Proceedings of the EURALI Workshop @LREC2020, pages 22–27
Marseille, 20 June 2022

How to Digitize Completely: Interactive Geovizualization of a Sketch Map from the Kuzmina Archive

Elena Lazarenko, Aleksandr Riaposov
Universität Hamburg
Hamburg, Germany
{elena.lazarenko, aleksandr.riaposov}@uni-hamburg.de

Abstract

This paper discusses work in progress on the digitization of a sketch map of the Taz River basin – a region that is lacking highly detailed open-source cartography data. The original sketch is retrieved from the archive of Selkup materials gathered by Angelina Ivanovna Kuzmina in the 1960s and 1970s. The data quality and challenges that come with it are evaluated and a task-specific workflow is designed. The process of the turning a series of hand-drawn images with non-structured geographical and linguistic data into an interactive, geographically precise digital map is described both from linguistic and technical perspectives. Furthermore, the map objects in focus are differentiated based on the geographical type of the object and the etymology of the name. This provides an insight into the peculiarities of the linguistic history of the region and contributes to the cartography of the Uralic languages.

Keywords: Uralic languages, language maps, digitization

1. Introduction

The aim of the long-term project INEL (Grammar, Corpora, Language Technology for Indigenous Northern Eurasian Languages)[1] is sustainable documentation and analysis of the resources in highly endangered indigenous Northern Eurasian languages and their varieties (Arkhipov and Däbritz, 2018). One of the tasks pursued by the project is documentation of linguistic data in the Selkup language (Brykina et al., 2021). A significant amount of Selkup data originates from the archive collected by Angelina Ivanovna Kuzmina (1924–2002) in years 1962-1977 (Tučkova and Helimski, 2010). The collection of hand-written notes and audio recordings from the archive has been extensively digitized by INEL and an access to the written part of the archive has been provided via a Kibana Dashboard[2] that allows for the interactive exploration of the materials. The content of Kuzmina's manuscripts is scanned and stored in PDF format and alongside with the TEI P5 compliant XML catalogue ingested into the project's Elastic cluster (Lehmberg, 2020). This provides a web-based access to the manuscripts with the possibility to create complex search queries and set filters based on such criteria as keywords, place of origin, speaker, and many others. While most of the the texts from the PDFs are available not only as scans but also in markup data formats via the published versions of the INEL Selkup Corpus and via a web-based Tsakorpus platform[3], one of the archive notebooks contains data that until now has been available only as a PDF scan. It is a hand-drawn sketch map that covers the region around the river Taz and depicts toponyms and hydronyms of this region. Cartography of languages of Northern Eurasia (e.g. Uralic) is not sufficiently researched and there are blank spaces when it comes to the language distribution within this area (Koriakov, 2020). Therefore this map could be of interest for researchers but since it cannot be used for any kind of analysis in its current form, we decided to create a digital searchable version of it. It will contribute to the information about the geographic objects of the region. Moreover, we pursue a task to transform this sketch map into a language map of the region by differentiation of the objects according to their name etymology in order to visualize the distribution of different languages in this area. This paper describes the steps that have been undertaken so far and the challenges of this task. At the moment, a working version of the digitized map can be found here: https://inel.corpora.uni-hamburg.de/portal/geo/kuzmina/map.html.

2. Background

The sketch map can be found in books 3 and 4 of the first archive volume. It consists of 18 A5 format pages with the total of ca. 580 objects depicted. At least 514 objects are unique, i.e. are marked only once. Vast majority of the names on the map are hydronyms such as rivers, lakes, swamps and smaller water objects. Another big group of objects are land objects, such as settlements and islands. Moreover, there are objects of unclear origin. Each A5 list of the map has a sequential number that allows to build a single image.

[1] The project is funded by the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

[2] https://inel.corpora.uni-hamburg.de/portal/kuzmina/

[3] https://inel.corpora.uni-hamburg.de/SelkupCorpus/search

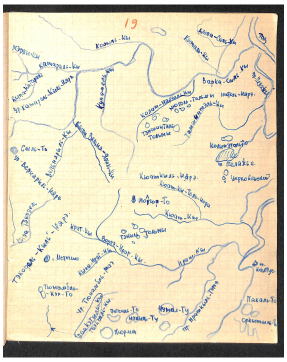

Figure 1: Example of the original sketch

2.1. Visualization challenges

At first sight it seems that the sketch depicts the geographical properties of the region with enough precision to start working with it directly, however already in the beginning we detected following challenges that created obstacles for a seamless and straightforward visualization of the map data with digital tools:

- **Absent coordinates**. No information about longitude and latitude and coordinate system is provided, which makes it impossible to use the sketch map in its original state as a reliable source of geolinguistic data.
- **Ambiguous scaling**. The distribution of the objects is arbitrary and does not follow scaling principles.
- **Inconsistent orthography and incomplete naming**. Some of the toponyms appear in unclear, hard-to-discern handwriting, for others only a part of the name is available. This makes it markedly difficult to find their direct counterparts on the modern maps of the region (see 2.2.2 for examples).
- **Poor naming convention**. A number of objects have abbreviations in front of their names, which pose problems due to the lack of a proper map legend and internal inconsistencies - e.g., о. might stand for either озеро (a lake) or остров (an island). The abbreviation ур. (for урочище) means a salient landmark of any kind, i.e. a swamp, a copse in an open field, a settlement, or some natural border; given such vagueness, we can only assume which objects were thus marked.

Together with scarce amount of precise geographical and geolinguistic data on the Taz basin, the issues listed above present an obstacle to transform the sketch map into a properly georeferenced and scaled map in a straightforward manner. Hence, we were unable to use existing GIS software packages for digitizing hand-drawn maps and had to develop a task-specific semi-structured workflow.

2.2. Workflow steps

The following workflow was developed in order to digitize the sketch:

- Merge the scattered pieces of the original sketch into a single image;
- Identify modern names of the objects represented on the sketch;
- Determine the latitude-longitude coordinate pair for each such object;
- Geovisulalize the objects with known coordinates;
- Classify the objects by type and, where possible, toponyms by language of origin.

The workflow is semi-cyclic, where after the visualization step we go back to step 2 to further improve accuracy of the previously identified coordinate pairs, thus through several iterations improving the visualization itself.

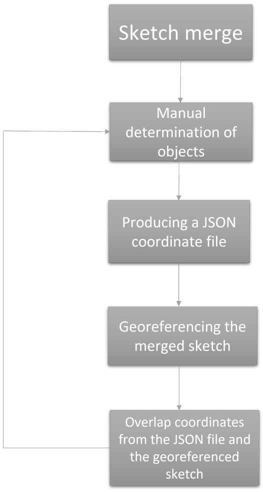
Sketch merge
Manual determination of objects
Producing a JSON coordinate file
Georeferencing the merged sketch
Overlap coordinates from the JSON file and the georeferenced sketch

Figure 2: Workflow steps

2.2.1. Sketch merge

Due to the fact that the archive map is distributed over multiple pages, it was reasonable to merge them together for further georeferencing and scaling; moreover, it was also important for the object identification - this way we could better understand the juxtaposition of the toponyms. To do so, the archive volume pages containing the sketch were extracted from the PDF and saved as separate PNG images.

Some challenges posited themselves forthwith. Namely, there is no instruction left by Kuzmina as to how one was to put the map together; the individual pages had to be aligned with each other in a way resembling a jigsaw puzzle until it "clicked". The settlements of Tol'ka and Krasnoselkupsk (Толька and Красноселькупск respectively in the original) were taken as base points for being quite compact geographically and easily identifiable on modern maps, then the rest of the map followed suit: names of some objects, mostly rivers, were spotted on different pages, providing a reason to place these pages alongside each other; after some trial and error we could trace the flow of the Taz and its tributaries, and the composite image was complete. In retrospect, there was a method to the map's madness - the pages, once ordered, displayed a numbering pattern; in addition to that the corners of some pages have markings consisting of a number from 1 to 5 followed by a Cyrillc letter Л, С or П, the meaning of which initially was a mystery. Apparently the numbers represent the latitudinal dimension with 1 being the southernmost and 5 - its northernmost counterpart, while letters stand for "left", "center" and "right" ("Левый", "Средний" and "Правый" in Russian); unfortunately, they were not of much use when piecing the map together as some "rows" of the composite image lie four pages abreast, thus leaving some pages unmarked, and Kuzmina's notation was not consistent as to which "column" assign as leftmost, rightmost, etc.

The resulting sketch, owing to the original's somewhat arbitrary scaling and no less arbitrary borders between pages, looks quite patchy; it was very instrumental however in allowing us to disambiguate some toponyms for which there are multiple objects with matching names in the general region around the Taz, and place them correctly on the digitized map.

2.2.2. Looking for objects

Mapping the entities one may find on the sketch turned out to be quite a strenuous task, stemming from a number of facts. First, the area around the Taz basin remains relatively poorly depicted on modern go-to sources of geographical data, prompting us to cross-reference a variety of resources such as Google Maps, Yandex Maps, Wikimapia and Wikipedia. Surprising as it may seem, the most fruitful resource was the Tatar Wikipedia where we could find names for many rivers identical or almost identical to what we have on the sketch map even if these rivers have been renamed recently. Second, the names used by Kuzmina in the original sketch displayed a plethora of issues - some toponyms would have multiple spellings (e.g. Поколь-Кы/Покаль-Кы[4]), other would abruptly end in the middle (e.g. Туне... for Тунелькы-Ягарт[5]); on top of that, this lack of rigour on Kuzmina's part is further compromised by name and/or spelling changes of varying drasticity the toponyms have underwent since 1960s. All this considered, we opted for manual lookup of each object from the original sketch, as automatizing the task was not anywhere near possible.

As a preliminary step to get ahold of Kuzmina's data, we comprised a dataset containing a list of all the geographical objects on the sketch, including duplicates and misspellings. From that we started to look for the latitude-longitude coordinate pairs of items on the list via the digital maps mentioned above, beginning with easily identifiable objects (e.g. the settlements of Tol'ka, Krasnoselkupsk, and Sidorovsk), then moving on to cases where some ambiguity arose. As a means to dispel that ambiguity and prepare the data for further processing, we created a custom Google Maps view and placed coordinate markers for successfully identified objects there, settling on using only one marker per object regardless of its type; thus, for watercourses such as rivers only one coordinate pair would be set. Such visual representation of the data allowed us to check whether the placement of an object with a doubtcasting name was indeed correspondent to its position on the sketch.

Unfortunately, we were not able to identify each and every object from the sketch due to the issues mentioned above, leaving 77 of 514 in limbo for the time being. The rest we exported from Google Maps in the KML format for further processing.

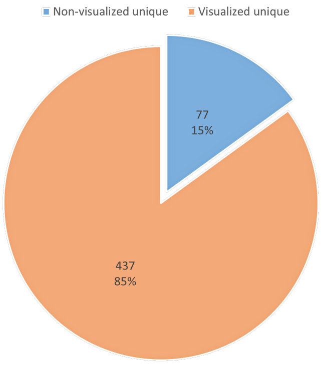

Figure 3: Distribution of identified objects

2.2.3. Technical details

No matter how convenient Google Maps were to collect the objects, it was only an intermediate step as we wanted abstain from using Google Maps services in the long-term perspective in favour of open-source solutions. Therefore for further visualization we settled to use open-source JavaScript library Leaflet[6] that works, among others, with OpenStreetMap[7] layers and is often a first-choice solution for geovisualization. Since Leaflet typically works with the JSON and GeoJSON formats, we transformed the KML data into JSON. However, the resulting JSON file was rife with irrelevant for our purposes remnants of KML data, which were manually removed. The resulting JSON file contains a FeatureCollection object where each geographical object represented as a feature with a list of attributes such as object name ("namemap",

[4] Pokol'-Ky/Pokal'-Ky

[5] Tune..., Tunel'ky-Yagart

[6] https://leafletjs.com/

[7] https://www.openstreetmap.org

"namekuz"), geographical type ("objtype") and name origin ("nameru", "namesel"). Some of the attributes (e.g. "objtype") were inherited from the original KML file, others were introduced later.[8] For testing purposes the fist version of the digitized map only showed the object distribution and the territory these objects covered. This way we could assure the quality of the JSON file, e.g. whether the latitude and longitude of points were correctly transferred from the KML structure and did not get swapped. Later we transformed it into a heatmap that colourfully depicted object clustering in order to get first impressions about the distribution of all the objects. After that we moved to classifying objects as per their type

2.2.4. Further differentiation

As soon as the quality of the JSON file was assured, we moved to more specific visualization tasks stemming from the goal to classify the objects. This was a task with gradually growing complexity. Our working hypothesis, which became a basis for further work on the data, is that geographical distribution of toponyms of different origin provides an overview of how the indigenous languages of the region were spread across the land in prerecorded history, as well as helps to determine areas of possible language contact. For that goal, we decided to group names from the sketch based on their etymology and their type - the idea behind the latter grouping being that, first, water streams such as large rivers flowing for up to 1400 km long in case of the Taz, might present a different distribution properties than compact objects such as lakes; second, we expected a different etymological outcome for settlement names since the Selkups, indigenous people of the region, preserved traditional nomadic lifestyle until well into the 20th century, at which point it is reasonable to expect local toponyms for relatively new-founded villages to display a considerable Russian influence. To begin with, as native speakers of Russian we could recognize and mark all the objects with the names of Russian origin; the respective attribute ("nameru") was also added to the JSON file. This allowed to visualize approximate distribution of the Russian versus non-Russian toponyms in the Taz basin. However, given the fact that the vast majority of the toponyms are of non-Russian origin, it did not provide us with enough information. We continued differentiating remaining toponyms by their language of origin. As the archive consists of entirely Selkup materials, it is highly likely that the source of many toponyms present on the sketch would be Selkup as well; this line of reasoning resulted in the choice of the Selkup language (provisionally omitting distinctions between its dialects) as the next direction of inquiry. To do so, we manually searched for the toponyms and their constituents in Selkup dictionaries (Bykonia et al., 2005; Kazakevič and Budianskaia, 2010), and introduced a respective attribute to the JSON file. Thus the attributes "nameru" or "namesel" would be used depending on the provenance of a toponym. Not only did we attempt to differentiate the toponyms based on their linguistic origin, but also to classify them based on the geographical type of the object. This was equally challenging due to naming convention problems specified in 2.1. To define the geographical type of the object we either evaluated its name compounds (for example, objects ending with "Кы" ("river" in Selkup) were classified as rivers), how it was depicted on the sketch map or based on the information retrieved from the modern map sources. Processing the sketch map data, the grouping we ended up settling for is presented on the Table 1.

Category	JSON "objtype"
Standing water	lake swamp
Watercourse	river creek anabranch
Land object	island settlement tract

Table 1: Types of objects

Each object then received relevant JSON attributes and each category received its own icon.

2.2.5. Putting everything together

From the technical perspective we considered it reasonable that each object type would be rendered separately in order to ensure easy map navigation. Therefore at the moment three overlays rendered from the same JSON file are loaded onto the map to display all the objects. Another layer that is also rendered from the same JSON file is the aforementioned heatmap; however, at the moment it does not appear on-load unlike other overlays and can be switched on if needed. Given the fact that many objects received only tentative coordinates and we were not fully sure about their position, to make the visualization more precise we intended to create another overlay from the sketch map. At first we needed to bring the PNG image to a state where it can be cut into layer tiles with coordinate information embedded. To do so, we used QGIS software package (long-term release version 3.22.4)[9]. One of its functions - georeferencing - allows to assign coordinate points to analogue maps, sketches, images, etc. We uploaded the PNG file to the georeferencing window and chose several objects, the coordinates of which were defined precisely. After assigning coordinates to these points, we ran a georeferencing tool with the transformation type "Thin Plate Sline" that would

[8] See 2.2.4 for further discussion of the introduced properties and types used.

[9] https://qgis.org

resample the data with the next neighbour method and use EPS:4326-WGS 84 cordinate system as the target one. We opted for Thin Plate Sline due to the fact that it can be used for images with unclear or absent scaling and no latitude-longitude information, whereas other transformation types are more suitable for analogue maps with a known coordinate system. This method requires to assign at least ten latitude-longitude pairs in order to calculate a georeferenced map. By doing so we obtained a georeferenced LZW-compressed TIFF image that was further cut into layer tiles with the maximum zoom depth of 10. Resulting tiles were overlayed on the existing Leaflet map together with the other overlays produced from the JSON file. However, the georeferncing task did not stop here. After bringing together icons from the JSON file and the sketch map tiles, we noticed several severe inconsistencies that were caused by mildly misleading orthography on the sketch map which, in turn, led us to assign coordinates to wrong objects. We updated coordinate data, purging errors from the JSON file, and afterwards introduced these points to the georeferencer and re-ran the calculations. The most up-to-date tile cut was based on 15 objects, such as lakes, islands and settlements. In order to make the map more navigable for users, we introduced features as grouped layer control where one can turn on and off each overlay separately, and on-click pop-ups displaying the name of an object in both the modern and Kuzmina's spelling. Moreover, there is a possibility to search through the whole FeatureCollection of the JSON object, allowing users to quickly find a toponym of interest.

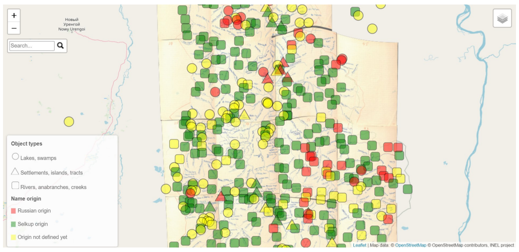
+
−
Новый
Уренгой
Nowy Urengoi
Search...
Object types
Lakes, swamps
Settlements, islands, tracts
Rivers, anabranches, creeks
Name origin
Russian origin
Selkup origin
Origin not defined yet
Leaflet | Map data: © OpenStreetMap © OpenStreetMap contributors, INEL project

Figure 4: Example of the digitized map.

3. Conclusion and future work

We built the first digitized variant of the sketch map from the archive of Angelina Ivanovna Kuzmina. After evaluating the quality of the source material and the challenges that come with it, we developed a data-specific workflow. The current version of the digitized map provides a good overview of the toponyms of the Taz river basin by the means of integration of modern digital maps and the original sketches. So far we have been able to assign geographical coordinates and visualize 443 objects, at least 437 of which have unique names. We have noticed that Russian toponyms is only a minor group with currently 63 visualized objects, clustering in the south-eastern and northern parts of the covered region. The differentiation of the objects of Selkup origin is still ongoing: at the moment 212 objects have been determined to emanate from Selkup. However, it is already visible that Selkup toponyms build the biggest group in the region. As the work on this visualization is still going on, we do not exclude the possibility of more changes being made to both GeoJSON and sketch map tiles. This will include deeper linguistic differentiation of the toponyms: picking out the remaining Selkup names and distinguishing them based on the dialect, as well as looking for toponyms etymologically coming from languages other than Selkup and Russian, e.g. from Evenki and Nenets. By doing so, we will turn our data into a language map and will be able to test our working hypothesis. Naturally, we also would like to bring more clarity into the remainder of the objects yet to be visualized: our task concerning these toponyms is to find coordinates and classify the objects by type. Moreover, as we find new points and double-check the existing ones, we intend to make the sketch overlay as finely-tuned and georeferenced as it can get, given irregular scaling of the original.

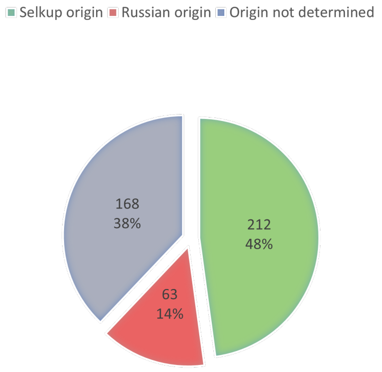

Figure 5: Etymology of the identified objects

4. Bibliographical References

Arkhipov, A. V. and Däbritz, C. L. (2018). Hamburg corpora for indigenous Northern Eurasian languages. *Tomsk Journal of Linguistics and Anthropology*, 21(3):9–18.

Bykonia, V. V., Kuznetsova, N. G., and Maksimova, N. P. (2005). *Sel'kupsko-russikii dialektnyi slovar'*. Tomsk State Pedagogical University, Tomsk.

Kazakevič, O. A. and Budianskaia, E. M. (2010). *Dialektologičeskii slovar' sel'kupskogo iazyka: Severnoe narečie*. Basko, Ekaterinburg.

Koriakov, Y. B. (2020). Kartografirovanie ural'skih yazykov. *Acta Linguistica Petropolitana. Trudy in-*

stituta lingvisticheskih issledovanii, 3(XVI):169–183.

Lehmberg, T. (2020). Digitale Edition des Kuzmina Archivs. *Finnisch-Ugrische Mitteilungen*, 44:123–130.

Tučkova, N. and Helimski, E. (2010). *Über die selkupischen Sprachmaterialien von Angelina I. Kuz'mina*, volume 5 of *Hamburger sibirische und finnougrische Materialien*. Institut fur Finnougristik/Uralistik der Universitat Hamburg, Hamburg.

5. Language Resource References

Brykina, Maria and Orlova, Svetlana and Wagner-Nagy, Beáta. (2021). *INEL Selkup Corpus*.