УДК [81’374(=811.511.2):004.89](045)
DOI: 10.26102/2310-6018/2026.57.6.001

Архитектурная модель аудиомодуля онлайн-словаря тундрового ненецкого языка

П.Е. Шняков✉, Е.С. Коканова

Северный (Арктический) федеральный университет имени М.В. Ломоносова, Архангельск, Российская Федерация

Резюме. Цифровизация языковых ресурсов малоресурсных языков требует формализованной организации сбора, описания, контроля качества и публикации аудиоданных. В связи с этим целью исследования является разработка модели данных аудиомодуля для ненецко-русского и русско-ненецкого онлайн-словаря и контура поддержки принятия решений, обеспечивающего отбор лексических единиц для записи, их постобработку и интеграцию в словарную систему. Материалы исследования составили корпусные и словарные ресурсы, учебные и тематические материалы, ранее созданные аудиоресурсы, а также результаты полевого исследования, выполненного в Нарьян-Маре в декабре 2025 года. Методическая основа работы включает системный анализ, формализацию информационных потоков, многокритериальную приоритизацию лексики и описание воспроизводимой процессной схемы работы с аудиоматериалами. В результате определены сущности модели данных аудиомодуля, контур контроля качества и контур поддержки принятия решений по развитию аудиопокрытия словаря. Для списка из 542 единиц выполнено профилирование по типам единиц, частям речи, тематикам и микротематикам; дополнительно охарактеризованы состав информантов, структура аудиоматериалов, правила именования файлов и статусы контроля качества. Предложенное решение может использоваться при разработке цифровых словарей и речевых ресурсов для малоресурсных языков.

Ключевые слова: информационная система, онлайн-словарь, аудиомодуль, поддержка принятия решений, метаданные, малоресурсный язык, тундровый ненецкий язык.

Благодарности. Авторы выражают признательность Марине Дмитриевне Люблинской за содействие, поддержку и предоставление доступа к русско-ненецкому озвученному разговорнику, а также носителям ненецкого языка, участвовавшим в пилотном полевом исследовании, за вклад в создание и развитие языкового ресурса. Работа выполнена в рамках проекта (гранта) «Сохранение и развитие тундрового ненецкого языка в цифровой среде»[1].

Для цитирования: Шняков П.Е., Коканова Е.С. Архитектурная модель аудиомодуля онлайн-словаря тундрового ненецкого языка. *Моделирование, оптимизация и информационные технологии.* 2026;14(6). URL: https://moitvivt.ru/ru/journal/article?id=2342 DOI: 10.26102/2310-6018/2026.57.6.001

Architectural model of the audio module of the Tundra Nenets online dictionary

P.Y. Shnyakov✉, E.S. Kokanova

Northern (Arctic) Federal University, Arkhangelsk, the Russian Federation

Abstract. The digitalization of language resources for low-resource languages requires a formal organization of audio data collection, description, quality control, and publication. In this context, the

[1] Фонд президентских грантов. *Сохранение и развитие тундрового ненецкого языка в цифровой среде.* URL: https://президентскиегранты.рф/public/application/item?id=9b85fc37-90eb-4571-88c8-e684b3a5dcee (дата обращения: 08.04.2026).

study aims to develop a data model for the audio module of the Tundra Nenets online dictionary, i.e. the Nenets-Russian and Russian-Nenets online dictionary, and a decision support framework for selecting lexical units for recording, post-processing audio materials, and integrating them into the dictionary system. The empirical base includes corpus and dictionary resources, educational and thematic materials, previously created audio resources, and the results of fieldwork conducted in Naryan-Mar in December 2025. The methodological framework combines systems analysis, formalization of information flows, multicriteria prioritization of lexical items, and a reproducible workflow for processing audio materials. The study identifies the core entities of the audio module data model, the quality control framework, and the decision support framework for expanding the dictionary's audio coverage. A list of 542 units was profiled by unit type, part of speech, theme, and microtheme; the paper also characterizes the composition of informants, the structure of audio materials, file naming conventions, and quality control statuses. The proposed solution can be applied to the development of digital dictionaries and speech resources for low-resource languages.

Keywords: information system, online dictionary, audio module, decision support, metadata, low-resource language, Tundra Nenets language.

Acknowledgements: The authors express their gratitude to Marina Dmitrievna Lyublinskaya for her assistance, support, and for providing access to the Russian-Nenets audio phrasebook, as well as to the Nenets language speakers who participated in the pilot field study for their contribution to the creation and development of the language resource. The study was carried out within the framework of the project (grant) "Preservation and Development of the Tundra Nenets Language in the Digital Environment".

For citation: Shnyakov P.Y., Kokanova E.S. Architectural model of the audio module of the Tundra Nenets online dictionary. *Modeling, Optimization and Information Technology.* 2026;14(6). (In Russ.). URL: https://moitvivt.ru/ru/journal/article?id=2342 DOI: 10.26102/2310-6018/2026.57.6.001

Введение

Существенную роль в процессах изучения и документации языка играет аудиокомпонент, так как корректность усвоения фонетической структуры речи, в частности, просодических элементов (интонации и словесного ударения), непосредственно обусловлена качеством звукового сопровождения. Одновременно аудиоданные предъявляют повышенные требования к управлению: необходимы регламенты записи, метаданные, контроль качества, а также воспроизводимость постобработки и интеграции в информационные системы.

Современное состояние цифровой инфраструктуры тундрового ненецкого языка характеризуется фрагментарностью: разработка корпусных данных, лексикографических баз и аудиоколлекций происходит гетерогенно и асинхронно. Выбор приоритетных направлений озвучивания и пополнения ресурсов при этом осуществляется преимущественно на основе экспертных оценок, без применения единой формализованной модели [1, 2]. В этих условиях приоритетной задачей становится проектирование управляемой информационной системы, обеспечивающей воспроизводимость, качество и интеграцию результатов.

Цель статьи – предложить архитектурную модель аудиомодуля онлайн-словаря тундрового ненецкого языка[2], включающую модель данных и контур поддержки принятия решений (СППР).

Ключевая прикладная задача аудиомодуля заключается в оптимизации аудиопокрытия лексического фонда при соблюдении ряда ограничений, обусловленных дефицитом информантов, ресурсоемкостью технической обработки,

[2] *Ненецко-русский и русско-ненецкий онлайн словарь.* URL: https://nenrusdict.narfu.ru (дата обращения: 30.03.2026).

диалектологической неоднородностью материала и лакунами в лексикографической базе данных.

Исходным лексическим массивом для формирования очереди записи служил базовый онлайн-словарь тундрового ненецкого языка, подготовленный в 2024 году [2]. Поверх этого словарного ядра выполнялись тематическая классификация единиц, приоритизация для полевой записи и последующее профилирование итогового списка.

Для задач цифровизации языковых ресурсов существенны не столько частные методы обработки текста, сколько принципы построения управляемых информационных систем, в которых объединяются данные, процедуры их обработки, контроль качества и механизмы принятия решений [3, 4]. Как отмечают В.А. Петров и А.А. Филиппов, в задачах классификации текстов «не существует универсального подхода» для решения проблемы [5], а выбор метода определяется структурой данных, постановкой задачи и требованиями к качеству результата [4]. В системах поддержки принятия решений это означает необходимость проектирования не только алгоритмического слоя, но и контуров управления данными, прозрачности критериев, объяснимости рекомендаций и воспроизводимости вычислительного процесса [6]. Воспроизводимость подобных систем повышается при стандартизированном описании структуры проекта, явной связи между данными, кодом, результатами и документацией [7].

С точки зрения системного анализа, устойчивость таких решений определяется не только характеристиками отдельных алгоритмов, но и формализацией входных данных, полнотой метаданных, прозрачностью процедур отбора и возможностью воспроизведения всех этапов обработки (Рисунок 1) [7, 8].

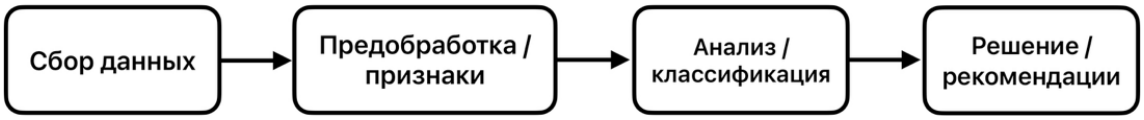

Рисунок 1 – Типовой контур научных работ по теме системного анализа
Figure 1 – Typical workflow in system analysis studies

Для разрабатываемого аудиомодуля эти положения имеют прямое значение. В настоящей работе контур поддержки принятия решений рассматривается как механизм, обеспечивающий отбор единиц для записи, фиксацию приоритетов, выявление проблемных позиций и планирование последующей обработки аудиоматериалов. Контроль качества, в свою очередь, трактуется не как отдельная техническая операция, а как встроенный элемент функционирования подсистемы, влияющий на приемку, повторную запись, сегментацию и публикацию аудиофайлов.

В контексте прикладных задач цифровизации языковых ресурсов информационная система (ИС) рассматривается как социотехническая система, интегрирующая данные, участников, процессы и инфраструктуру. В качестве методологической основы предлагается использовать рамку прикладного системного анализа Ф.П. Тарасенко[3], позволяющую формализовать описание границ системы, ее подсистем, моделей и механизмов управления. Данный подход обеспечивает переход от эмпирической работы с данными к построению формализуемых моделей информационных потоков, контроля качества и поддержки принятия решений.

Классификация Р. Акоффа применяется для позиционирования разрабатываемого решения как поиска «наилучшего варианта функционирования» в условиях ресурсных ограничений, характерных для языков КМНС (малые данные, вариативность, кадровые

[3] Тарасенко Ф.П. *Прикладной системный анализ.* Москва: КноРус; 2017. 322 с.

и организационные ограничения) [9]. Отсюда следует целесообразность ориентации на проектирование управляемого, а не стихийно развивающегося процесса.

В исследовании использовались четыре группы источников: корпусные ресурсы[4], словарные ресурсы[5], учебные[6,7] и тематические[8,9] материалы, а также полевые и ранее созданные аудиоресурсы. Корпусные и словарные ресурсы использовались для отбора и профилирования лексики, учебные и тематические материалы – для выявления тематически релевантной и культурно маркированной лексики, а полевые и ранее созданные аудиоресурсы – для апробации процедур записи, постобработки и интеграции аудио.

В качестве значимого текстового ресурса рассматривается ненецкий корпус на платформе «Корпусы ИЭА РАН»[10], где представлены тексты с метаданными (диалектология, подкорпус/стиль, исполнитель, год записи) и статистическими показателями (количество предложений, словоупотреблений, словоформ; частоты словоформ).

Дополнительным аудиоресурсом является русско-ненецкий озвученный разговорник[11] (Russian-Nenets Audio Phrasebook), включающий около 550 русских фраз, распределенных по 21 тематическому разделу, и несколько вариантов озвучивания, представляющих основные ненецкие говоры. В настоящее время ресурс недоступен онлайн; доступ к его офлайн-копии требует отдельного согласования с правообладателями. Данный пример показывает, что даже ранее созданные аудиоданные могут оказаться ограниченно доступными при отсутствии устойчивой модели хранения и распространения.

Таким образом, существующие цифровые ресурсы предоставляют основу для отбора и описания лексического материала, однако одновременно выявляют ограничения, требующие специальной организации аудиокомпоненты словаря.

Полнота и прозрачность метаданных имеют методологическое значение, поскольку оценка репрезентативности корпуса и интерпретация его ограничений зависят от того, насколько последовательно описаны состав данных, их происхождение и условия получения [8].

Материалы и методы

Процедура формирования списка для пилотного полевого исследования строится как многоэтапный процесс, в котором автоматическое ранжирование дополняется экспертной фильтрацией. В данной статье этот процесс концептуализируется в виде контура поддержки принятия решений. Входными данными для него выступает множество кандидатных лексикографических единиц, включающее слова, словосочетания, фатические фразы, информативные вопросы, диалогические единицы и служебные единицы. На выходе формируется список для пилотного полевого исследования, а также набор метаданных, приоритетов и маркеров контроля качества.

[4] Budzisch J., Wagner-Nagy B. *INEL Nenets Corpus.* Universität Hamburg. URL: https://www.fdr.uni-hamburg.de/record/16518 (дата обращения: 30.03.2026).

[5] Бармич М.Я. *Русско-ненецкий словарь.* Санкт-Петербург: Алмаз-Граф; 2023. 832 с.

[6] Ханзерова В.А. *Поговорим на ненецком.* Нарьян-Мар: Принт; 2018. 191 с.

[7] Бармич М.Я. *Картинный словарь ненецкого языка.* Санкт-Петербург: Филиал издательства «Просвещение»; 2006. 182 с.

[8] Щербакова А.М. *Ненецкие сказки. Репринтное издание 1960 года.* Ижевск: Принт; 2019. 90 с.

[9] Явтысый П.А. *Вына' я.* Таллин; 2005. 127 с.

[10] *Корпусы ИЭА РАН.* URL: https://corpora.iea.ras.ru/corpora (дата обращения: 30.03.2026).

[11] Sherstinova T. *Russian-Nenets Audio Phrasebook.* Academia.edu. URL: https://www.academia.edu/2468553/Russian_Nenets_Audio_Phrasebook (дата обращения: 30.03.2026).

Лингвистическая приоритизация лексики для аудиозаписи осуществляется на основании комплекса параметров, отражающих различные уровни языковой системы и функционирования единиц в речи:

– фонетический уровень: учет сегментных и супрасегментных особенностей, включая наличие артикуляторно сложных сочетаний, акцентную вариативность, темпоритмическую организацию высказывания, а также прогнозируемую потребность в дублях для достижения эталонного качества;

– стилистический уровень: анализ стилистической маркированности единиц, их регистровой принадлежности, роли в системе речевого этикета и контекстуальной обусловленности употребления;

– коммуникативно-прагматический уровень: оценка дидактического и культурного потенциала лексики, ее связи с типовыми ситуациями общения и степени риска прагматических ошибок при восприятии вне контекста.

Таким образом, параметры используются в работе как операциональные признаки приоритизации, влияющие на порядок записи, требования к контексту и глубину последующей верификации. Формализация этих признаков позволяет обеспечить повторяемость процедуры отбора и прозрачность обоснования приоритетов. Для языков с небольшой популяцией носителей такая прозрачность становится критически важной, поскольку количественные процедуры здесь сталкиваются с методологическими ограничениями [10].

Процедура формализованного отбора и профилирования лексического материала для аудиозаписи реализуется как многостадийный алгоритм (Таблица 1).

Таблица 1 – Процедура отбора и профилирования лексического материала
Table 1 – Procedure for the selection and profiling of lexical material

Этап	Описание
Компиляция первичного корпуса	Консолидация данных из разнородных источников в единый массив (~1500 единиц) с унификацией структуры полей (лемма, глосса, тип единицы)
Семантическая индексация	Присвоение каждой единице меток темы и микротемы, где микротематика выступает инструментом тонкой семантической дифференциации и оптимизации очереди записи
Прагматическая классификация диалогических единиц	Определение коммуникативной природы единицы (клишированная/свободная) и ее иллокутивной функции (приветствие, запрос, реакция и др.)
Валидация и фильтрация	Применение набора эвристик для автоматического выявления записей, нуждающихся в доработке: дефектные тематические метки, редкие микротемы, сложные для озвучивания многословные конструкции, единицы, нуждающиеся в контекстной подсказке
Генерация выходных форм	Формирование результирующего списка пилотного полевого исследования, присвоение идентификаторов и подготовка сессионных листов по тематическому и типологическому принципу

При проектировании речевого корпуса целесообразно задавать цели сбора, критерии отбора информантов, технические параметры записи, правила аннотирования и показатели качества до начала сессии записи и контролировать их на протяжении всего жизненного цикла ресурса [11].

Полевая запись рассматривалась как регламентированный процесс, предполагающий четкое распределение функций между участниками. В рамках записи

были выделены функции координатора, технического специалиста, лингвиста-верификатора и носителей языка; для контроля хода работы велась таблица статусов, содержавшая идентификатор единицы, сведения об информанте, дубле и комментарии. Случаи ошибок, оговорок или постороннего шума помечались как требующие повторной записи.

Современные подходы к созданию речевых корпусов подчеркивают, что техническое качество записи, пригодность ресурса для последующего использования и прозрачность критериев оценки должны рассматриваться совместно, а не как независимые этапы [11].

Постобработка аудиоматериала включала шумоподавление, отбор наиболее качественного дубля, сегментацию и сохранение отдельных файлов с трассируемыми именами и сопроводительными метаданными. В качестве основного инструмента постобработки использовалось программное обеспечение Amadeus Pro[12]. Программное обеспечение также позволяло выделять участок фонового шума, формировать на его основе шумовой профиль и применять шумоподавление по принципу минимально необходимого вмешательства, что позволяло минимизировать вероятность артефактов. При выборе дубля учитывались следующие параметры: отсутствие клиппинга и иных технических артефактов, минимальный уровень фонового шума, четкость артикуляции, отсутствие оговорок, а также соответствие интонационных и темпоральных характеристик целевым задачам использования и языковой единице.

До начала записи были подготовлены и подписаны формы информированного согласия, предусматривавшие разрешение на звукозапись, хранение, обработку и последующее использование голосового материала в составе цифрового ресурса.

Цифровые ресурсы тундрового ненецкого языка, включая базовый онлайн-словарь и связанные с ним данные, могут быть рассмотрены в составе более широкой информационной системы анализа и поддержки языковых ресурсов. В настоящей работе предметом рассмотрения является не система в целом, а аудиомодуль онлайн-словаря, в котором объединяются три уровня: данные, процессы и пользовательский доступ (Рисунок 2). На уровне данных фиксируются лексемы, информанты, записи, сессии и статусы контроля качества. На уровне процессов поддерживаются приоритизация, полевая запись, постобработка, приемка и публикация. На уровне пользовательского доступа аудио связывается со словарной статьей и становится доступным через интерфейс поиска и проигрывания.

Общая логика работы подсистемы представлена на Рисунке 2.

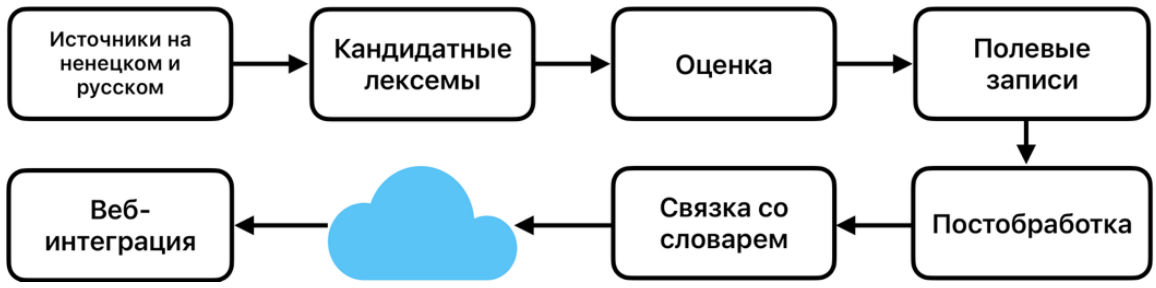

Рисунок 2 – Воспроизводимая процессная схема формирования аудиоресурсов
Figure 2 – Reproducible workflow for building audio resources

Схема включает компиляцию кандидатов для записи, их оценивание по набору признаков, полевую сессию, постобработку, формирование связей с базой словаря и

[12] *Amadeus Pro.* Amadeus. URL: https://www.hairersoft.com/pro.html (дата обращения: 30.03.2026).

публикацию. Процесс позволяет фиксировать не только конечный аудиофайл, но и основания для включения единицы в очередь записи.

В соответствии с системной интерпретацией цифровые языковые ресурсы тундрового ненецкого языка рассматриваются в рамках общей информационной системы анализа и поддержки языковых ресурсов (Рисунок 3).

Рисунок 3 – Модель «Черного ящика»
Figure 3 – Black-box model

На данном уровне рассмотрения граница системы задается функционально: она обеспечивает информационную и аналитическую поддержку процессов сбора, обработки и публикации языковых ресурсов (Рисунок 4).

Рисунок 4 – Модель состава системы
Figure 4 – System composition model

В рамках общей архитектуры аудиомодуль рассматривается как специализированный компонент, обеспечивающий хранение аудиофайлов, ведение метаданных, контроль качества, интеграцию аудиоматериалов со словарными единицами и их представление в структуре онлайн-словаря. Модель аудиомодуля может быть представлена как система с входами (лексические единицы, информанты), выходами (аудиофрагменты с метаданными), управляющими воздействиями (приоритизация, контроль качества) и обратной связью (контур качества).

Поскольку базовый онлайн-словарь реализован на реляционной СУБД MySQL[13], предлагаемая архитектура рассматривает аудиомодуль как расширение существующей

[13] *MySQL*. URL: https://www.mysql.com (дата обращения: 30.03.2026).

схемы данных путем введения дополнительных сущностей для обеспечения трассируемости аудиоданных (Таблица 2). В рамках данной модели WAV-файлы размещаются во внешнем файловом хранилище, тогда как в базе данных сохраняются метаданные. Метаданные аудиофрагмента представлены минимальным набором атрибутов, обеспечивающих идентификацию, трассируемость и управление жизненным циклом записи: *audio_id, lexeme_id, speaker_id, take и qc_status*, используемыми в качестве имен атрибутов модели. Указанные поля позволяют однозначно связать аудиофрагмент с лексической единицей и информантом, зафиксировать конкретный дубль записи и отразить результат процедуры контроля качества, что делает возможными приемку, повторную запись и последующую интеграцию аудио в словарную систему.

Такое разделение потоков данных позволяет снизить нагрузку на реляционную базу, упростить резервное копирование и обеспечить независимое масштабирование бинарного хранилища и слоя метаданных.

Таблица 2 – Сущности аудиомодуля
Table 2 – Audio module entities

Наименование	Описание	Аргументы
AudioItem	Аудиофрагмент	ID, ссылка на файл, технические параметры, показатели качества, связи с лексемой, информантом и дублем
Speaker	Информант	ID, пол, возрастная группа, диалект/говор, регион/происхождение
RecordingSession	Сессия записи	Дата, место, оборудование, ответственные роли, условия
Annotation/QC	Разметка и контроль качества	Кто проверял, когда, статус: принято/дубль/перезаписать, комментарий

Для описания функционирования аудиомодуля и его интеграции в общую информационную среду выделяются два взаимосвязанных, но функционально различных контура.

Контур качества данных ориентирован на контроль технических параметров аудиофайлов, полноты и непротиворечивости метаданных, лингвистическую верификацию качества произношения и уровня шума, а также на контроль корректности интеграции аудиофрагментов со словарными единицами. Такой набор процедур соотносится с подходами к стандартизированному описанию речевых ресурсов и сопровождающих их метаданных [12].

Контур поддержки принятия решений по развитию ресурсов ориентирован на формирование отчетов о частотных списках и профилях лексикографических единиц, степени аудиопокрытия словаря, включая выявление лексем, не обеспеченных аудиозаписями, дефиците возрастных и диалектных групп, а также трудоемкости обработки. В его рамках вырабатываются рекомендации по приоритизации лексики для следующей сессии записи, выявлению единиц, требующих дополнительной лингвистической верификации, и планированию работ по привязке аудио к словарным статьям и распределению процедур контроля качества по тематическим пакетам. Существенным требованием к подобным рекомендациям является прозрачность оснований для приоритизации и возможность интерпретации критериев выбора пользователем системы [6].

Результаты

В результате автоматического ранжирования[14] на основе собрания параллельных и монолингвальных корпусов Nenets Language Datasets[15] (тундровый и лесной) и экспертной фильтрации был сформирован список объемом 542 единицы.

Фрагмент списка, иллюстрирующий принципы типологической и тематической разметки, приведен в Таблице 3; количественный профиль списка представлен на Рисунках 5–7.

Таблица 3 – Примеры единиц из списка (фрагмент)
Table 3 – Examples of units from the list (fragment)

Единица – перевод	Тип и тематическая разметка
ваӈг – берлога	Слово; Природа / Ландшафт и природные объекты
вато – закон	Слово; Человек и общество / Социально-правовые понятия
едэй юн – новость	Словосочетание; Активность и атрибуты / Действия и процессы
Сава яля ӈэя! – Добрый день!	Фатическая фраза; Речь и коммуникация / Приветствие
Нюмл ӈамгэ? – Как Вас зовут?	Информативный вопрос; Речь и коммуникация / Знакомство / Самопрезентация
нись — без (приставка, предлог)	Служебная единица; Речь и коммуникация / Морфемы и служебные слова

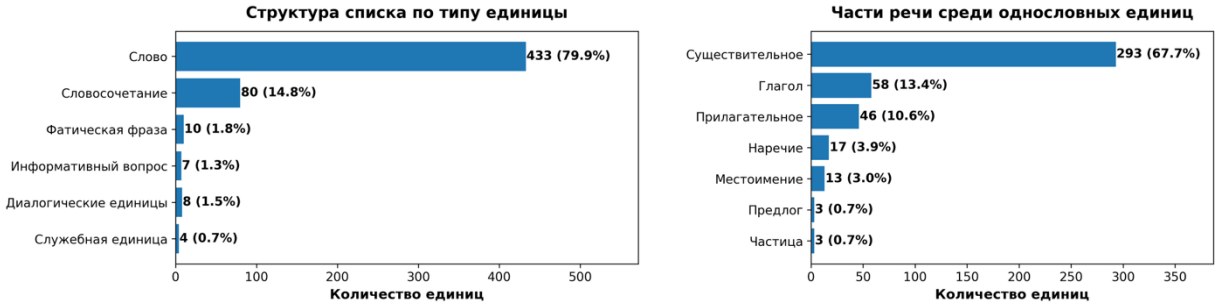

Рисунок 5 – Структура списка
Figure 5 – Structure of the list

Рисунок 6 – Тематический профиль списка
Figure 6 – Thematic profile of the list

[14] Shnyakov P. *Nenets Language Processing.* Zenodo. URL: https://doi.org/10.5281/zenodo.17369083 (дата обращения: 30.03.2026).

[15] Shnyakov P. *Nenets dataset.* Hugging Face. URL: https://doi.org/10.57967/hf/6728 (дата обращения: 30.03.2026).

В списке преобладают однословные единицы – их доля составляет 79,9 %. Среди них чаще всего встречаются имена существительные: 67,7 % от числа однословных единиц, или 54,1 % от всего списка. Для многословных конструкций ключевыми параметрами становятся сегментация и темпоритмическая структура, тогда как для диалогических единиц приоритетны определение коммуникативного статуса и прагматического типа.

Тематическое ядро образуют: «Активность и атрибуты», «Природа», «Человек и общество» и «Речь и коммуникация». Внутри них наиболее частотны микротематики: «Качества и состояния» (64 единицы), «Действия и процессы» (57), «Ландшафт и природные объекты» (45) и «Время, счет и деньги» (39). Среди малопредставленных микротематик выделяются «Пространство и ориентирование» (1 единица), «Рыболовство» (2), «Песни и фольклор» (2), «Охота» (4), «Рыбы» (4), «Транспорт и передвижение» (4) и «Еда и напитки» (5); указанные зоны могут рассматриваться в качестве приоритетных при формировании следующей очереди записи и тематических пакетов контроля качества.

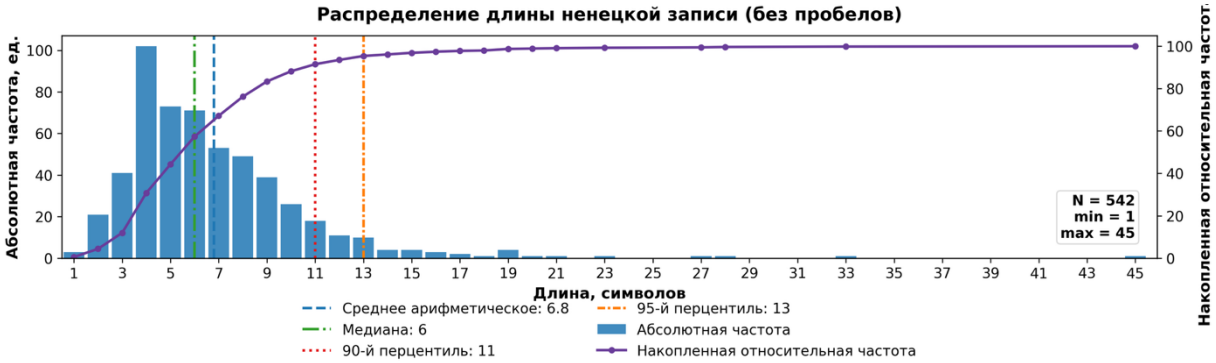

Рисунок 7 – Распределение длины ненецкой записи в списке
Figure 7 – Distribution of Tundra Nenets entry length in the list

Длина ненецкой записи варьирует от 1 до 45 символов; при этом 90 % записей не превышают 11 символов, а 95 % – 13. Этот показатель может использоваться как дополнительный критерий сегментации и предварительной оценки трудоемкости постобработки.

В пилотном полевом исследовании участвовали 9 информантов (Рисунок 8); в метаданные сессии были включены сведения о малоземельском, большеземельском и канинском говорах, а также о междиалектном влиянии у одного участника.

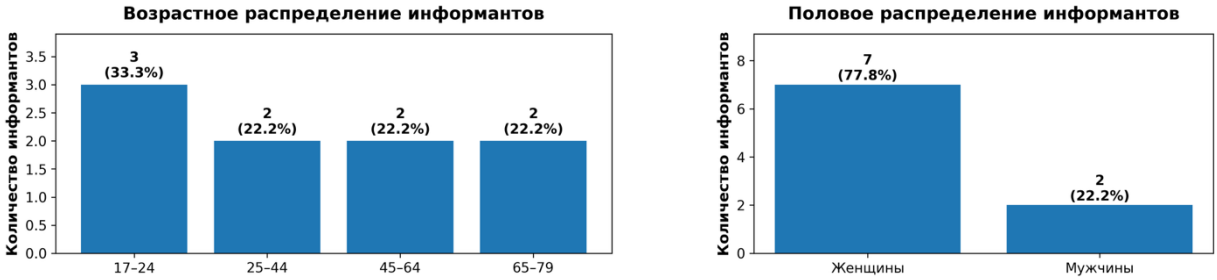

Рисунок 8 – Возрастно-половой профиль информантов
Figure 8 – Age and sex profile of the informants

Пользовательский уровень аудиомодуля обеспечивает доступ к связанным со словарными статьями аудиоматериалам через типовые операции поиска и навигации в

онлайн-словаре. В числе поддерживаемых функций предусматриваются фильтрация по параметрам информантов, сортировка словарного состава и строковый поиск.

Обсуждение

Доступные цифровые ресурсы характеризуются ограниченным и неоднородным объемом данных, различаясь по жанровому составу, степени разметки, полноте метаданных и объему представленного материала. Это влияет на устойчивость статистических оценок и требует фиксации версии и даты выгрузки используемых источников. Дополнительным фактором сложности выступает диалектологическая вариативность – привлечение носителей разных говоров повышает репрезентативность аудиоматериала, но одновременно усложняет описание метаданных, критерии приемки и сценарии представления аудио в онлайн-словаре [12].

Существенным ограничением остаются авторские и лицензионные условия использования внешних текстовых и аудиоресурсов, поскольку не все материалы допускают свободное повторное использование и публикацию. Это ограничивает возможность прямой интеграции отдельных материалов в состав словарного ресурса и требует отдельного согласования правового статуса данных. Наряду с этим одним из наиболее ресурсоемких этапов работы остается сегментация и постобработка аудиофайлов, что делает критически важными приоритизацию очереди обработки, формализацию контроля качества и последующую автоматизацию отдельных операций.

Практическая значимость исследования состоит в том, что предложенный подход может использоваться для формирования очередей записи с учетом лингвистического риска, дефицитных тематик и состава доступных информантов, для организации проверки аудиофайлов, отслеживания проблемных единиц и подготовки тематических пакетов для контроля качества, а также для воспроизводимой привязки аудио к словарным статьям. Кроме того, структурированные, тематически размеченные и снабженные метаданными аудиоматериалы могут использоваться как основа для дальнейших задач автоматической обработки речи, включая распознавание, синтез и учебные цифровые приложения [13].

Заключение

В рамках исследования предложена и эмпирически апробирована совокупность архитектурных и процессных решений, образующих аудиомодуль онлайн-словаря тундрового ненецкого языка. Системный анализ реализованного подхода показывает, что построение аудиокомпонента для малоресурсного языка требует сквозной согласованности на всех этапах жизненного цикла: от отбора единиц и полевой записи до постобработки, контроля качества и интеграции в штатный контур словаря.

Процедура отбора лексических единиц для озвучивания может быть реализована как формализованный механизм поддержки принятия решений, опирающийся на тип единицы, тематическую разметку и признаки лингвистического риска; такой подход позволяет уменьшить зависимость от экспертной интуиции и повысить воспроизводимость планирования. Профилирование отобранного списка (542 единицы) по типам единиц, частям речи, тематикам и микротематикам создает основу для перехода от эвристического к структурированному планированию очередности записи, что соответствует принципам управляемости ресурсами в системах с ограниченным доступом к информантам.

Полевое исследование, проведенное в Нарьян-Маре 11–13 декабря 2025 года, подтвердило применимость предложенного процессного шаблона: в работе участвовали 9 информантов в возрастном диапазоне 17–79 лет, а по его итогам был сформирован

массив данных из 71 файла слов и 11 жанровых аудиозаписей длительного формата. Полученные эмпирические данные свидетельствуют о функциональной состоятельности организационной схемы. Наиболее ресурсоемким элементом процессной архитектуры аудиомодуля остается постобработка: ручная сегментация и отбор дублей занимают 25–50 минут на каждые 30 слов, что выявляет критическую зону операционной эффективности и требует строгой организации очереди обработки и последующей автоматизации трудоемких операций.

Системная эволюция аудиомодуля предполагает расширение аудиопокрытия словаря по тематическим блокам, диалектным и возрастным группам информантов, а также разработку механизмов массовой привязки аудиоматериалов к словарным статьям в рамках единого технологического контура.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Епимахова А.С., Коканова Е.С. Ненецкий язык в цифровом пространстве. *Журнал Сибирского федерального университета. Гуманитарные науки.* 2025;18(10):1924–1931. (На англ.).
   Epimakhova A.S., Kokanova E.S. Nenets language in digital environment. *Journal of Siberian Federal University. Humanities & Social Sciences.* 2025;18(10):1924–1931.
2. Коканова Е.С., Шняков П.Е. Специфика разработки ненецко-русского и русско-ненецкого онлайн-словаря. *Этнопсихолингвистика.* 2025;(3):61–75.
   Kokanova E.S., Shnyakov P.Ye. Specifics of designing the Nenets-Russian and Russian-Nenets online dictionary. *Ethnopsycholinguistics.* 2025;(3):61–75. (In Russ.).
3. Malashina A.G. Possibility of Recovering Message Segments Based on Side Information about Original Characters. *Doklady Mathematics.* 2023;108(S2):S282–S292. https://doi.org/10.1134/S106456242370151X
4. Макарова Е.А. Обработка слабоструктурированных текстовых данных для использования в моделях анализа. *Информационные и математические технологии в науке и управлении.* 2023;(1):178–189. https://doi.org/10.25729/ESI.2023.29.1.015
   Makarova E.A. Processing of semi-structured text data for use in data analysis models. *Information and Mathematical Technologies in Science and Management.* 2023;(1):178–189. (In Russ.). https://doi.org/10.25729/ESI.2023.29.1.015
5. Петров В.А., Филиппов А.А. Анализ методов классификации текстов на естественном языке. *Вестник Ульяновского государственного технического университета.* 2024;(3):40–44.
   Petrov V.A., Filippov A.A. Analysis of natural language text classification methods. *Bulletin of Ulyanovsk State Technical University.* 2024;(3):40–44. (In Russ.).
6. Onwujekwe G., Weistroffer H.R. Intelligent Decision Support Systems: An Analysis of the Literature and a Framework for Development. *Information Systems Frontiers.* 2025;27(5):2027–2058. https://doi.org/10.1007/s10796-024-10571-1
7. Van Kampen A.H.C., Mahamune U., Jongejan A., et al. ENCORE: a practical implementation to improve reproducibility and transparency of computational research. *Nature Communications.* 2024;15(1). https://doi.org/10.1038/s41467-024-52446-8
8. Dirdal H., Johansen S.H., Durrant Ph. Representativeness and metadata presentation in learner/child corpora: Lessons from the GiG and TRAWL corpora. *Research Methods in Applied Linguistics.* 2024;3(3). https://doi.org/10.1016/j.rmal.2024.100145
9. Ackoff R.L., Magidson J., Addison H.J. *Idealized Design: Creating an Organization's Future.* Upper Saddle River: Wharton School Publishing; 2006. 336 p.

10. Гренобль Л. Новые горизонты в исследовании эвенского языка. *Северо-Восточный гуманитарный вестник.* 2024;(3):23–31. https://doi.org/10.25693/SVGV.2024.48.3.002
Grenoble L. New Horizons in Research on the Even Language. *North-Eastern Journal of Humanities.* 2024;(3):23–31. (In Russ.). https://doi.org/10.25693/SVGV.2024.48.3.002
11. Wieczorkowska A. Methodology for Obtaining High-Quality Speech Corpora. *Applied Sciences.* 2025;15(4). https://doi.org/10.3390/app15041848
12. Gibbon D., Moore R., Winski R. *Handbook of Standards and Resources for Spoken Language Systems.* Berlin, New York: Mouton de Gruyter; 1997. 886 p.
13. Сабуров А.А., Никифоров А.С., Минчук О.В. Состояние сохранности ненецкого языка в Ненецком автономном округе: по материалам социологического исследования. *Арктика и Север.* 2023;(50):189–210. https://doi.org/10.37482/issn2221-2698.2023.50.189
Saburov A.A., Nikiforov A.S., Minchuk O.V. Preservation of the Nenets Language in the Nenets Autonomous Okrug: Based on Sociological Survey. *Arctic and North.* 2023;(50):189–210. (In Russ.). https://doi.org/10.37482/issn2221-2698.2023.50.189

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Шняков Павел Евгеньевич, аспирант, ассистент кафедры информационных технологий, Северный (Арктический) федеральный университет имени М.В. Ломоносова, Архангельск, Российская Федерация.
e-mail: p.shnyakov@narfu.ru
ORCID: 0009-0004-5147-6647

Pavel Y. Shnyakov, Postgraduate, Assistant Professor at the Department of Information Technology, Northern (Arctic) Federal University, Arkhangelsk, the Russian Federation.

Коканова Елена Сергеевна, кандидат филологических наук, доцент, заведующий базовой кафедрой технологий и автоматизации перевода в бюро переводов «АКМ-Вест», Северный (Арктический) федеральный университет имени М.В. Ломоносова, Архангельск, Российская Федерация.
e-mail: e.s.kokanova@narfu.ru
ORCID: 0000-0001-6623-5636

Elena S. Kokanova, Candidate of Philological Sciences, Docent, Head of the Department of Translation Technology and Practice at AKM-WEST, Northern (Arctic) Federal University, Arkhangelsk, the Russian Federation.

Статья поступила в редакцию 08.04.2026; одобрена после рецензирования 27.05.2026; принята к публикации 09.06.2026.

The article was submitted 08.04.2026; approved after reviewing 27.05.2026; accepted for publication 09.06.2026.