Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

HighTech and Innovation Journal

Vol. 4, No. 1, March, 2023

New Technologies and Innovative Solutions in the Development of Multimedia Corpus of Mezen Robinsons Texts

Tatiana V. Shvetsova [1*], Veronika E. Shakhova [2], Svetlana A. Dulova [2]

[1] *Department of Literature and Russian Language Studies, Northern (Arctic) Federal University Named after M.V. Lomonosov, Russian Federation.*

[2] *Department of Planning and Support of Scientific Projects, Northern (Arctic) Federal University Named after M.V. Lomonosov, Russian Federation.*

Received 08 December 2022; Revised 14 February 2023; Accepted 23 February 2023; Published 01 March 2023

Abstract

Objective: New Technologies and Innovative Solutions in creating a multimedia corpus of texts about the "Mezen Robinsons" aims to preserve the memory of an event that occurred in the 18th century and to study the history of Spitsbergen development. This article presents a multimedia corpus of Russian-language texts about the "Mezen Robinsons" written in 1766–2022. Observations show that the history of the survival of the Mezen hunters on Edge Island in 1743–1749 has repeatedly attracted the attention of specialists from various fields of knowledge: historians, archaeologists, publicists, professional writers, translators, etc. The corpus unites texts, audio, video, and multimedia resources. *Methods:* continuous sampling was used to collect the material; when analyzing and describing the data, we applied a descriptive method, a biographical method of studying literature, statistical data processing, philological analysis, observation, assessment, and corpus modeling methods. *Findings:* the methodology and technology of building an independent multimedia corpus, its architecture, and its design are described. *Novelty:* the multimedia corpus is a contribution to the development of a new approach to studying the subjectology of Russian literature. *Practical significance:* the findings can become the basis for studying the biographies and creativity of various authors who built their works on the plot of the Mezen industrialists and for further comparison of various interpretations of one event from the history of the development of the Arctic.

Keywords: Corpus Linguistics; Multimedia Text Corpus; Mezen Robinsons; Corpus-Based Research.

1. Introduction

Modern technologies and innovative solutions provide significant practical contributions to the creation of a multimedia text corpus. The use of audio and video materials, interactive elements, hyperlinks, and other tools enriches the text and makes it more accessible to the user. One example of new technologies is speech recognition, which allows automatic conversion of audio files to text, making it easier to work with large volumes of information. The use of software for creating interactive elements such as slideshows or graphic diagrams is also worth noting. It helps the user better understand the material and remember it. Some innovative solutions include the use of artificial intelligence to create synthesized speech or translate text into various languages. This greatly speeds up the process of working with multiple language versions of the same material. In general, new technologies and innovations make it much easier to create a multimedia corpus of text, making it more interesting and accessible to users.

A multimedia corpus of texts is a collection of various types of texts, such as written and oral materials, audio, and video files, which are used for linguistic research. They contain information about the language (its use and variability)

* Corresponding author: shvecova_tatiana@mail.ru

http://dx.doi.org/10.28991/HIJ-2023-04-01-07

in different sociocultural contexts. Multimedia text corpora provide access to a variety of natural language resources, including spelling, grammar, and lexical information. In addition, they allow linguists to study linguistic phenomena in the context of actual language use.

The use of multimedia text corpora provides the possibility of conducting a detailed analysis of various language forms and structures. For example, changes in vocabulary can be studied, or the differences between spoken and written language can be compared. Multimedia text corpora are widely used in linguistics. They are used to create dictionaries, grammars, and other resources that can help with language study. In addition, they are used to determine the national and regional characteristics of the language as well as the sociolinguistic and cultural aspects of linguistic behavior.

According to contemporary researchers, "the growing application of corpus-based research can be attributed to the fact that representative corpora are useful in obtaining quantitative data on the units of analysis and answering many questions about the texts, storylines, and key topics" [1]. Developing various text corpora has been a trend in the humanities for the last decade. Modern philological scholarship presents examples of text corpora united by the same genre [2], same author [3], or same subject [4]. The corpus of texts is a generic concept of many quite diverse objects. These objects tend to have a common generic name.

Multimedia is the sum of technologies that allow the computer to input, process, store, transmit, display, and output such data as text, graphics, animation, still digitized images, video, sound, and speech. Multimedia corpora are being created on the basis of systematization, generalization, and computer representation of a large variety of thematic and multimedia materials reflecting the most diverse aspects of the state and development of national languages [5-8] and literature [9]. Such corpora enable the placement of photographs, video clips, audio recordings of dialect speakers' speech, etc., and most importantly, reference and historical materials. Telecommunications technology, which provides wide access to sites for users with different levels of training, plays a special role in the development of this new area.

Multimedia corpora are mainly used in linguistics for the analysis of dialog patterns and the relationship between speech and nonverbal communicative means (gestures, facial expressions, eye movements, etc.) [10]. Such corpora are convenient for teaching languages and literature [11], in translation and interpretation studies [12], and in age psychology. One of the modern directions of research is the application of corpora for creating computer systems when the behavior and communication of real people are analyzed by videos and transferred to computer characters or to robots interacting with humans [13].

The multimedia corpus is a new tool for the study of genre and discourse; it provides an opportunity to analyze oral performance, such as poetry, and compare different interpretations of texts; it also allows us to compare the written text with its sound, opening up great opportunities for verse studies and studies of poetic syntax and semantics. Despite widespread development in this direction, methodological developments in this area are sparse.

1.1. Aims and Goals

One of the most important innovative areas in the development of modern philology is the inclusion of multimedia methods and tools in its toolkit, with a simultaneous transition to interactive principles for using works of fiction. Our research goal is to create a set of machine-readable texts that will present the variety of the phenomenon under study with maximum objectivity and give an unbiased picture of this phenomenon in the practice of native speakers of the Russian language. The phenomenon being studied in our case is the embodiment of Russian fisher folk wintering on an island of the Spitsbergen archipelago for 6 years and 3 months in different discourses. The task of the corpus, as a unique verbal unity, is to give a picture of the representation of the mentioned episode in the Russian artistic, journalistic, and scientific discourse of the 18th–21st centuries.

The criteria for selecting texts for the corpus are: a common elementary plot in all texts; mentioning the words of the "thematic grid Arctic Robinsonade." The task of the multimedia corpus of texts about Mezen Robinsons is to reflect the existence of historical and cultural phenomena in public speech practice. The multimedia accompaniment of the texts in the corpus is chosen to provide factual, historical, and culturological commentary on one or another lexeme, which sets the topic of the Mezen Robinsons. The focus of the study is on the aspects and principles of the multimedia corpus structure in relation to the formation of Mezen Robinson's folk narrative.

1.2. Computer Innovations Used for Processing Texts in Linguistics

Modern linguistics is impossible without computer technologies. One of the most important tasks for them is analyzing and studying language processes. The most important innovations are software tools for automatic analysis and text processing. Such tools include, for example, morphological and syntactic analyzers that allow the determination of the grammatical structure of a sentence.

Other useful innovations are software tools for creating text corpora—large collections of texts unified by the same topic or the same genre. The text corpora approach can be used for studying various linguistic phenomena (lexicon, writing style, and so on). Another computer innovation is editing software, which provides quick and convenient formatting and editing of texts according to user requirements. In general, computer innovations are undoubtedly accelerating all types of text processing that can be required within the tasks of linguistic research.

1.3. Creating a Text Corpus

Nowadays, creating multimedia text corpora is one of the most important tasks for both linguistics and computer science. This type of corpora allows researchers in all scientific areas to access multiple data and contributes to the spheres of machine learning and AI. The project of creating a multimedia text corpus of "Mezen Robinsons" is an example. The rationale for creating this multimedia text corpus is to preserve unique local cultural information and provide an opportunity for efficient studies in the future by linguists investigating the texts devoted to the historical event of "Mezen Robisons" and therefore unified by certain topics.

1.4. Computer Technologies used in Corpus Linguistics

Corpus linguistics is a scientific discipline that studies natural languages using corpora—sets of texts, optionally presented in digital collections. It requires analyzing great data arrays and text volumes; therefore, applying computer technologies in this sphere is extremely useful.

One of the main technologies used in corpus linguistics is machine learning. It allows creating models for automatic processing and classification of texts according to various parameters, such as word frequency or writing style. Another widely used tool is part-of-speech analyzers. They are engaged in parsing sentences into their constituent elements and determining the grammatical characteristics of each word (part of speech). Computational lexicography, as another example, is a distinctive area of computer technologies in corpus linguistics that helps create dictionaries. Finally, the data visualization technologies are worth noting. Tools for creating graphs and charts can help illustrate and visualize the results of the analysis of text corpora.

1.5. The Research and Product Design

Creating a corpus of texts requires bringing them into a machine-readable format. This requires converting journals and newspapers downloaded from the public domain, sometimes in pre-revolutionary orthography, into a suitable text format. This work has been done: all the source materials have been converted from pdf- or html-format to Word document format. The resulting set of texts is posted on the hosting site *"mezrob29.ru"* [14] in the free public domain. The total number of documents in the published version of the dataset was 61 (as of December 2022), with 790 typewritten A4 sheets.

The technological map of multimedia product development includes the following stages (see Figure 1):

Stage 1: Development of the project;

Stage 2: Collection and preparation of necessary materials: scanning of source materials, converting them into digital form;

Stage 3: Selecting software;

Stage 4: Development of the layout and design, development of the structure and content of the thematic blocks: design of the database (DB);

Stage 5: Filling the database with data, synthesizing the layout of the multimedia product;

Stage 6: Development of a user guide for teachers and for students;

Stage 7: Placement of the database on the Internet;

Stage 8: Design of the search interface.

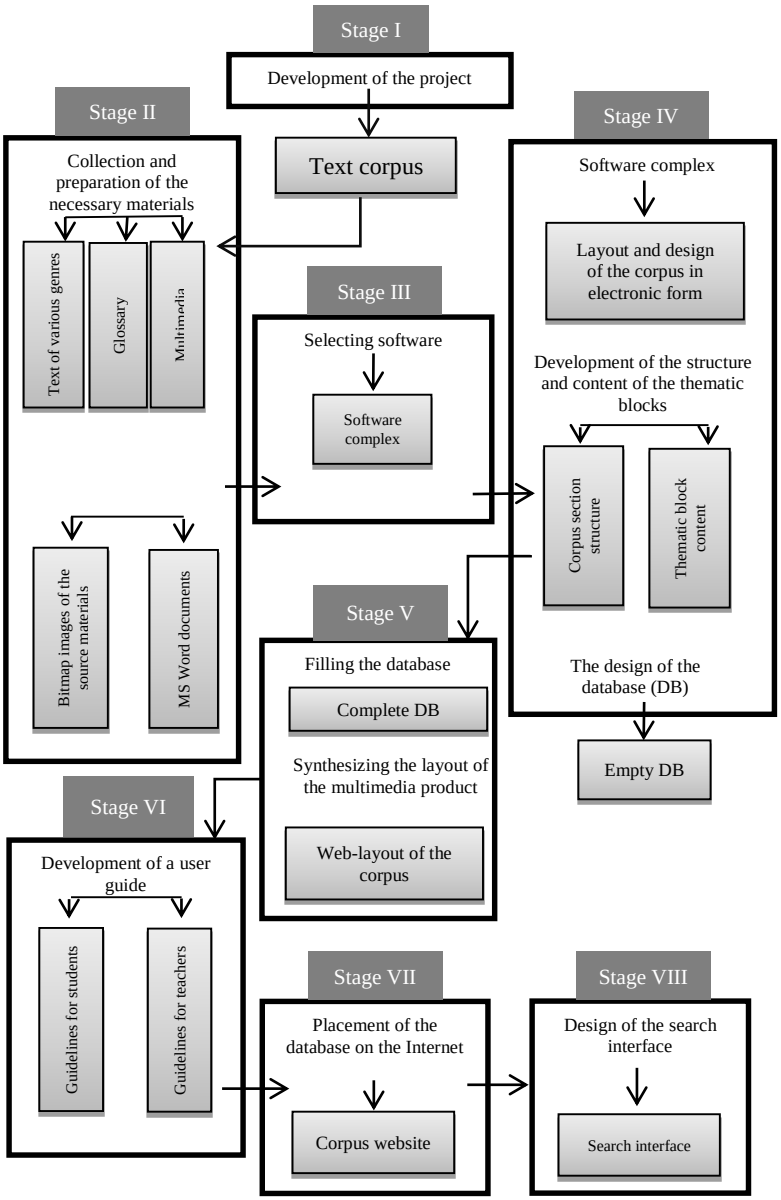

Figure 1. The general technological diagram of the project to create a multimedia corpus of Mezen Robinsons texts

2. Methods

In the first phase of the project, extensive work was carried out on the collection of material for the corpus of texts, namely:

- Studies in the archives of Arkhangelsk and St. Petersburg to attribute the manuscript "Historical Description of the Journey to Spitsbergen in 1743–1749 of Four Mezen Sailors: Alexey and Ivan Khimkov, Stepan Sharapov and Fyodor Virugin" [15], and to analyze the content and historical and cultural circumstances of the origin of the said manuscript; including research of the antique documents, such as 18th century's atlas of the Arkhangelsk Province.
- Studies in the archives of the local history museums in Arkhangelsk and Mezen.
- Obtaining acquaintance with the family tree of Mezen inhabitants who were in distress on Edge (Edgeøya) island, meetings with descendants of helmsman Alexey Khimkov; the meetings were documented and recorded on audio and video media.
- Work with collections of the Dobrolyubov Regional Library and K.S. Badigin Inter-Settlement Library in Mezen; making search queries in the Russian State Library, the Russian National Library, and the National Library of the Republic of Karelia; analysis of information in the WorldCat bibliographic database.

The collected printed materials were digitized; texts from 19th century magazines and newspapers were digitaliZed into MS Word format: *Syn Otechestva* (Son of the Fatherland), 1822; *Severnaya Pchela* (Northern Bee), 1846; *Arkhangelskiye Gubernskiye Vedomosti* (Gazette of the Arkhangelsk Province), 1846; *Rus*, 1846; *Zhurnal Dlya*

Chteniya… (Magazine for Reading to Cadets of Military Schools), 1846; *Detskoye Chteniye* (Children's Reading), 1871; *Drug Naroda* (A Friend of the People), 1876; *Arkhangelskiye Gubernskiye Vedomosti* (Gazette of the Arkhangelsk Province), 1896; *Russkaya Zemlya* (Russian Land), 1899. In total, 23 texts that have not been republished in the 20th or 21st century (790 pages of A4 format) were digitized.

Online resources were scanned using keywords ("Mezen Robinsons", "Polar Robinsons", "Arctic Robinsons", "Pomor Robinsonade", "Helmsman Alexey Khimkov", and names of other persons involved in history) through the following search engines: Yandex, Google, Rambler, Lycos, Bing, and Yahoo. In total, 79 resources were identified.

From April to October 2022, an online survey was conducted in the VKontakte (VK) social network to determine the extent of familiarity with the plot among network users living in Mezen and Mezen District. The questionnaire was distributed through the "VKontakte" social network and by random daily mailing to those users who indicated Mezen as their geolocation. A total of 510 questionnaires were collected.

Field work has been organized. During the expedition to Mezen, the opportunity arose to communicate with local people about the preservation of the story of the six-year survival on Spitsbergen in the 18th century in the cultural memory. 32 oral narratives were recorded and 15 audio recordings were compiled and introduced in our corpus. The results of the expedition were presented in the form of a report at an all-Russian conference.

At this (first) stage, the preparation and creation of the text corpus was an important component of the work. The following tasks were accomplished:

1) Preparation of the documents in three stages:

- Documents in an archive or library were scanned or photographed; scanned copies were ordered from libraries;
- The texts of the following types were manually re-typed in MS Word format: works of fiction, newspaper and magazine publications, and scholarly journal articles;
- Web publications were converted into the MS Word format.

2) Preparation of texts for populating the corpus:

- Compiling a text ticket with source metadata;
- Marking up a text for the sake of a contextual search;

3) Gathering information about the authors of the texts.

4) Compiling a glossary of keywords and tags.

5) Research specific literature to elaborate commentaries for particular words.

6) Selecting multimedia (collection of videos, pictures, and audio files).

7) Choosing a domain and hosting.

8) Preparing the design of electronic product.

7) Uploading documents to the system.

An electronic corpus of texts [14] has been created; it comprises five sub-corpuses: works of fiction, periodicals, scientific-journalistic, Web, and oral texts. The texts are digitized and converted to MS Word format in modern orthography. Each source in the corpus is meta-textually tagged. A glossary has been compiled. The primary commentary of lexemes forming the plot about Mezen Robinsons has been prepared. Illustrating multimedia were selected.

Another important element of the research is to find possibilities for combining scientific and educational tasks; identification of didactic ways of using the corpus in schools. To address this matter, a visual novel was developed as a way to study the texts of the 18th century. Then, several chapters in a multimedia tutorial for organizing a course on Russian (Native) Literature (namely the works of writers K.S. Badigin and S.B. Radzievskaya) on the Tilda platform were prepared. In addition, the concept and scenario of a computer game based on the plot are under development.

2.1. Methods of Data Analysis

The Adventures of four Russian sailors brought to the island of Ost Spitsbergen by a storm, where they lived for six years and three months by Le Roy [16] is a relatively small text first published in 1760 that spawned extensive literature, including research and commentary works, artistic responses, and fantastical interpretations. Particularly prominent in this volume are retellings and works of fiction based on Le Roy's story.

The multimedia corpus of Mezen Robinsons texts is a large corpus of Russian written texts of various genres from

the 18th century to the present day.

A corpus is a collection of texts based on single idea that unites the texts (topicality, author, location etc.). In this case, it is a historical episode of the survival of the Mezen people on Spitsbergen between 1743 and 1749. For selection, the texts were categorized. The story was selected according to parameters such as the source of the text, the type of publication (journal, book, Internet publication), time of publication, genre, volume, and language.

The collected texts belong to different genre categories: retelling ("Son of the native land"), translation from an old German journal ("Northern Bee"), essay (A. Zubkovsky), fable (O. Belomorsky), short story (K. Badigin, S. Radzievskaya), historical novel (Z. Davydov), translation (M. Arkhangelskaya), and review (V. Popov).

These works were written at different times, though there are chronological localizations: starting from Le Roy's life period in 1760; journal editions in Russia (1822, 1846, 1864–69); collections (1899, 1900); book editions: 1933 (Z. Davydov), 1955 (Le Roy and K. Badigin), S. Radzievskaya (2016, 2021), O. Scherbatov (2020). Since 2000s, there are also emerging relevant publications on forums, blogs, PDF versions in digital libraries, and magazines. In the journal versions of the 19th century, the text has different titles, namely, "Adventures," "Journey," "Voyages," "Disasters of Russian Sailors," which is of interest for scholarly discussion.

An important selection factor is the language of this work. For the most part, we focused on the Russian-language versions. However, the fact that the story originally appeared in German and was then translated into various languages, including Russian, gives reason to turn to versions in foreign languages as well as works by foreign authors [17].

The decision to fill the corpus orally or in writing was fundamental. Our corpus included documents digitized and converted into the MS Word format and audio recordings collected from the residents of Mezen. The purpose of the audio recordings was: 1) to determine to what extent the story of the Mezen fisherfolk is preserved in the cultural memory of our contemporaries, 2) to see what elements of the story people reproduced in their oral narratives (for example, women reproduce the episode about how A. Khimkov's wife, seeing her husband alive and unharmed, fainted and fell into the water; men insist that the crew of sailors was unprofessional: they were "drunks who took on debt, got a job on the ship, and went to the Arctic to work off their debts"). Some respondents transposed Mezens' story to their personal seagoing experience.

In its current state, the corpus contains texts with more than 200 thousand words in total. Extralinguistic factors such as the authors of the texts (their gender, age, profession, and nationality), the place, the subject matter, the date of publication, age, and size of the intended audience, etc. were considered when the texts were put into the corpus.

2.2. Methodology for Making a Corpus

This research draws on previous corpus-based studies built on the premise that collecting and analyzing large numbers of samples of discourse is an effective tool for understanding how the Russian language evolves, for example: “Dream Story Corpus” created in Russian State Russian State University for the Humanities [18]; “St. Petersburg Corpus of Hagiographic Texts” [19]; “Lived through: Personal stories in an electronic corpus of diaries and memories: a project of the European University” [20]; “Verbatim: Politics and Literature. A Digital Archive of Literary Organizations, 1920–1930” [21]. By analogy, the authors have created a corpus designed to understand how a certain plot develops synchronically and diachronically, to analyze the specific perception of historical fact at different stages of the formation of the Russian historical and literary process, to highlight structural elements and to understand the semantic overtones of the plot, characteristics for a certain historical period, and a certain picture of the world.

The texts compile the multimedia corpus of Mezen Robinsons are intended for reading and studying as there are quite a few of them that have not been republished. The task ahead is to process these texts specially — to introduce the necessary information — markup, summaries, and design a search interface. After the corpus has been processed, any necessary information (the date, the place, length, authorship, the use of a particular word or grammatical structure) may be searched in it. Special programs — concordances — are used to process the information. They search the text in the same way as engine searches do for information on the Web and generate a concordance, i.e., a list of all contexts in which a word or phrase occurs in the text under study.

2.3. Input Data

Analysis of the information on the Internet showed that users post/repost in their social networks the real story of the six-year wintering of Russian "Robinsons" on one of the islands of the Spitsbergen archipelago and upload digitized books about it to children's e-libraries, the LiveJournal portal, etc. The history of the narrative is presented in the following section.

3. Results

3.1. Composition and Structure of the Corpus

The selected texts are arranged in modules according to the types of discourse: fiction texts, documentary texts (newspaper and scientific papers), media texts, and everyday speech.

The electronic multimedia web-resource contains video, audio (sound files), photos, and textual information. Options such as reference to audio files (speech of Mezen inhabitants), availability of images (details of Pomor ship, details of Mezen manufacturers' clothes, details of landscape and relief of the island), and hyperlinks, etc. are implemented.

This case enables the user, at his request, to obtain information about Pomor culture and way of life, the structure of a Pomor ship, Mezen industrialists' routes in the 18th century, Mezensky dialect, and human survival possibilities in extreme Arctic conditions. The designed multimedia resource is planned as a reliable tool for studying the history of literature, literary local history, and literary geography in the conditions of total "reading crisis."

3.2. A Sub-Corpus of Fiction Texts in Russian

The corpus of fiction texts is a corpus of Russian written prose and verse works created between 1762 and 2022. The texts are presented in Russian in modern orthography. This corpus includes, in certain proportions, various genres (essays, fables, novels, and translations into the Russian).

The corpus may be divided into two main arrays with their own features: modern written texts (early 20th - early 21st century) and early texts (mid-18th - late 19th century). Formally, the boundary between these arrays is not drawn, and by default, they are searched simultaneously.

3.2.1. Texts of Early 20th - Early 21st Century in the Corpus

A corpus of fiction texts with contextual markup forms the core of the main corpus. This corpus includes various types of texts representing the contemporary Russian literary (written) language:

- Fiction prose of various genres and trends;
- Memoirs and biographical literature;
- Journal papers;
- Newspaper entries;
- Academic texts.

The sources of the texts included in the Corpus for published books, magazines, and newspaper texts are, as a rule, their electronic versions.

3.2.2. Texts of the Mid-18th and the Late 19th Centuries in the Corpus

The texts of the mid-18th and the late 19th centuries in the Corpus represent various genres of prose (fiction, journalism, archival document). For this period (up to the end of the 19th century), translated texts can be included in the main corpus.

Texts originally written and/or published in the old orthography (before 1918) are more often given in the post-reform orthography. Multiple texts are included in the corpus based on original editions without preserving the orthography. The volume of the collection of texts in the pre-reform orthography as of 2022 is more than 200 000 words.

This corpus includes texts, reproduced from the 19th-century journals: *Syn Otechestva* (Son of the Fatherland), *Severnaya Pchela* (Northern Bee), *Arkhangelskiye Gubernskiye Vedomosti* (Gazette of the Arkhangelsk Province), *Rus*, *Zhurnal Dlya Chteniya…* (Magazine for Reading to Cadets of Military Schools), *Detskoye Chteniye* (Children's Reading), *Drug Naroda* (A Friend of the People).

3.3. A Sub-Corpus of Newspaper Entries

The newspaper corpus comprises articles beginning from 1876 (the newspaper "Drug Naroda") to 2022 (the newspaper "Arkhangelsk").

The corpus of newspaper texts includes texts of printed newspapers and magazines as well as digitized newspapers: *Arkhangelskiye Gubernskiye Vedomosti*, *Pravda Severa*, *Drug Naroda*, *Mayak Kommunizma*, *Sever*, *Pravda Severa*, *Arkhangelsk*, *Pomorskaya Stolitsa*, *Russky Vestnik Spitsbergena*, etc. The annual addition to the corpus was planned to continue.

3.4. A Sub-Corpus of Scientific Journalism

The sub-corpus comprises papers from scholarly journals on issues related to the analysis of genesis, poetics, and the functioning of works of fiction uploaded to the corpus. These are papers in traditional format, electronic scientific publications, monographs, collections of scientific articles and conference proceedings, and scientific publications in private blogs. These issues are in focus:

- Biographies of the participants of the voyage to Spitsbergen in 1743–1749 and the persons connected with it: the navigator Alexei Khimkov, Khrisanf Khimkov, Fyodor Sharapov, Stepan Verigin, Amos Kornilov, M.V. Lomonosov, P.-L. Le Roy, Solomon Vernizober, the Evreinovs, P.I. Shuvalov, etc.;
- Biographies of the authors of the publications;
- The history of the creation of a particular work;
- The history of the sealing and whaling industry in the Russian North;
- Archaeological excavations on Spitsbergen;
- The history of Spitsbergen's development;
- The history of travels and voyages to Spitsbergen by scientists and explorers from various countries;
- The history of polar expeditions;
- The study of Pomors' settlements on Spitsbergen;
- Equipment of Chichagov's secret expedition;
- The study of the Atlas of Arkhangelsk Province.

3.5. A Sub-Corpus of Internet Texts

This sub-corpus is a collection of Internet publications from the LiveJournal, from online publications (the online edition of the newspaper *Komsomolskaya Pravda*), from the official WEB-pages of various organizations, institutions and electronic libraries, posts on social networks, notes on forums. A survey of the Internet space was conducted using keywords ("Mezen Robinsons," "Polar Robinsons," "Arctic Robinsons," "Pomor Robinsonade," "helmsman Alexei Khimkov" and other persons involved in history) in the search engines Yandex, Google, Rambler, Lycos, Bing, Yahoo. According to preliminary estimates, there are about 40 pieces. The publication range is 2007–2022.

3.6. An Oral Sub-Corpus

The subcorpus of oral texts implies the inclusion of sounding texts in Russian, recorded on the territory of the ancestral residence of the Mezen Robinsons (Mezen District, Arkhangelsk Region). The corpus now has 15 tracks. The text will be provided to the user in the form in which it was originally recorded, including phonetic transcription with preservation of accents.

3.7. Translations

The corpus offers a collection of translations of works about the Mezen Robinsons into French, German, Dutch, and Italian. There are also books of T. Griesinger [17] and D. Roberts [22] translated into Russian.

3.8. Quantitative Distribution of Texts in the Corpus

The quantitative distribution of publications about Mezen Robinsons relative to the time of their appearance is shown in the diagram (Figure 2).

Figure 2 presents a chronological straight line on which annual ranges are marked: 1766–1800 – the period of translations of Le Roy's book into different languages; 1800–1900 – the epoch of the story in newspaper and magazine periodicals; 1900–2000 – the epoch of the story embodied in the art form; 2000–2022 – the period of the story in the digital environment and the time of the appearance of studies of Mezen Robinson.

The color indicates the style of the text: fiction texts and publications – dark blue; articles in scientific journals – red; newspaper publications – green; publications in new media – purple; translations – light blue. The numbers in the columns denote the number of published texts about Mezen wintering on Edge Island.

As it shows, the peak of the story's popularity among readers falls on our days: the story of the "Russian Robinsons" is presented in different formats - newspaper articles, scientific publications, translated texts, Internet texts, and poetic and prose texts. At the same time, most of them are on the Internet due to modern reality and the era of "big data" and speed.

It is worth mentioning that in the middle of the 19th century, editors of Russian children's publications were paying attention to the story of the Mezen people's survival in the Arctic. The major part of the texts of this period are journal translations of Le Roy's text. In 1840-1860s, literary magazines were the primary source of literature.

The Soviet era was not inferior to the previous century in the number of literary texts. In the 30-50s of the 20th century, the Mezen story became an independent plot and embodied the genre of the adventure novel (Z. Davydov, K. Badigin, S. Radzievskaya, etc.). This is the time of struggle for the Arctic, the assertion era of the idea of indestructibility and strength of domestic navigators.

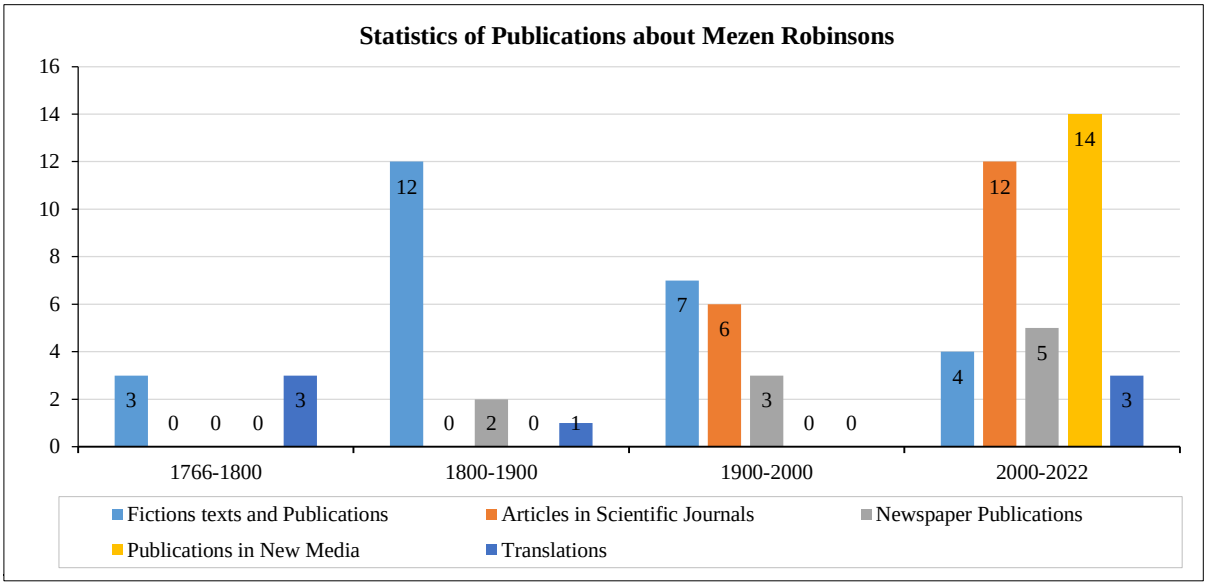

Figure 2. Statistics of publications in the corpus depending on the year of writing

3.9. Corpus Interface

For user convenience, the corpus includes the construction of cards with meta-information on each source (Figure 3). The card specifies the author, title of the work, the year of publication, scope, genre of the text, and type of text.

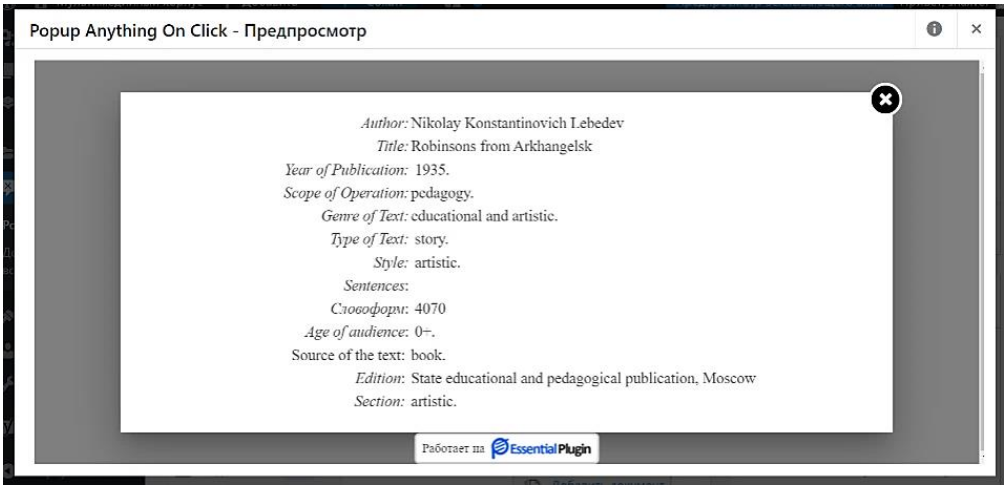

Figure 3. An example of a card with metainformation

The texts in the corpus menu are now divided into four categories: fiction texts and publications, newspaper publications, articles in scientific journals, publications in new media, audio speech, and translations into other languages (Figure 4). This division and, however, is purely conventional and does not prevent any text from one category from being compared with any text from another category. The order of an arrangement of texts in the form is free.

If desired, the user can use the built-in Voyant Tools. This resource helps establish the lexical density of the document, perform semantic analysis of the text, establish the most frequent words in the document, perform automatic calculation of different parts of speech in a particular text, and build concordance (Figure 5).

At the moment, the corpus has a context search set up. A glossary of reference words composing the thematic field "Mezen Robinsons" (210 words) was compiled in advance. These words are highlighted in blue in each document. When you click on a certain word, a box with commentary pops up (Figure 6).

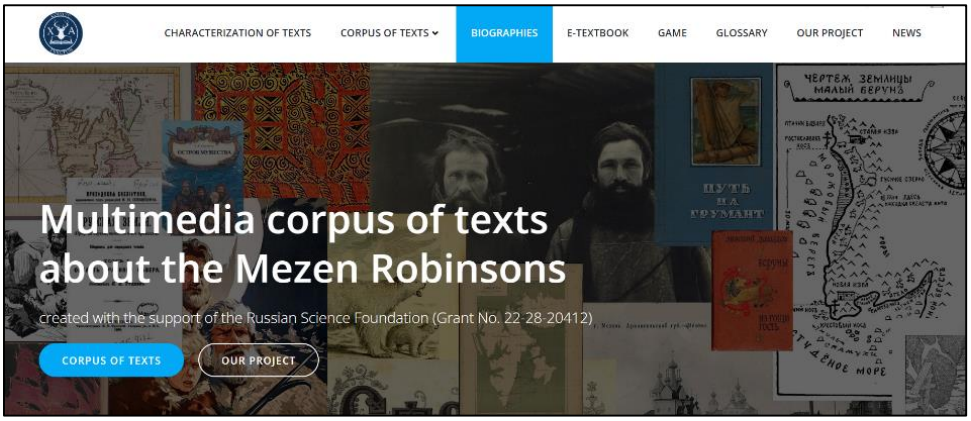

Figure 4. Multimedia corpus menu

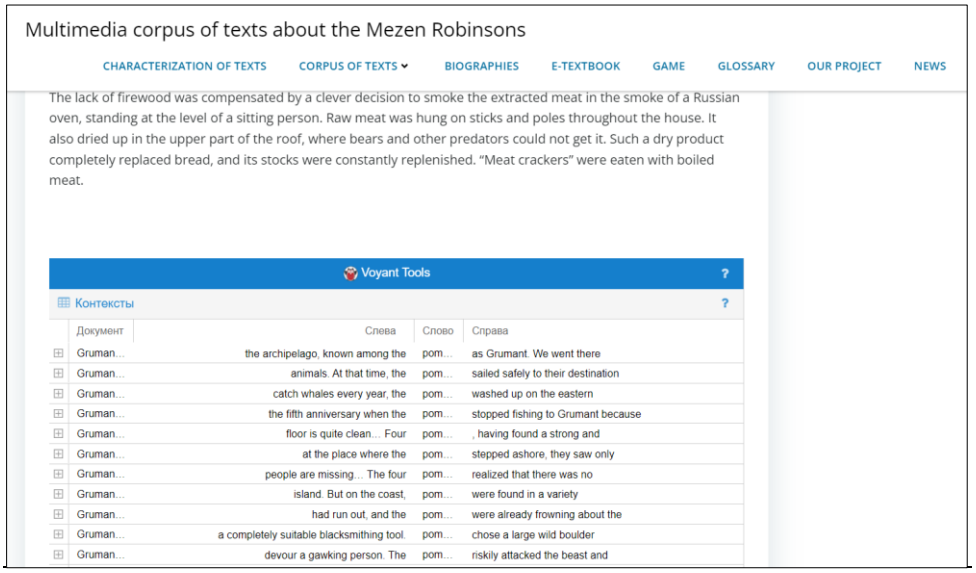

Figure 5. An example of Voyant Tools usage

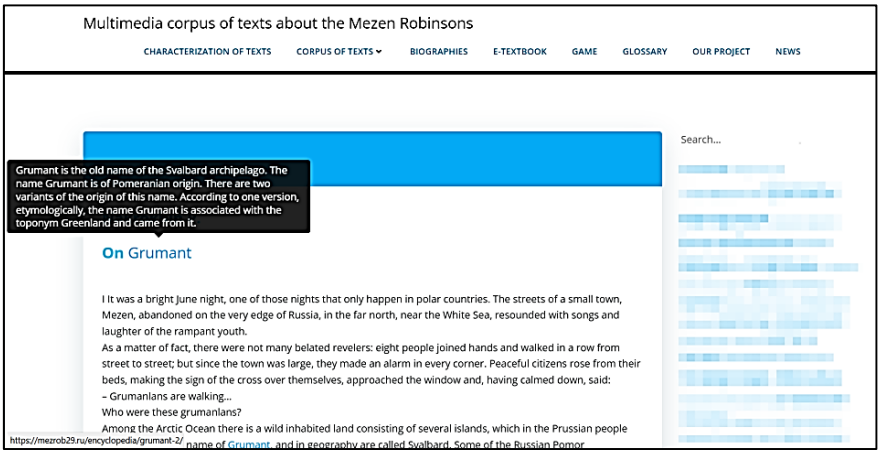

Figure 6. A glossary of a reference word (example)

Using the visualization method (word cloud), it is possible to highlight significant text points. Such visual representation is convenient for quick perception of any text, and in our case the whole plot is contained in a large array of texts (Figure 7).

Figure 7. Word cloud on the basis of the story about Mezen Robinsons

The word cloud is compiled on the basis of the thematic glossary in the corpus of texts. It allows us to systematize and synthesize information (data volume) in a simple and attractive way. Thanks to the word cloud, the main thematic points of the story about Mezen Robinsons are clearly distinguished. These are the names of the main figures (Alexei is the main character of the story, the steersman and senior representative of the hunter's dynasty), their naming words ("Robinsons", "Grumlan", "Hunters"), the names of survival supplies, the names of loci and toponyms, etc.

Multimedia support involves images, pictures, and video files that open when you click on the word (Figure 8).

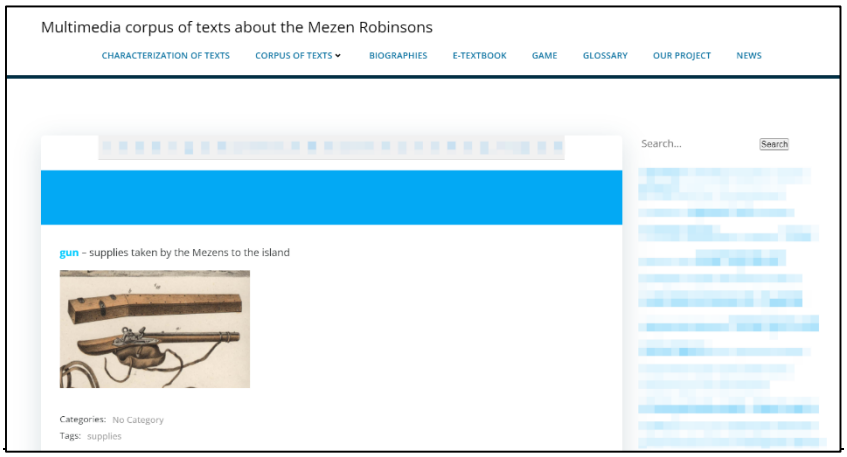

Figure 8. An example of multimedia support for a word

3.10. Corpus Architecture

The multimedia corpus of Mezen Robinsons texts offers various sections of information: "Corpus Characteristics," "Corpus of Texts," "Biographies", "Multipark," "News," "Tutorial," "Novella," "About the Project" (see Figure 9).

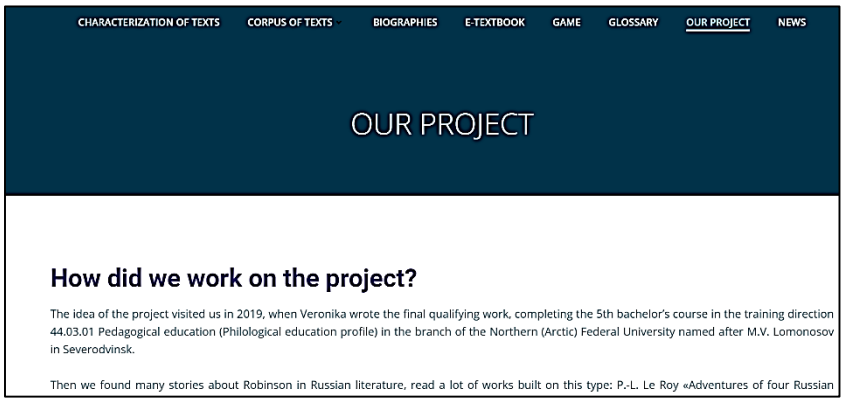

Figure 9. Sections of the multimedia corpus

The "Corpus Characteristics" section provides general information about the various materials on the resource. The "Corpus of Texts" section contains links to the five subcorpora that make up the corpus. The "Biography" section accumulates information about the lives of the authors who were somehow involved in the genesis and development of the Mezen "Robinsons" story (P.-L. Le Roy, M.V. Arkhangelskaya, A. Zubkovsky, N.K. Lebedev, Z.S. Davydov, K.S. Badigin, N. Zaitsev, S.B. Radzievskaya, K. Konichev, M.Yu. Starchikov, O. Scherbatov, and others).

Multipark collects screenings of works, scientific and popular programs about this event from the history of the Russian fur seal industry. "Stale sea" is a legendary feature film, filmed in 1954 by director Y. Yegorov based on the literary work "The Road to Grumant" by K. S. Badigin.

The "News" section informs users about the latest events related to the preparation and presentation of the corpus. The "Tutorial" section describes the possibility of using the corpus materials for educational purposes. The Tilda platform presents chapters of multimedia tutorials with the development of tutorials for schoolchildren on the works of K.S. Badigin, S.B. Radzievskaya, Z.S. Davydov, and others. The materials include video lectures on the authors of the books, search tasks for texts, tests, and games.

The "Novella" section is an adaptation of P.-L. Le Roy's book in the form of a visual novel. The novella combines the text of the work of French people of the 18thcentury, exclusive illustrations, and musical accompaniment. The section "About the Project" presents the history of the project chronologically. The corpus contains 61 sources in Russian, including: 1 archival document, 23 works of various genres (translations, retellings, fables, short stories, novels) in the time range 1762–2021, 9 newspaper and journal publications between 1876–2022, 16 scholarly publications, 15 tracks of audio speech.

4. Discussion

The diachronic corpus with a narrow textual focus reflects the changes in the cultural description model of the issue regarding Spitsbergen and Arctic exploration in the literary texts written by authors of different nationalities. Texts about the Mezen "Robinsons" are distributed in the collection according to the genre and chronology of writing. On the basis of the presented corpus, the principles of commenting documentary and fiction texts about Mezen "Robinsons" were developed, considering the widest possible historical and literary context, described (including) within the framework of the undertaken study of the main array of foreign and Russian prose.

Many works of contemporary researchers in this field focus on compiling a collection of oral and written texts [23-26] where the keywords are highlighted, document characteristics are given, or principles for generating annotations are developed That is to say, such corpora consist of texts of the same genre, the same volume, and created approximately at the same time. Our corpus is stylistically heterogeneous, containing texts of many genres and categories. We do not set the task to measure actual changes in the semantics of words or to analyze the word frequency of a particular thematic group.

The multimedia corpus described in the article expands the analytical tools for studying and describing foreign- and Russian-language works about the voyage of the Mezen "Robinsons" of 1766–2022 (extracting and describing the "elementary plot", typologizing genre and narrative models, etc.). It makes it possible to create a "rich" analytical description of the story about the Mezen "Robinsons" within the specified chronological framework. On the basis of this corpus, it is possible to trace the genesis of the story about the Mezen "Robinsons", its development, filling with new elements or loss of such elements. It is also possible to establish the factors influencing the expansion or reduction of the plot as well as to look through the changes in literary techniques used by the writers of Mezen "Robinsons".

The methodology of creating a multimedia corpus is effective for tracing the time of changing the topic. It also allows you to identify dynamic patterns of topical fluctuations: which topos is demanded by readers of a particular epoch (staying on the island and its description – 18th century; struggling with difficulties – 20th century; salvation – 21st century).

This study focuses on the dynamics of changes in the theme of the island adventure in the target audience and genre embodiment of the story. The corpus is essentially an encyclopedia on the survival history of the Mezen fishers. The authors collected 69 printed and electronic texts in pre-revolutionary and modern orthography, unified them, described metadata, prepared a glossary (words related to the story of the Mezen industrialists' survival on the island), and each word was accompanied by a commentary, a picture (photo, video). In the corpus, along with the texts distributed by genre blocks, there is a visual novel (a game on the content of the work of the 18th century), which attracts pupils and students to the text of the 18th century due to its modern format. Thus, students get information about the event of a distant epoch. In addition, the corpus has the "Multimedia tutorial" tab, which contains useful training materials (such as video lectures, audio fragments, test assignments on the content of the novels) for students and teachers conducting classes in the "Russian Literature" course. The corpus also contains oral narratives (live speech audio files) of Mezen residents about the plot.

4.1. Practical Application of the Multimedia Corpus of Mezen Robinsons

The constructed multimedia corpus of texts presents interesting material for use in research and education. Examples of the subject corpus application to different spheres can be characterized as shifting from research and teaching to forensic linguistic [27, 28]. Corpus methods and corpora in general have been actively used in sociolinguistic research since the formation of these concepts. The created corpus is of interest to sociologists because the oral sub-corpus contains audio recordings of Mezen residents' oral narratives about a historical episode of the manufacturers' voyage to Spitsbergen.

A sociolinguistic analysis (Table 1) of the narratives recorded in Mezen led to the following conclusions:

- The story of the Mezen "Robinsons" is relevant and attractive to different social groups (students, workers, retirees, representatives of the scientific community);
- The respondents obtained information about the Mezen manufacturers from various sources;
- The main sources are excursions to the local history museum and stories of teachers - 25 people (25.3%), literary works - 20 people (20.2%);
- 66 people (74.3%) out of 98 heard similar stories about sailors who were shipwrecked and wintered in the polar Arctic;
- Part of the respondents (13.1%) know the history of the Mezen people and their descendants in detail; they have communicated with scientists who study this topic.

This research allows us to determine the degree of the preservation of the story in the cultural memory of the Mezen people.

The multimedia corpus of texts about Mezen Robinsons is a set of texts united by a common plot. The availability of electronic texts by the same author (for example, editions of Z.S. Davydov's novel of 1933, 1955) makes it possible to expand the range of tasks traditionally solved by stylistics and stylometry. The analysis of the structure of cognitive metaphors in the texts of N.K. Lebedev, Z.S. Davydov, K.S. Badigin, S.B. Radzievskaya, M.V. Arkhangelskaya, and others seems productive.

Based on the texts introduced into the corpus, it is possible to trace which linguistic and stylistic means were used by different authors in the 19th, 20th, 21st centuries for constructing the plot model, the description of plot situations in the story about Mezen "Robinsons". The study of the story about Mezen "Robinsons" with the help of structural and semantic analysis allows to identify an "elementary plot" in the text, to detect modifications and transformation of the story elements at different stages of literary history.

A collection of publications from 19th-century newspapers and magazines (Syn Otechestva; Journal for Reading...; Podsnezhnik; DetskoyeChteniye; Pravda Severa; MayakKommunizma; etc.) allows us to trace the development of the genre paradigm where the story of the Mezen manufacturers is presented. This is of interest to literary scholars and journalists.

One of the important principles of corpus formation is comparativism. Comparison of interpretations of the Mezen "Robinsons" story by the authors speaking different languages allows us to discover the story peculiarities in different national world views and the peculiarities of the vision of this story by representatives of different linguistic cultures. An interesting example is the comparison of the interpretation of the event that happened in the 18th century by the French historian, the witness to a historical fact, Le Roy [16], and the interpretation by the Russian writer M.Y. Starchikov [29].

The electronic corpus can be used as an effective tool for translators. It is of interest to compare translations of the story into other languages in the 19th and 20th centuries. It is also worth considering this phenomenon from the viewpoint of different types of art: the search for a suitable translation of a story from the language of literature into the language of cinema or painting is noteworthy.

Specific research points can attract specialists in history in the aspect of analyzing the Russian exploration of the North, the development of trades in the Russian North in the 18th century, missionary activity on Spitsbergen, and the fate of historical figures associated with the above-mentioned historical circumstances (P.I. Shuvalov, S.S. Vernizober, M.V. Lomonosov, P.-L. Le Roy, A. Kornilov, and others). In the example of literary texts, we can consider the ethnic stereotype, embodied in the story and represented by a set of traditions, customs, beliefs, superstitions, etc. This aspect is important to cultural researchers, ethnographers, and folklorists. A multimedia corpus of texts about Mezen "Robinsons" can be used during literature classes at schools because it contains scenarios for interactive lessons dedicated to reading the literary text.

5. Conclusions

This research on the construction of a multimedia corpus of texts about the Mezen "Robinsons" contributes to solving the problem of modeling text corpora. Corpus analysis tools make it possible to consider a large amount of data, support a hypothesis or conclusion with reliable textual evidence, make new observations, or refute the intuitive assumptions of a text researcher. A multimedia corpus of texts about the Mezen "Robinsons" is necessary for corpus analysis of the language and text. The possibilities of the corpus allow for linguo-culturological commenting on the source. The practical application of the corpus provides a way out of the linear reading of texts about the Mezen "Robinsons", contributes to their structural understanding and adequate scientific interpretation.

Using the presented corpus of texts, it is possible to find answers to questions that often arise before a literary critic:

1. What is the semantic richness of the title of the text?
2. What connections are indicated between the names and titles of the protagonists of a literary text?
3. What ideas and concepts are leading in the text?
4. What motives, details, and images determine the integrity and fullness of the artistic world of the work?
5. What intertextual connections are determined in this work?

The results of corpus analysis using the tools of the multimedia corpus of texts about the Mezen "Robinsons" are optimal and represent original scientific material. The multimedia corpus contains original texts, documents in Russian, information about the date of the events described in the document, information about the authors and categories with which the document is associated, and it is available for downloading and use on devices without Internet access.

The source of the corpus data is paper and digitized editions in pre-revolution and modern orthography. The vocabulary of the corpus included 150,600 unique word forms. Based on the above, we can say that the multimedia corpus of Mezen Robinsons texts, offering a collection of different editions, can be an excellent basis for further research on the story by different authors, as well as a basis for research on the comparison of translations of this story into different languages.

5.1. Limitations and Suggestions for Future Research

The limitations of this study are mostly related to the object itself being rather narrow, since only the texts dedicated to a particular historical event, rather obscure one, are relevant for the study. The analyzed text corpus currently has only 95 texts (including 18 pieces of oral narratives); although adding new samples to the collection is possible since the story about Mezen Robinsons keeps receiving new iterations in works of literature, Internet publications, and even modern oral narratives of Mezen town.

This study also has the limitation of the impossibility of producing an experiment. The authors constructed the multimedia text corpus (implemented as a web resource); collected the texts available in libraries and the Internet; united the form of the gathered materials; created the glossary; and introduced the context search. In the future perspective, such additions are expected to perform semantic search and therefore create topical clusters.

The results obtained in our study can serve as a good basis for performing corpus analysis of lexemes and word forms from such semantic fields as “Arctic” and “Arctic Robinsinade”.

6. Declarations

6.1. Author Contributions

Conceptualization, T.V.S.; methodology, T.V.S., V.E.S., and S.A.D.; software, S.A.D.; validation, T.V.S.; formal analysis, V.E.S.; investigation, V.E.S. and S.A.D.; resources, S.A.D.; data curation, T.V.S.; writing—original draft preparation, V.E.S. and S.A.D.; writing—review and editing, T.V.S.; visualization, V.E.S.; supervision, T.V.S.; project administration, T.V.S.; funding acquisition, T.V.S. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The research was funded by the Russian Science Foundation (Project No. 22-28-20412 "Multimedia corpus of Mezen robinsons texts: Ideas for creating and spreading", implemented at the Northern (Arctic) Federal University named after M.V. Lomonosov).

6.4. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. References

[1] Monogarova, A., Shiryaeva, T., & Arupova, N. (2021). The Language of Russian Fake Stories: A Corpus-Based Study of the Topical Change in the Viral Disinformation. Journal of Language and Education, 7(4), 83–106. doi:10.17323/JLE.2021.13371.

[2] Dios, P. S. (2022). Veiga: a Multimedia Corpus of Film Subtitling for Multimodal Analysis. New Trends in Translation and Technology, 4-6 July, 2022, Rhodes Island, Greece.

[3] Russian Shakespeare (2007). Information and Research Database. Shakespeare Commission of the Russian Academy of Sciences, Russia. Available online: https://rus-shake.ru/ (In Russian).

[4] Orekhov, B. (2023). The Tale of Igor's Campaign: Corpus. Available online: http://nevmenandr.net/slovo/pro.php (accessed on April 2023). (In Russian).

[5] Pelcz, K. (2022). A Multimedia Corpus for Language Teaching Purposes: the MagyarOK Video Corpus. Studi Finno-Ugrici, ns, 2, 1-20. doi:10.6093/1826-753X/9863. (In Italian).

[6] Zhang, Y., Hu, W., & Liu, L. (2022). The Construction And Application Of The Multimedia Corpus Of Bisu Language: Taking The Study On Measure Words As An Example. Journal of Positive School Psychology, 6(10), 3902-3914.

[7] Al-Maadeed, S., AlJa'am, J., Khalifa, B., & Elsaud, S. A. (2021). MOALLEMCorpus: A Large-Scale Multimedia Corpus for Children Education of Arabic Vocabularies. 2021 IEEE Global Engineering Education Conference (EDUCON), Vienna, Austria. doi:10.1109/educon46332.2021.9453983.

[8] Wu, H. (2021). Multimedia Interaction-Based Computer-Aided Translation Technology in Applied English Teaching. Mobile Information Systems, 2021, 1–10. doi:10.1155/2021/5578476.

[9] Khokhlova Maria, V. (2023). Learner corpora: relevant information and an overview of the existing frameworks. Terra Linguistica, 51(1), 57-69.

[10] Ahmed, S., Sadeq, N., Shubha, S. S., Islam, M. N., Adnan, M. A., & Islam, M. Z. (2020). Preparation of bangla speech corpus from publicly available audio & text. Proceedings of The 12th language resources and evaluation conference, 13-15 May, 2020, Marseille, France.

[11] Zhang, J., Wang, C., Muthu, A., & Varatharaju, V. M. (2022). Computer multimedia assisted language and literature teaching using Heuristic hidden Markov model and statistical language model. Computers & Electrical Engineering, 98, 107715. doi:10.1016/j.compeleceng.2022.107715.

[12] Gomez Guinovart, X. (2019). Enriching parallel corpora with multimedia and lexical semantics. Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications, 90, 141–158. doi:10.1075/scl.90.09gom.

[13] Shen, Y., Yang, H., & Lin, L. (2022). Automatic Depression Detection: an Emotional Audio-Textual Corpus and A Gru/Bilstm-Based Model. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi:10.1109/icassp43922.2022.9746569.

[14] RSF. (2023). Multimedia Corpus of Texts about Mezen Robinsons. Russian Science Foundation, Russia. Available online: https://mezrob29.ru/ (accessed on March 2023).

[15] State Archive of the Arkhangelsk Oblast (2019). Historical Description of the Journey to Spitsbergen in 1743–1749 of Four Mezen Sailors: Alexey and Ivan Khimkov, Stepan Sharapov and Fyodor Virugin. Fond 6, Inventory 17, Case 1, 1-8.

[16] Leroy, P.-L. (1933). The adventures of four Russian sailors, to the island of Spitsbergen, a storm brought. All-Union Arctic Institute, Leningrad, Russia.

[17] Griesinger, T. (1894). In the Far North: Travel and Adventures in the Polar Lands. Oehmigke, Leipzig, Germany. (In German).

[18] Spokencorpora (2023). Stories about Dreams and Other Speech Corpora. Stories about Dreams. Available online: http://spokencorpora.ru/showcorpus.py?dir=00dreams (accessed on April 2023). (In Russian).

[19] Project Phil (2023). Saint Petersburg Hagiographic Text Corpora. Available online: http://project.phil.spbu.ru/scat/page.php?page=project (Accessed on April 2023). (In Russian).

[20] Prozhito (2023). European University at St. Petersburg. Prozhito" (Lived thorugh). Available online: https://prozhito.org/ (accessed on April 2023). (In Russian).

[21] Stenogramma. (2023). Politics and Literature. The Digital Archive of Literary Organizations in 1920-1930. Available online: http://stenogramma.imli.ru/ (accessed on April 2023). (In Russian).

[22] Roberts, D. (2005). Four against the Arctic: Shipwrecked for six years at the top of the world. Simon and Schuster, New York, United Sates.

[23] Wilbur, W. J., Rzhetsky, A., & Shatkay, H. (2006). New directions in biomedical text annotation: Definitions, guidelines and corpus construction. BMC Bioinformatics, 7(1). doi:10.1186/1471-2105-7-356.

[24] Weismayer, C., & Pezenka, I. (2017). Identifying emerging research fields: a longitudinal latent semantic keyword analysis. Scientometrics, 113(3), 1757–1785. doi:10.1007/s11192-017-2555-z.

[25] Webber, R., & Stroud, D. (2013). How changes in word frequencies reveal changes in the focus of the JDDDMP. Journal of Direct, Data and Digital Marketing Practice, 14(4), 310–320. doi:10.1057/dddmp.2013.19.

[26] Riedhammer, K., Gropp, M., Bocklet, T., Hönig, F., Nöth, E., & Steidl, S. (2013). Lmelectures: A multimedia corpus of academic spoken english. First Workshop on Speech, Language and Audio in Multimedia, 22-23 August, 2013, Marseille, France,

[27] Bloshchynskyi, I., Bahrii, H., Nanivska, L., Tsviak, L., Isaieva, I., Skyba, K., ... & Mishchynska, I. (2022). Gender Characteristics of Individual's Linguistic Behavior in the Context of Future Translators' Professional Training. Emerging Science Journal, 6, 199-208. doi:10.28991/ESJ-2022-SIED-014.

[28] Kopotev, M., & Mustayoki, A. (2003). Principles of the Creation of the Helsinki Annotated Corpus HANCO in the Internet. Scientific and Technical Information. Series 2. Information Processes and Systems, 6, 33–36.

[29] Starchikov, M.Yu. (2021). Polar Robinsons. Multimedia corpus of texts about the Mezen Robinsons. Available online: https://mezrob29.ru/mihail-yurevich-starchikov/ (accessed on April 2023). (In Russian).