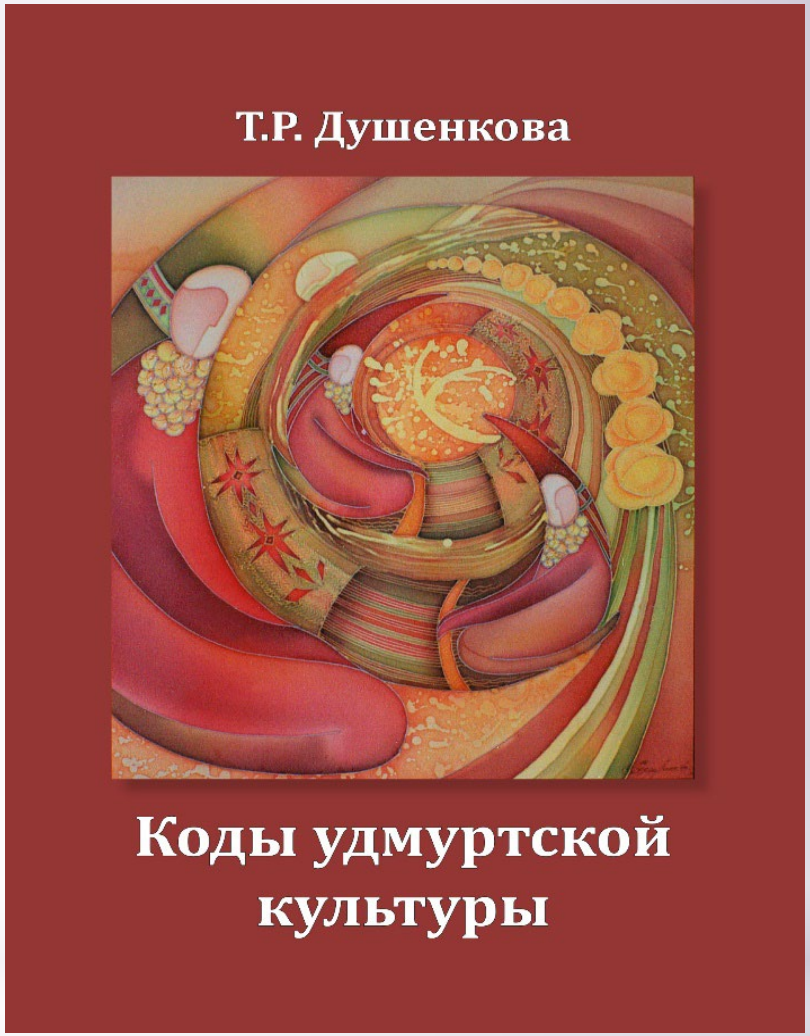
Удмуртский культурный код: практики использования цифровых технологий

Поздеев Игорь Леонидович, к.и.н.,
руководитель Удмуртского института истории, языка и литературы Удмуртского федерального исследовательского центра Уральского отделения РАН

Основные функции культурного кода в рамках культуры этноса:

- Этнодифференцирующая,
- Символическая/ценностная,
- Поведенческая.



Душенкова Т. Р. Коды удмуртской культуры: монография / УдмФИЦ УрО РАН. Ижевск, 2024.

Язык – основное средство отражения и сохранения культурного кода

Этапы цифровизации языков народов РФ

I. Создание сообществ в социальных сетях.

II. Создание раскладок клавиатуры для Windows, Linux, MacOS, Android, iOS.

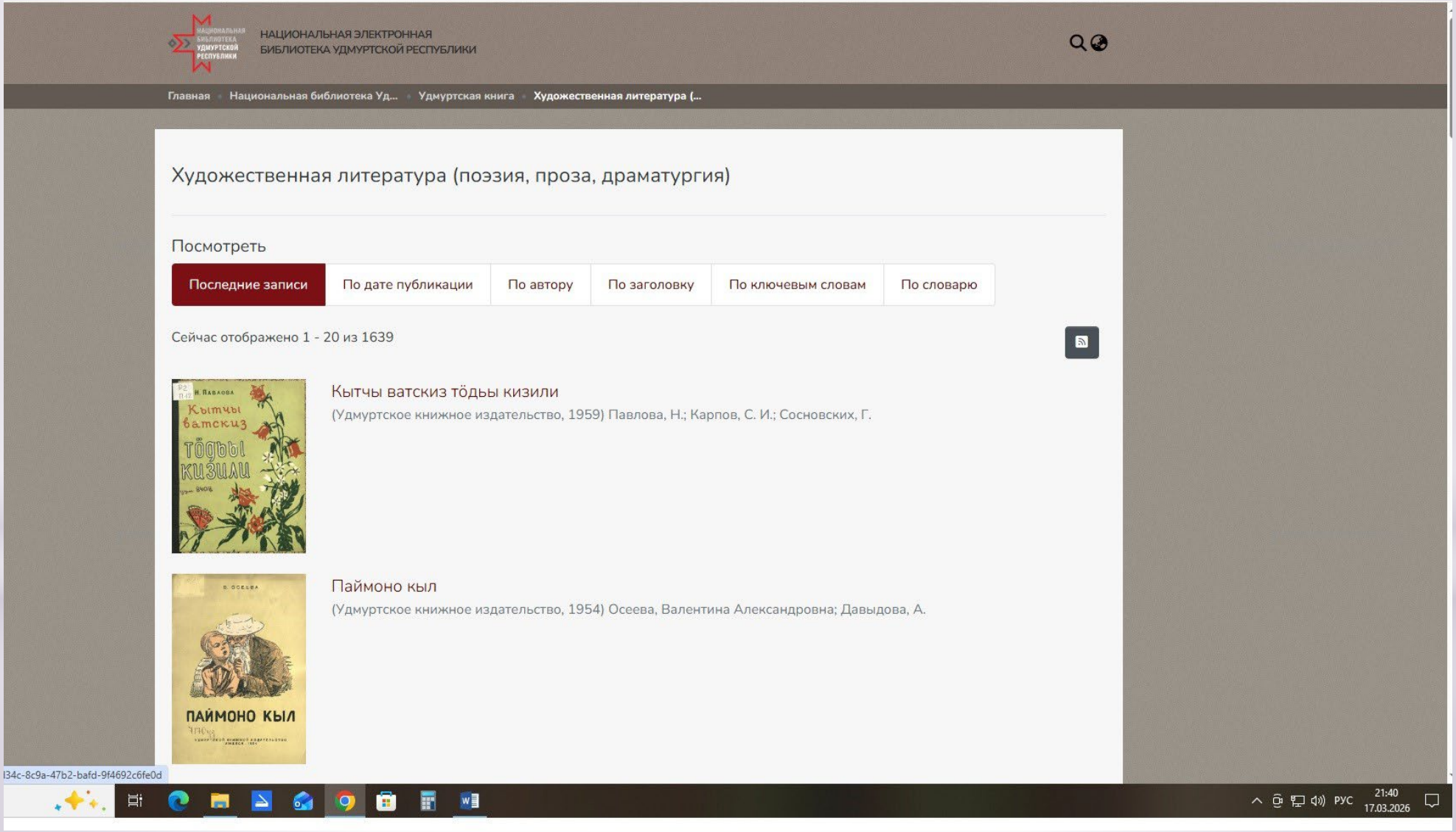
III. Подготовка к сбору корпуса (создание электронных словарей, коллекций сканов книг, программ распознавания текста и т. д.).

IV. Создание корпусов (монокорпус, параллельный корпус, корпус диалогов и т. д.).

V. Создание программ на основе искусственного интеллекта (машинный переводчик, синтезатор речи, распознаватель речи и т. д.).

VI. Создание продуктов на основе искусственного интеллекта (браузерные перевода сайтов и видео в сети Интернет, умная колонка расширения для и т. д.)

Представленность удмуртского языка в сети Интернет

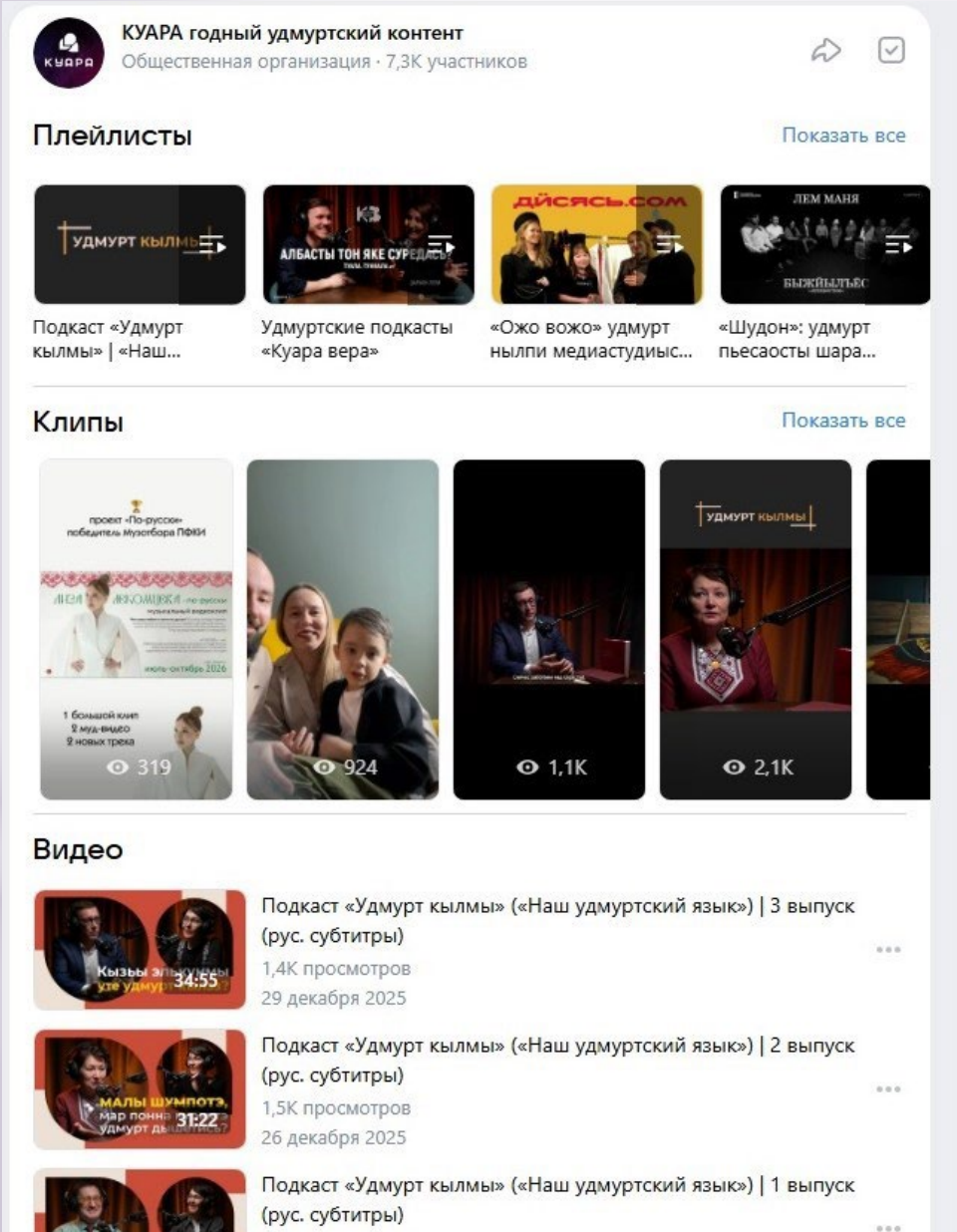


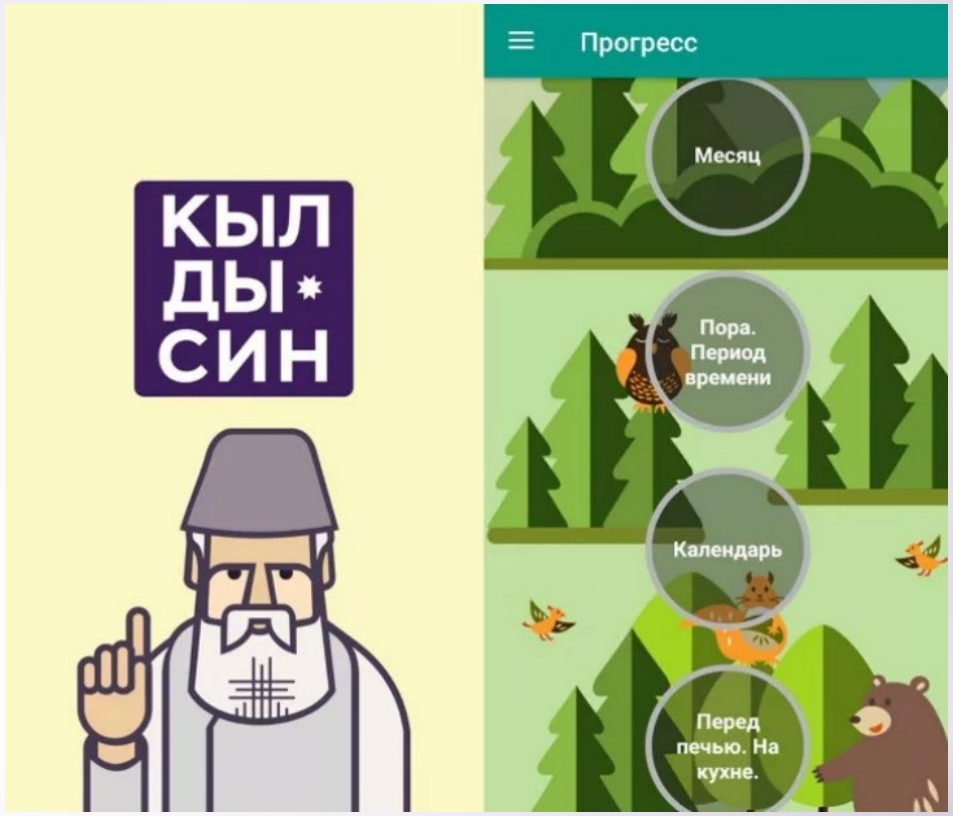
Представленность удмуртского языка в сети Интернет

Представленность удмуртского языка в сети Интернет

Информационно-образовательный портал по обучению удмуртскому языку и литературе «УДМУРТ КЫЛДУННЕ»

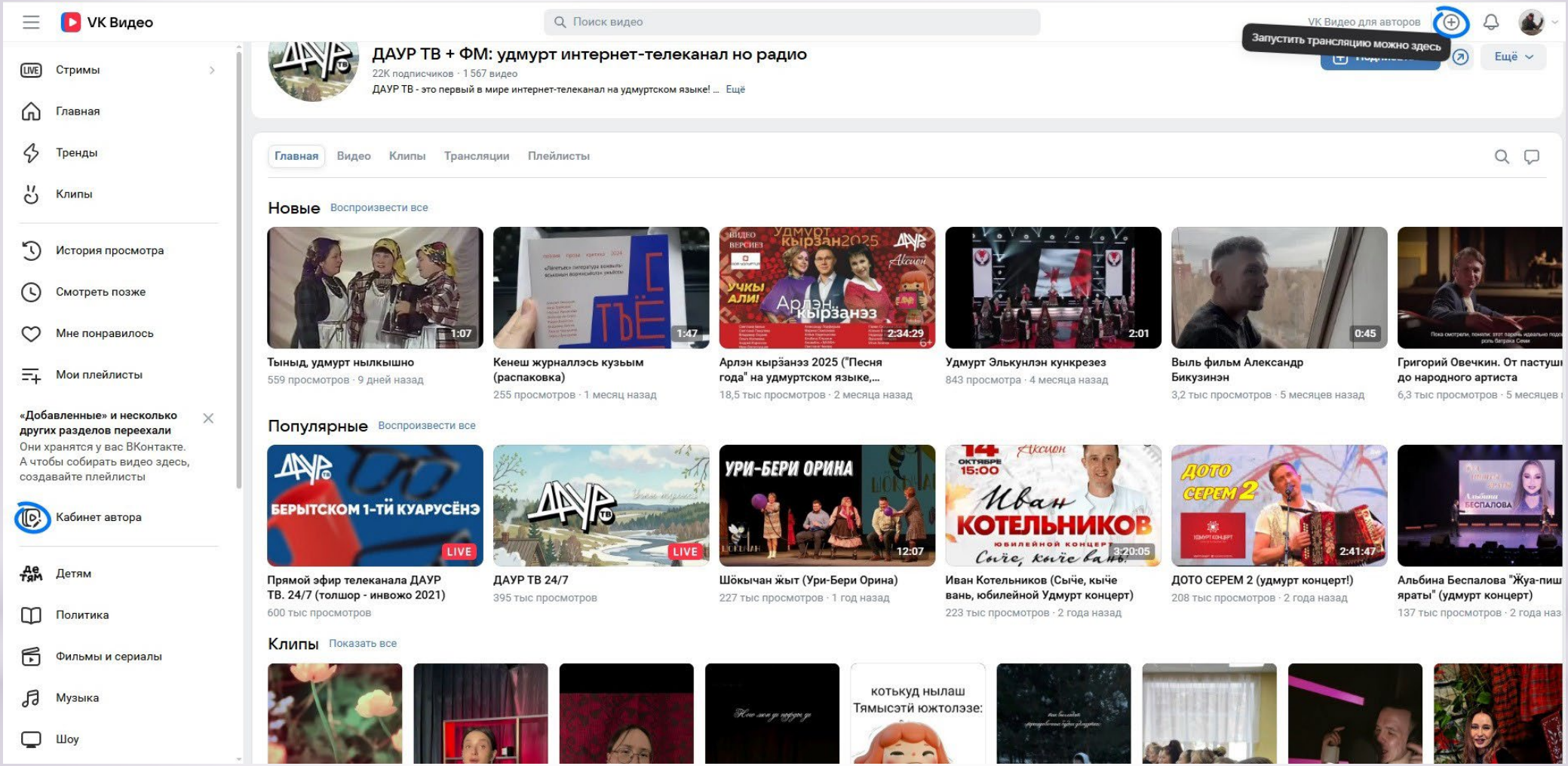
Создание и продвижение современного цифрового контента на удмуртском языке





Удмурт лякетъёс
(Удмуртские стикеры)

Создание и продвижение современного цифрового контента на удмуртском языке



В формате прямого эфира в социальной сети ВК создаются и показываются плейлисты с различными видео на удмуртском языке, ведутся трансляции мероприятий, проводятся ежедневные авторские передачи из студии.

Национальный корпус удмуртского языка

Удмурт йӧскалык кылшыкыс

Национальный корпус удмуртского языка — это информационно-справочная система, основанная на собрании удмуртских текстов в электронной форме. Корпус предназначен для всех, кто интересуется вопросами, связанными с удмуртским языком: профессиональных лингвистов, преподавателей языка, школьников и студентов.

СЛОВАРЬ

ОСНОВНОЙ КОРПУС

КОРПУС ПИСЬМЕННЫХ ПАМЯТНИКОВ

РУССКО-УДМУРТСКИЙ ПАРАЛЛЕЛЬНЫЙ КОРПУС

Участники проекта

Удмуртский федеральный исследовательский центр Уральского отделения Российской академии наук

Министерство национальной политики Удмуртской Республики

Национальная библиотека Удмуртской Республики

ООО "Нооматика"

https://udmcorpus.udman.ru

Модернизация национального корпуса удмуртского языка

1. Осуществляется регулярное пополнение основного корпуса (11 млн. словообразований на конец 2025 г.)
2. Разработан корпус письменных памятников удмуртского языка.
3. Разработан POS-тэггер (Part-of-Speech tagger) для удмуртского языка на основе ИИ – программа автоматического определения частей речи на основе контекста.

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2026614037

Программа для морфологической разметки текстов на удмуртском языке

Правообладатель: *Федеральное государственное бюджетное учреждение науки «Удмуртский федеральный исследовательский центр Уральского отделения Российской академии наук» (RU)*

Автор(ы): *Лебедев Егор Михайлович (RU)*

Заявка № 2025692156
Дата поступления 19 ноября 2025 г.
Дата государственной регистрации
в Реестре программ для ЭВМ *11 февраля 2026 г.*

Руководитель Федеральной службы по интеллектуальной собственности

Ю.С. Зубов

«Русско-удмуртский параллельный корпус: разработка инструментария и сопоставительные исследования» (грант РНФ № 24-78-10041, 2024–2027 гг.)

1. Подготовлен параллельный русско-удмуртский корпус (объемом 200 тысяч пар предложений).
2. Разработана система по автоматическому выравниванию русских и удмуртских предложений.
3. Подготовлена основа нейросетевой модели для исправления ошибок распознавания символов (OCR).
4. В 2026 г. запланирована разработка спеллчекера (программы проверки орфографии) на базе искусственного интеллекта.

Соглашение о научно-техническом сотрудничестве УдмФИЦ УрО РАН и ООО «Яндекс» (25.11.2024)

Задачи:

1) повышение качества русско-удмуртского и удмуртско-русского перевода в сервисе Яндекс Переводчик;

2) разработка и внедрение в сервисы Яндекса технологий синтеза и распознавания речи на удмуртском языке;

3) выполнение других проектов, направленных на расширение использования удмуртского языка в Интернете и разработку технологических решений, связанных с удмуртским языком (распознаватель удмуртского текста).

Глобальная задача в цифровизации удмуртского языка – создание программ на основе искусственного интеллекта!

УДМУРТСКИЙ ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР УРО РАН

Яндекс

Результаты обновления модели перевода для удмуртского языка в Яндексе (март 2026 г.)

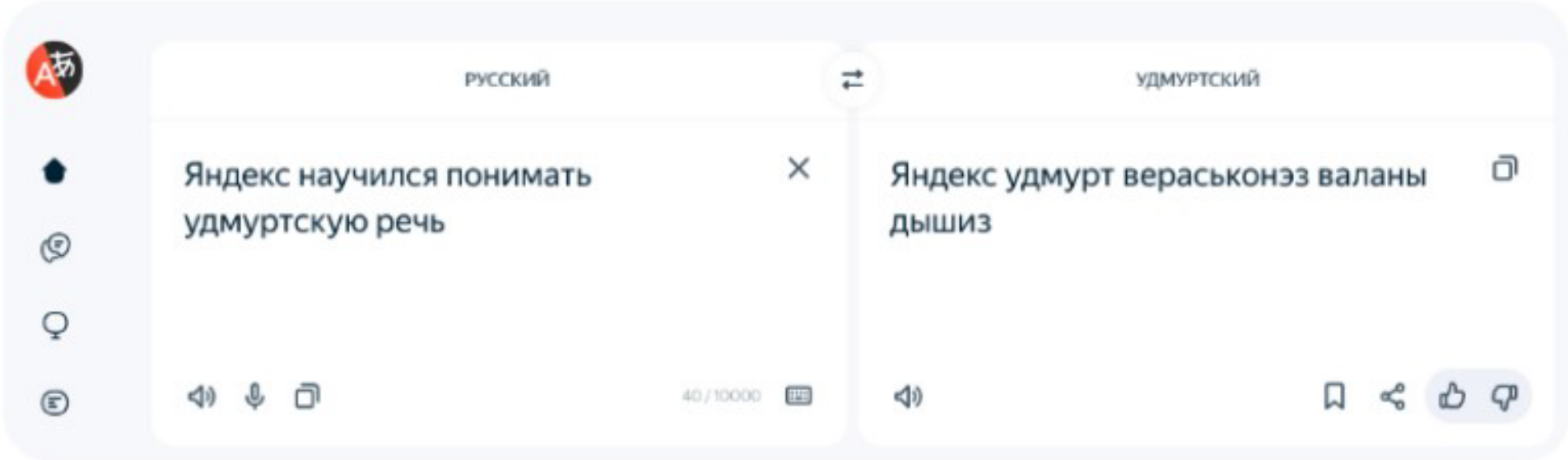
По замерам качество перевода с удмуртского на русский язык (udm-ru) и с русского на удмуртский язык (ru-udm) значительно улучшилось. Были получены следующие результаты:

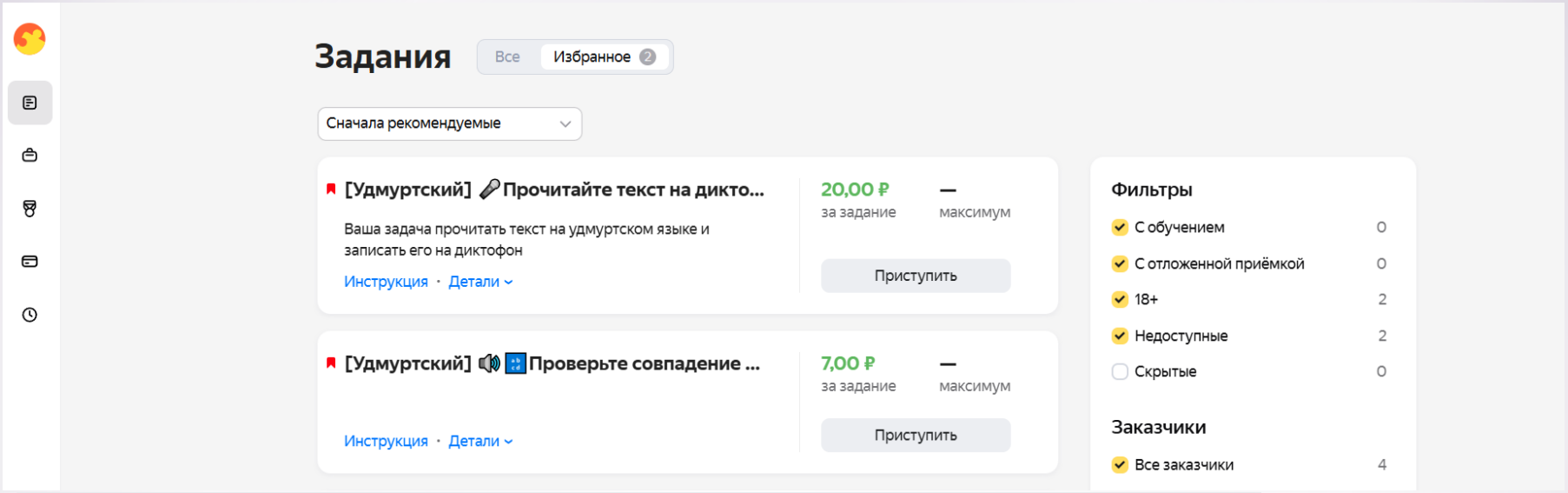
udm-ru: 6.86 -> 31.56 [BLEU]

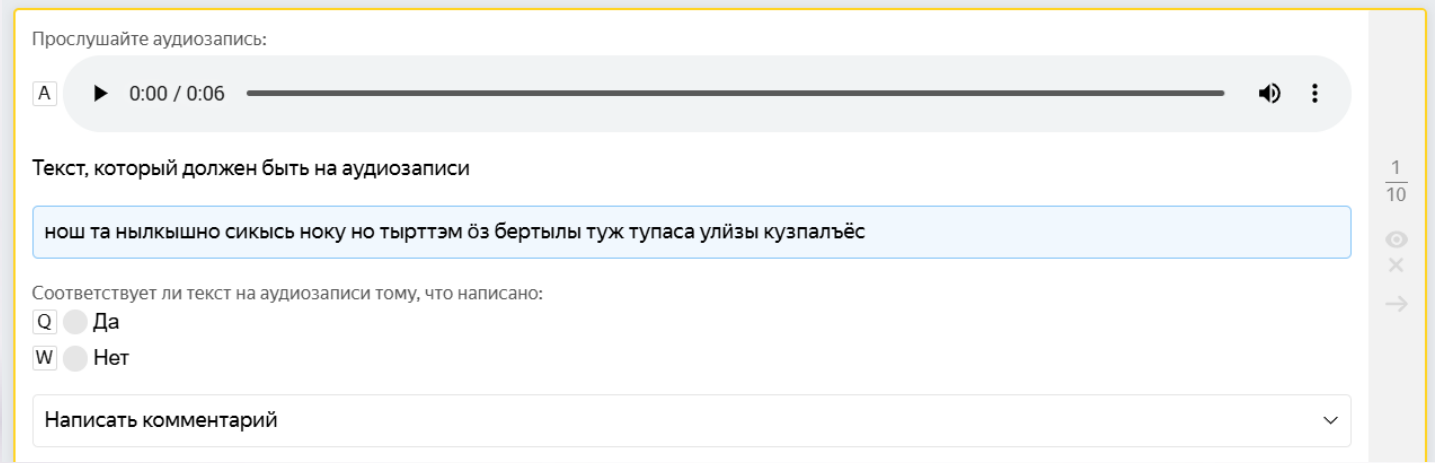
ru-udm: 2.45 -> 19.25 [BLEU]

Качество перевода с удмуртского на русский язык увеличилось в 4,5 раза, а с русского на удмуртский – в 7,86 раз.

Ко Дню языков народов России Яндекс добавил в свою технологию синтеза и распознавания речи поддержку удмуртского языка. Теперь пользователи Переводчика, Браузера и Клавиатуры могут переводить диалоги с удмуртского в режиме реального времени или использовать голосовой ввод, когда нет возможности печатать. Удмуртский стал первым языком среди финно-угорских и четвёртым среди всех языков народов России, для которого доступна такая возможность. Технология Яндекса также поддерживает три тюркских языка — татарский, чувашский и башкирский.







*Для **распознавателя удмуртской речи** сотрудники института осуществляли отбор носителей удмуртского языка с учетом диалектного разнообразия, проверку удмуртских фраз для озвучивания и оценку качества озвученных фраз.*

*При подготовке речевого датасета для **синтезатора удмуртской речи** осуществлялось исправление ошибок в текстах для озвучки, контроль процесса аудиозаписи.*

Удмуртский синтезатор говорит голосом Светланы Пикулевой

План дальнейшей работы в рамках цифровизации удмуртского языка

Пополнение Национального корпуса удмуртского языка

Проведение сопоставительных исследований и создание новых словарей (фразеологический словарь)

Создание продуктов на основе ИИ

Работа с Яндексом по повышению качества русско-удмуртского и удмуртско-русского перевода

Работа с Яндексом по созданию новых программных продуктов (распознаватель аудиозаписей, OCR)

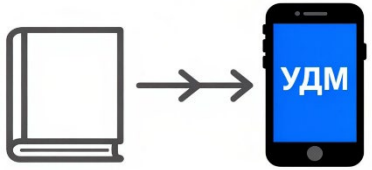
Основные проблемы цифровизации языков: взгляд из глубинки

– Трудности с наполнением датасета (для создания продуктов на основе ИИ нужны хорошие базы данных).

– Недостаточность финансирования.

– Сложность в организации работы волонтеров (многие просто не знают, где и как приложить свои усилия).

– Недостаточное количество высококвалифицированных IT-специалистов, готовых и умеющих работать с языками народов РФ.



ТАУ!
СПАСИБО!