unified view for media files. URL: https://www.mediaelementjs.com/.

[19] OwlCarousel: a jQuery plugin, which allows to create sliders and carousels. URL: https://owlcarousel2.github.io/OwlCarousel2/.

[20] J. Saarikivi, The divergence of Proto-Uralic and its offspring. A descendent reconstruction, in: M. Bakro-Nagy , J. Laakso, E. Skribnik (Eds), Oxford Guide to the Uralic Languages, Oxford University Press, forthcoming.

Appendix 1

Language page structure

Language page section	Section content	Technical implementation
Basic information in tag format	The section contains 4 tags: “Native speakers”, “Area”, “Family” and “EGIDS Status” (for more details on the EGIDS scale, see [17]).	Each tag occupies a separate cell in the database, which makes it possible to use the tag as a filter or a criterion for sorting. At the layout level, the tags are packaged in a separate HTML block, which receives a graphic design different from the rest of the text using CSS styles.

Brief information	The section contains a brief introductory text that provides information about areal distribution, number of speakers, dialect structure, traditional lifestyle of the people, etc.	The page contains only a few lines of text and the "read more" button, which unfolds the rest of the text field below. In the unfolded state, the "collapse text" button appears after the text. After clicking on it, the text collapses to its original position. This mechanism is built on JavaScript. The layout is performed using HTML/CSS. The graphical solution intuitively allows the user to understand how the mechanism works.
Genealogy	The section describes the genetic affiliation of the language and its dialect structure. The genealogy is visualized as an interactive diagram.	See §4.1.
Areal distribution	The section provides the areal characteristics of the language, reflected on the interactive map.	See §4.2.
Language contacts and multilingualism	The section describes the languages that are (or historically have been) in contact with the language in question, as well as the multilingualism among its speakers.	The section is technically implemented in the same way as the section “Brief information” (see above).

Language functioning	The section provides the following information about the language: legal status, writing system, language standardization. The domains of language usage are also described here. This section is one of the most important sections of the site.	At the database level, each section element occupies a separate cell, which allows to work with individual data elements, as well as to apply them as filter and sort criteria. The section "Language functioning" is divided into 4 subsections: "Legal status", "Writing system", "Language standardization", and "Domains of language usage". These subsections are represented by tabs, each of which contains a corresponding subsection. The tabs are designed in a separate HTML block, highlighted in color. Inside the block, a panel with the names of four subsections is displayed. These subsections serve as buttons for switching between the tabs. The button corresponding to an open tab is highlighted in the active color. The content of only one active tab is displayed under the button panel. This solution allows to present a large amount of information in a compact way, as well as to immediately provide the list of all the subsections. The mechanism is built by means of JavaScript and CSS. See §4.3 on the domains of language usage.
Dynamics of language usage	The section provides information on the changes in the language usage based on the census data (in the format of an interactive diagram) and on the degree of the language vitality at present.	See §4.4.

Language structure	The section provides brief information on the language structure within 4 main levels (phonetics, morphology, syntax, vocabulary). This section is planned to be expanded in the future.	The section is presented as a grid of four blocks corresponding to the subsections "Phonetics", "Morphology", "Syntax" and "Vocabulary". Each block contains the name of the subsection, a short description, and the "more details" button, which opens the corresponding information in a pop-up window. The pop-up window mechanism is implemented using JavaScript, and the visual solution is implemented using CSS styles. See §4.5 on the phonetic data.
Language research	The section provides information on the history of language research and indicates the relevant experts in this language and the centers where the language research is conducted.	The data are presented in a table with two columns. This format was chosen for compact data representation. The left column contains the following information: a photo (if absent, an abstract image in the form of a human silhouette is used), the full name and the affiliation of the expert, and a link to the expert's personal page. The right column contains a summary of the expert's research activities. This mechanism is implemented by means of HTML/CSS.
Main publications	The publications are grouped into the following sections: grammars and grammatical essays, dictionaries, selected works on certain aspects of grammar, publications of texts in the language, works on sociolinguistics and ethnology.	Information is presented in the "accordion" format used in the web layout for compact data representation. This section contains a list of clickable titles. Only one header can remain open: when you click on any title in this view, the previous open one is closed. This mechanism is implemented using JavaScript and CSS3.

Resources	The section provides links to available electronic resources (corpora and text collections, dictionaries, etc.).	The data are presented as a text field and a table with two columns. This format was chosen for compact data representation. The left column contains the following information: the logo (if available) and the name (with a link to the resource). The right column contains brief information about the resource. This mechanism is technically implemented using HTML/CSS.
Administrative and public support	The section provides information on the administrative and / or public support for the language, including the support from non-governmental organizations and language activists.	The information is presented as a text field and a table with two columns. This format was chosen for compact data representation. The left column contains the following information: a logo (if available) or a photo (in the case of a person) and a name or a full name (in the case of a person). The right column provides a summary of their activities. This mechanism is technically implemented using HTML/CSS.
Data source	The section indicates the experts who provided the data on the language.	The data is wrapped in an individual HTML block and is stylistically highlighted in color. The implementation is built on HTML/CSS tools.

Text	The section contains video and / or audio recordings of the texts with transcript and morphological annotation. Each text is accompanied by detailed metadata (general data about the text, information about the recording, transcript and morphological annotation).	Video files are integrated from YouTube. The website has its own channel. Audio files can be integrated with SoundCloud. A corresponding template is written for this purpose. Audio files can also be uploaded directly to the template on the website in OGG, MP3, or WAV formats. In this case, the library MediaElement.js is responsible for displaying the files [18]. Metadata markup for the texts is presented as separate HTML blocks, stylized in such a way that each element and its caption are distinguished. At the database level, each element has a separate cell. The transcript and morphological annotation are uploaded as PDF files. This block is implemented using JavaScript, HTML, and CSS. SoundCloud and YouTube embed integration tools as well as the library MediaElement.js were used in the development process.

Photographs	The section contains photographs of native speakers, settlements, traditional household items, clothing, etc.	In this section we employ a "carousel", which is one of the well-known ways of displaying images. It consists of multiple slides and navigation elements between them. Each slide contains an image and a caption. Only one slide is shown at a time. There are buttons on the right and on the left sides of it to switch between the slides. There is a panel with dots under the slide. The number of dots corresponds to the number of slides. When a dot is clicked, the user is transferred to the corresponding slide in the array, and the dot is highlighted in the active color. When you click on a slide, the image opens without cropping. Its original proportions are preserved, although it is limited in weight and scale. The mechanism is implemented using the jQuery library owl.carousel.js [19]. The markup and styling are performed by means of HTML/CSS. Image files with .png, .gif, .jpg or .jpeg resolutions and any aspect ratio can be uploaded to the template. The images are then passed through a program that automatically processes them for displaying in the carousel.