

## **Электронные базы данных языков и фольклора народов Сибири и Дальнего Востока**

*И.В. Силантьев, Н.Б. Кошкарёва, Н.Н. Федина, И.В. Шапошникова,  
Е.Н. Кузьмина, А.А. Гриневич, П.С. Шахов, Е.Л. Тирон*

**Аннотация.** В Институте филологии СО РАН созданы и развиваются разные типы фольклорных и лингвистических электронных ресурсов, включающих материалы по языкам коренных и переселенческих народов Сибири и Дальнего Востока. Они предназначены для разных целей: во-первых, для хранения образцовых текстов, представляющих собой эталоны вербальной культуры в их исконной форме; во-вторых, для лингвистической обработки больших массивов текстов, представленных, в соответствии с международными стандартами, с глоссированием и поморфемным переводом, что обеспечивает возможность их использования в сравнительно-сопоставительных и типологических исследованиях; в-третьих, для обеспечения исследования вербальных и этномузыкальных культур оригинальным интерактивным инструментарием. Все электронные ресурсы направлены на многоаспектное отражение вербальных культур Сибири и Дальнего Востока, способствуют сохранению и развитию фольклорных традиций и языковой культуры в её лучших локальных проявлениях и укреплению этнических идентичностей.

**Ключевые слова:** языки, фольклор, этномузыкальные и вербальные культуры, коренные и переселенческие народы Сибири и Дальнего Востока, русский язык в Сибири, электронные ресурсы, фольклорные и лингвистические базы данных.

В Институте филологии Сибирского отделения РАН разрабатываются разнообразные электронные ресурсы и базы данных, включающие материалы по языкам и фольклору коренных и переселенческих народов Сибири и Дальнего Востока. Они находятся в свободном доступе на сайте института<sup>1</sup>. Некоторые из них подготовлены в рамках мегапроекта «Культурные универсалии вербальных традиций народов Сибири и Дальнего Востока: фольклор, литература, языки» по гранту Минобрнауки России (2019-2021 гг.) и являются результатом работы Лаборатории вербальных культур Сибири и Дальнего Востока, другие созданы в результате выполнения грантов, преимущественно РНФ.

Отличительной особенностью электронных ресурсов Института филологии СО РАН является их комплексный характер: все они интегрируют различные направления деятельности института в русле комплексных исследований языков, фольклора, литературы, вербальной культуры народов Сибири и Дальнего Востока.

---

<sup>1</sup> <https://www.philology.nsc.ru/resources/index.php>

## Фольклорные базы данных

В институте разрабатываются специализированные электронные ресурсы, в которых представлены многообразные фольклорные традиции народов сибирско-дальневосточного региона.

Одним из крупнейших электронных собраний фольклора народов Сибири и Дальнего Востока является электронная версия академической серии «Памятники фольклора народов Сибири и Дальнего Востока»<sup>2</sup> (руководитель Е.Н. Кузьмина). Это фундаментальный проект, представляющий собой новый тип публикации и исследования произведений устно-поэтического творчества. Тексты представлены на языке оригинала и в русском переводе, без каких-либо искажений и литературного вмешательства, с обязательным сохранением подлинной народной речи, диалектных форм языка. Аналитические статьи, сопровождающие публикацию текстов, включают филологические, музикноведческие и этнографические комментарии, что делает данное собрание ценным источником сведений о разных сторонах вербальных культур разных народов.

Электронный портал «Фольклор народов Сибири»<sup>3</sup> (руководитель А.А. Гриневич) является виртуальной исследовательской средой для проведения сравнительно-исторических исследований в области фольклора. Портал предоставляет инструментарий для сравнительного изучения фольклорных произведений – семантическую разметку текстов на основе онтологического подхода. Для образца и упрощённого начала работы с порталом разработано несколько онтологий, которые доступны любому зарегистрированному пользователю. Метаданные размечаются на основе онтологии фольклорных жанров, позволяющей охарактеризовать каждый публикуемый ресурс с точки зрения жанровой характеристики, указать национальный термин, которым традиционно обозначаются произведения подобного типа, отнести фольклорный образец к универсальной жанровой категории.

На материале бурятского фольклора разработана модель описания эпических текстов. Она содержит параметры для характеристики действующих лиц, их облика, еды, жилища, предметов, психологических состояний, событий и пр. Эта сеть взаимосвязанных понятий может быть применена и к другим родственным традициям.

Размещённая на портале онтология «Культурные универсалии народов Сибири» предлагает ряд концептов, характеризующих общие явления в различных сибирских языковых традициях. Например, в ней представлена модель описания эмоций и чувств, которые в фольклоре традиционно связаны с частями тела. Чтобы отобразить эту особенность в онтологии, между классами «Эмоции» и «Части тела» установлена связь, позволяющая указывать названия частей тела при описании эмоций. Все разработанные онтологии являются масштабируемыми и могут быть дополнены при необходимости. Инструменты

<sup>2</sup> <https://www.philology.nsc.ru/resources/pamyatniki.php>

<sup>3</sup> <https://www.philology.nsc.ru/resources/portal.php>

портала позволяют пользователям создавать собственные семантические сети взаимосвязанных понятий, посвящённых любому исследовательскому вопросу в области фольклора. Одни и те же сущности, размеченные в разных текстах, отображаются в едином реестре, который показывает параллели в разных языковых культурах и служит для навигации по ресурсам базы данных.

В настоящее время база данных портала содержит большую коллекцию на русском языке (фольклор русских в Сибири) и развивающиеся коллекции на бурятском, корякском и алюторском языках общим объёмом более 1500 текстов и иных ресурсов.

Сотрудники института разрабатывают уникальные ресурсы, посвящённые музыкальным культурам народов Российской Федерации, аналогам которых в научной практике изучения этномузыкального наследия фактически нет.

Ресурс «Песенная традиция овюорских тувинцев в условиях сельско-городской миграции» (руководитель Е.Л. Тирон; грант РНФ № 22-28-01475) демонстрирует трансформацию тувинской песенной традиции в контексте усилившейся миграции сельского населения в город. В 2022 г. проведено полевое исследование состояния песенной традиции овюорских тувинцев в сёлах Овюрского района Республики Тыва и в г. Кызыле. Экспедиция 2022 г. прошла по маршруту разведывательной музыкально-этнографической экспедиции Института филологии СО РАН 2009 года. Проведение повторных экспедиций на регулярной основе позволяет осуществлять мониторинг состояния традиционной музыкальной культуры.

В проведении полевого исследования активно использовались возможности социальных сетей. Так, для оповещения местного населения о музыкально-этнографической экспедиции были задействованы группа проекта в социальной сети VK, официальные группы администрации района, местные СМИ, а также местные блогеры. В публикациях была размещена информация о цели, сроках и населённых пунктах работы учёных, представление команды собирателей. Такая форма помогала при поиске информантов – знатоков песенной традиции.

Подготовленная в рамках проекта коллективная монография «Песенная традиция в контексте сельско-городской миграции» (Е.Л. Тирон, А.В. Байыр-оол, И.С. Тарбастаева) сопровождается аудиовизуальным приложением (около 600 аудиовидеозаписей), размещённым на одном из популярных видеохостингов<sup>4</sup>, а также в социальной группе проекта в VK. Такое представление результатов исследования обеспечивает их доступность для широкой общественности, а также способствует живому интересу местных жителей к своей традиционной культуре.

Проект «Фольклор автохтонных и переселенческих традиций народов Поволжья в современных записях и исторической динамике: интерактивный атлас звучащих текстов» (руководитель П.С. Шахов; грант РНФ № 19-78-10113) посвящён сопоставительному изучению автохтонных и переселенческих народов Поволжья – чувашей, мордвы и удмуртов поволжского и сибирского бытования. Научный коллектив объединяет фольклористов Института фило-

---

<sup>4</sup> <https://rutube.ru/pls/567250/>

## Труды Отделения историко-филологических наук

логии СО РАН и специалистов из поволжских научных центров – Саранска (Научно-исследовательский институт гуманитарных наук при Правительстве Республики Мордовия), Ижевска (Удмуртский институт истории, языка и литературы УдмФИЦ УрО РАН) и Чебоксар (Чувашский государственный институт гуманитарных наук).

В результате выполнения проекта сформирован круг разновременных аудиоисточников (архивных и современных полевых записей) по локальным песенным фольклорным традициям – сибирским и поволжским, связанным с местами выхода сибирских переселенцев.

Выявленные аудиоисточники разновременных фольклорных записей 1970-х – 2020-х гг. используются для создания интерактивного электронного атласа. Создание такого ресурса на фоне угасания фольклорных традиций имеет большую социокультурную значимость.

Электронный атлас фольклорных текстов<sup>5</sup> носит экспериментальный характер, так как связан с точечным картографированием взамен фронтальному районированию, принятому в традиционной картографии. Новаторский характер атласа связан также с созданием обобщённых унифицированных указателей по разным фольклорным традициям, каждая из которых представляет собой относительную целостность. В электронном атласе каждый фольклорный текст подаётся на языке оригинала с параллельным переводом на русский язык и сопровождается аудиоисточником. К тексту добавлены общие сведения (этнос, язык, место записи, исполнители, собиратели, год записи, место хранения источника, номер коллекции), содержательные характеристики (жанр, этнографический комплекс, тематика), аналитические данные (структура стиха, музикально-поэтическая композиция, фактура, слоговая музикально-ритмическая форма и др.), а также сведения о вариантах текста из электронного атласа.

Часть параметров доступны в качестве фильтров на интерактивной карте. Геометка на карте показывает место фиксации фольклорного текста, цветом обозначен этнос и добавлена информация о количестве представленных образцов, записанных в данном населённом пункте. При выборе какого-либо параметра из набора фильтров на карте остаётся определённая комбинация геометок, к которым привязаны фольклорные тексты, соответствующие фильтру (например, Красноярский край представлен чувашскими, удмуртскими и мокша-мордовскими материалами). Карта переключается в списочный формат отображения текстов (например, в Красноярском крае – 35 образцов). В атласе также реализована возможность полнотекстового поиска по заголовкам и поэтическим текстам.

В настоящее время в атласе размещены более 200 текстов, распределённых более или менее равномерно между этносами.

Атлас снабжён слогоритмическим модулем для описания особенностей напевов представленных песен. С помощью виртуальной клавиатуры этномузыкологи Института филологии СО РАН осуществляют набор музикально-слогового ритма напевов по трём уровням – ритмические группы / формулы, ритмические периоды, слогоритмические композиции в целом. После

<sup>5</sup> <http://folkmap.philology.nsc.ru>

введения данных появляется возможность поиска по этим трём уровням через картографические фильтры.

Все электронные ресурсы, разрабатываемые сотрудниками Института филологии СО РАН, носят междисциплинарный характер. Так, один из проектов молодых учёных под руководством Н.Н. Фединой, поддержанный РНФ (№ 17-78-20185), был посвящён тексту как явлению культуры и его значению для сохранения идентичности народов сибирско-дальневосточного региона. Работа велась по четырём направлениям: лингвистическому, фольклористическому, этномузыковедческому и литературоведческому.

Лингвистический аспект исследования направлен на отражение звучащего текста в письменной форме. Были проанализированы первые письменные тексты на алтайском, чалканском и тувинском языках в сопоставлении с современными текстами-самозаписями, что позволило выявить изменения, произошедшие в языке в неглубокой диахронии. Фольклорные исследования посвящены хакасскому героическому эпосу (алыптых нымах) с целью выявления типичных образов в этнической культуре. Проведено этномузыковедческое исследование фольклорной традиции тувинцев-тоджинцев и выявлены дифференцирующие признаки протяжных и скорых песен (жанров ыр и кожамык). Литературоведческие исследования связаны с изучением стиля произведений корякского писателя середины ХХ в. Кеца Кеккетына (на корякском языке).

В рамках проекта создан электронный ресурс «Тексты в этнической культуре народов сибирско-дальневосточного региона»<sup>6</sup>, где размещены произведения на национальных языках: алтайском, теленгитском, чалканском, хакасском, тувинском, тувинском-тоджинском, корякском-чавчувенском, алаторском, эрзя-мордовском. Собранные материалы продемонстрировали значительные отличия многих исследованных территориальных разновидностей языков от литературных норм, поэтому в базе данных некоторые алтайские, тувинские и корякские идиомы представлены самостоятельно. В настоящее время в базе данных проекта размещены более 150 текстов на 12 языках сибирских народов.

В базе данных реализована система поиска по этническому (или субэтническому) признаку, типу порождения текста (авторский, фольклорный), форме бытования (устный, рукописный, книжный), а также по жанровым типам, таким как автобиографические нарративы, нарративы-пересказы, сказки, исторические предания, песни, рассказы, очерки, стихотворения. Классификация также учитывает внутрижанровые разновидности. Например, песенные тексты подразделяются на лирические, круговые, личные и др., очерки – на событийные и портретные и т. п.

На сайте описываемого электронного ресурса представлены расшифровки устных, рукописных, книжных текстов на исследуемых языках, многие из них сопровождаются нотированием и глоссированием. Глоссирование выполнено по международным стандартам, что даёт возможность использовать представленный языковой материал для типологических исследований. Объём глосси-

---

<sup>6</sup> <https://text.philology.nsc.ru/about.php>

рованного корпуса составляет 25 текстов, 1346 предложений, 10048 словоупотреблений. Для скачивания доступны нотные тексты и сканированные копии источников. Более половины текстов на национальных языках сопровождаются аудиофайлами.

В рамках проекта изучалась также актуальная для малочисленных народов проблема вариативности национальной орфографии, которая препятствует нормализации письменной формы языка. В результате выполнения проекта формулируются рекомендации, направленные на стабилизацию письменных норм, что в конечном итоге способствует развитию литературных норм для языков коренных малочисленных народов РФ, обеспечивает стабильность и преемственность в сфере образования на родных языках.

### Лингвистические базы данных

Лингвистические базы данных разрабатываются на материале языков коренных народов РФ, а также на основе русского языка применительно к особенностям его бытования на территории Сибири и Дальнего Востока. Все базы данных также носят комплексный характер, их материалы демонстрируют не только собственно лексические и грамматические особенности языков, но и отражают культурные процессы на синхронном срезе и в неглубокой диахронной перспективе, актуальной для текущего момента.

Одним из таких проектов является электронный диалектологический атлас уральских языков, распространённых на территории Ямalo-Ненецкого автономного округа<sup>7</sup>, а также развивающийся на его основе электронный словарь языков народов Сибири и Дальнего Востока (руководитель Н.Б. Кошкарёва), включающий в настоящий момент данные по 12 урало-алтайским языкам (около 19000 лексических единиц), среди них такие тюркские языки, как алтайский, чалканский, тувинский, хакасский, шорский, барабинско-татарский, якутский. По тунгусо-маньчжурским языкам введены данные по ульчскому языку, ведётся подготовка к введению данных по удэгейскому, эвенкийскому и орокскому языкам. Из финно-угорских языков представлены языки казымских, приуральских, шурышкарских и сургутских хантов, а также ямальские говоры коми-зырянского языка. Самодийские языки представлены ненецкими идиомами – языками тундровых и лесных ненцев, а также северными диалектами селькупского языка. Подавляющее большинство материалов в словаре является результатом полевых исследований сотрудников Института филологии СО РАН.

Отличительной особенностью словаря является его полнотекстовый характер и значительный объём словарных статей. В них даётся максимально исчерпывающее описание лексического значения слова: его прямые и переносные значения, примеры словосочетаний и предложений, иллюстрирующих употребление, фразеологические единицы, отсылки к вариантам данного слова в других диалектах, указываются по возможности синонимы и

<sup>7</sup> <https://www.philology.nsc.ru/resources/atlas.php>

антонимы. Для некоторых языков, таких как хантыйский и ненецкий, объём словаря насчитывает несколько тысяч слов (от 4 до 6 и даже более тысяч слов). В словаре ставится цель показать лексическое богатство сибирских языков, зафиксировать максимально полно возможности реализации значений в разнообразных контекстах.

Поскольку существование подавляющего большинства языков Сибири и Дальнего Востока находится под угрозой, в словарных статьях размещаются аудиофайлы, представляющие собой записи отдельных слов, сочетаний и выражений, что обеспечивает возможность использования данного словаря как источника для фонетических исследований, а для носителей языка позволяет соотнести написание и произношение. Это тем более важно, что многие исследуемые языки являются младописьменными, в которых орфографические и графические традиции всё ещё не устоялись, и орфографическая вариативность оборачивается трудностями на письме и при чтении текстов.

В словаре предусмотрены возможности поиска по разным основаниям: по частям речи, по ключевым словам перевода, по грамматическим и семантическим пометам и т.д. Также реализована возможность отражения данных на лингвистических картах, которые в перспективе помогут создать многомерные описания ареальных контактов на территории Сибири и Дальнего Востока, уточнить границы распространения тех или иных диалектных явлений, построить новые изоглоссы. Функционал словаря постоянно дополняется и обновляется. В настоящее время сотрудники Сектора языков народов Сибири работают над онтологическими разметками, которые позволяют представить информацию о лексическом фонде в идеографическом ключе.

Другой тип электронных баз – это текстовые базы данных, размеченные при помощи специальной программы FIELD WORKS, находящейся в свободном доступе и предназначеннной для автоматической разметки текста. В рамках мегапроекта созданы базы данных для сибирского говора вепсского языка, для хантыйского и ненецкого языков. Представление лингвистической информации сопровождается традиционными метаданными о месте и времени сбора материала, об исполнителе и т. д. Тексты приводятся в глоссированной форме. В первой строке даётся запись предложения, во второй – разбивка слова на морфемы, в третьей – поморфемный перевод, в четвёртой – факультативно, при необходимости, буквальный перевод на русский язык, и в последней строке – литературный перевод. Такой способ представления позволяет использовать материалы сибирских языков для сравнительно-сопоставительных и типологических исследований, обеспечивает возможность сохранения языковых данных и их последующего использования в учебных целях, для написания академических грамматик и словарей. Общая база текстов составляет 24000 слов, 15750 предложений, 4200 морфем в словаре. База текстов на сургутском хантыйском – 10044 слов, 7139 предложений, 1176 морфем в словаре. База текстов на тундровом ненецком – 13072 слов, 8440 предложений, 2260 морфем в словаре. База текстов на вепсском (сибирский диалект) – 1023 слов, 164 предложений, 686 морфем в словаре.

Ещё один крупный проект, который осуществляется в секторе русского языка в Сибири под руководством И.В. Шапошниковой, – «Русская региональная ассоциативная база данных: Сибирь и Дальний Восток» (СИБАС)<sup>8</sup>. Проект выполняется на материалах эмпирической базы ассоциативно-вербальных данных, пополняемой в сотрудничестве с Новосибирским государственным университетом.

Феномен русской языковой личности (РЯЛ) эксплицируется в эксперименте на модель ассоциативно-вербальной сети, которая позволяет изучать смысловую динамику языковой личности, в первую очередь молодого поколения, и показывает обрастание историко-культурной основы значенияй слов инвариантами личностных смыслов.

Расширение экспериментальной базы изучения РЯЛ в пределах нескольких поколений даёт динамическую системную модель векторов смыслообразования, репрезентативную по отношению к образу мира носителей русского языка и российской культуры. Увидеть закономерности в изменении доминант у разных поколений россиян можно только на материалах больших, репрезентативных для целой страны и её регионов баз данных, таких как СИБАС.

СИБАС содержит основной корпус и подкорпусы. Основной корпус ассоциаций создаётся как пополняемая эмпирическая база для построения ассоциативно-вербальной модели русской языковой личности студенческой молодёжи Сибири и Дальнего Востока XXI в. Наличие такой эталонной для азиатских регионов РФ базы данных создаёт условия для изучения активной грамматики и флуктуаций актуальных смыслов как по временной оси – «по вертикали», так и «по горизонтали» – в сравнении с данными ассоциативных баз других регионов России. Кроме того, массовый характер и значительный охват по времени проводимых экспериментов позволяет не только исследовать динамику смысловых акцентаций, но и выявить ассоциативные доминанты русской языковой личности.

В структуре основного корпуса ассоциаций и подкорпусов СИБАС две части: 1) массив эмпирических данных, полученных экспериментальным путём, который находится в открытом доступе для пользователя; 2) исследовательская часть, включающая информацию об уже завершённых и текущих проектах, в которых используются эмпирические материалы СИБАС. Обе части постоянно обновляются по мере накопления новых данных и результатов реализации конкретных проектов. В настоящее время доступны для исследования основной корпус и подкорпусы со всем встроенным инструментарием обработки данных, а также новейшие материалы в формате ассоциативных словарей.

Общий объём базы – более 5000 слов-стимулов и более 1000000 словоупотреблений – обратных реакций.

### Заключение

Использование интерактивных технологий позволяет в режиме реального времени пополнять данные, интегрируя в единое информационное пространство результаты текущих и предшествующих исследований.

<sup>8</sup> <http://adictru.nsu.ru>

В Институте филологии СО РАН созданы и развиваются разные типы электронных ресурсов: одни из них предназначены для хранения образцовых текстов, представляющих собой эталоны вербальной культуры в их исконной форме; другие отражают результаты лингвистической обработки больших массивов текстов, представленных, в соответствии с международными стандартами, с глоссированием и поморфемным переводом, что обеспечивает возможность их использования в сравнительно-сопоставительных и типологических исследованиях и расширяет фактическую основу для проведения крупномасштабных лингвистических проектов; трети содержат оригинальный, разработанный сотрудниками института инструментарий для анализа языка и фольклора.

Особую ценность представляют ресурсы, показывающие синхронное состояние этномузыкальных культур, которые пользуются большой популярностью у носителей этих языков и культур. Собирая в полевых условиях фактический материал, исследователи обрабатывают его и представляют на сайте института в доступном для носителей культурных традиций виде. Тем самым сайт Института филологии СО РАН становится важным источником информации не только для учёных, но и для языковых и культурных сообществ на территории Сибири и Дальнего Востока.

Создаваемые в Институте филологии СО РАН ресурсы, направленные на многоаспектное отражение вербальных культур народов, населяющих территорию Сибири и Дальнего Востока, способствуют сохранению и развитию фольклорных традиций и языковой культуры в её лучших локальных проявлениях и укреплению этнических идентичностей.