ЯЗЫКИ КОРЕННЫХ НАРОДОВ:

НА ПУТИ В ЦИФРОВОЕ ПРОСТРАНСТВО

РУКОВОДСТВО ПО ЦИФРОВИЗАЦИИ ЯЗЫКА

TRANSLATION COMMONS

2022-2032 | INTERNATIONAL DECADE OF Indigenous Languages

Языки коренных народов: на пути в цифровое пространство

Авторы: Дебора В. Андерсон, Ли Коллинз, Крейг Корнелиус, Крейг Каммингз
Редакторы и соавторы: Эндрю Оуэн, Джулиа Ни, Лоуренс Вольф-Сонкин, Анна Луиза Дайно, Джули Андерсон, Дэниел Богре Юделл.
Оформление и подготовка к выпуску: Метте Аттар, Джоанна Бем, Леонидас Папас
Координация проекта: Эстер Перез, Джаннет Стюарт

1. ПРЕДИСЛОВИЕ

Translation Commons — некоммерческое сообщество волонтеров, деятельность которого направлена на поддержку процесса перевода языков в цифровой формат, наставничество представителей языковых профессий, а также предоставление образовательных и информационных ресурсов в сфере языковых услуг.

Одной из центральных программ Translation Commons является «Инициатива по цифровизации языков», которая дает возможность заинтересованным языковым сообществам получить доступ к инструментам цифровых технологий. В мире существует около 6 000 языков с незначительным присутствием в информационном цифровом пространстве или вовсе в нем не представленных. В «Инициативе по цифровизации языков» предлагается стратегический план, следуя которому любое сообщество сможет осуществить цифровизацию своего языка.

Для привлечения внимания к коренным народам и вопросам цифровизации их языков Translation Commons сотрудничают с ЮНЕСКО в рамках инициативы «2019: Международный год языков коренных народов». Частью миссии «Инициативы по цифровизации языков» является обеспечение равных возможностей доступа к цифровым технологиям для носителей языков коренных народов и других языковых меньшинств с целью повышения сетевой активности таких сообществ и предоставления им возможности пользоваться современными компьютерными приложениями на своем родном языке. В созданных руководствах и информационных ресурсах по цифровизации письменности и увеличению доли присутствия языков коренных народов в сети Интернет сообществам даются необходимые знания для самостоятельного преобразования языка в цифровой формат. Помимо руководств, Translation Commons предоставляет обучающие материалы, проводит семинары и помогает языковым сообществам установить рабочие связи со специалистами отрасли, которые способны направить их в процессе стандартизации.

1.1 На пути в цифровое пространство

Данный документ входит в серию методических руководств под названием *«На пути в цифровое пространство»*, в которой комплексно рассматривается деятельность по цифровизации языков. Авторами этих руководств выступили эксперты в сфере языковых технологий и лингвистики. Этот документ предназначен для любого языкового сообщества, для которого возможное применение родного языка в цифровых системах представляет интерес.

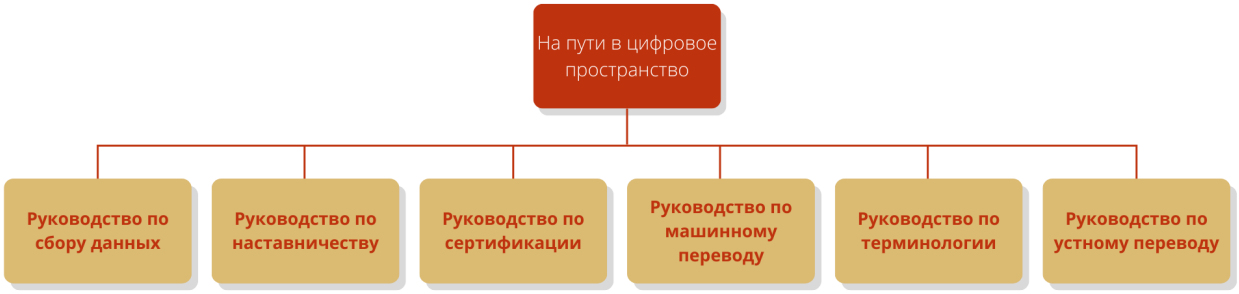
Присутствие языка в сети Интернет открывает новые каналы общения для его носителей. В разделе *приложения «Преимущества цифровизации языков»* подробно рассматривается, каким образом от присутствия языка в цифровом

пространстве выигрывает как коренной народ-носитель языка, так и мировое сообщество в целом.

Более подробная информация о процессе цифровизации языка приведена в разделе «На пути в цифровое пространство. Как привести ваш язык в Интернет». На вебсайте Translation Commons в подразделе проекта Language Digitization (Цифровизация языков) в разделе Resources (Ресурсы) размещены дополнительные материалы, посвященные «Инициативе по цифровизации языков» (руководства, презентации, видео и прочие документы).

На нижеследующем рисунке приведен перечень всех руководств, созданных для помощи коренным народам в данном вопросе.

Рисунок 1. Серия руководств «На пути в цифровое пространство»



В настоящем документе подробно рассматриваются способы поддержки письменного языка мобильными и настольными приложениями. Следуя данным рекомендациям, носители языков коренных народов смогут наладить общение в сети Интернет, обмениваться знаниями и документами, а также пользоваться ранее недоступными им приложениями и устройствами.

Документ адресован:

- Членам общин коренных народов, которые хотели бы иметь возможность пользоваться программным обеспечением для мобильных устройств и компьютеров на родном языке;
- Техническим специалистам, занимающимся цифровизацией одного или нескольких языков;
- Организациям, желающим создать языковые сообщества.

Целью данного документа является предоставление справочной информации о необходимых инструментах цифровизации, а также инструкций по их использованию. Дополнительно в документе приведена информация о том, какие возможности уже существуют для взаимодействия в сети Интернет на вашем родном языке.

Отвечая на вопрос о способах взаимодействия в сети Интернет на родном языке, важно отметить существование множества возможных ответов. Для начала использования языка в цифровом пространстве необходимо знать о существовании нескольких уровней в сфере информационных технологий как для веб-приложений, так и для мобильных версий.

Настоящий документ фокусируется на использовании письменной формы языков в сети Интернет с целью общения, чтения или творчества. В эту категорию входят обсуждения, общение с помощью текстовых сообщений, электронной почты, социальных сетей и блогов. Главная задача документа состоит в том, чтобы помочь разработчикам веб-сайтов и создателям разнообразного контента укреплять каналы общения как местных сообществ, так и языковых диаспор по всему миру. У каждого языка может быть несколько шрифтов или общепринятых норм. Наша цель не в том, чтобы предписывать способы использования цифровых технологий, а в том, чтобы предоставить возможность использования описанных технологий для повышения престижа, общественного восприятия и расширения сферы применения языков коренных народов по всему миру. Несмотря на то, что традиционные формы документирования, такие как грамматическая система языка и словари, являются важными составляющими лингвистических исследований и процессов стандартизации, они отнюдь не являются необходимым условием для использования языка в сети Интернет.

Настоящая коллегиальная группа состоит из Координационного комитета, курирующего реализацию, узкоспециальных групп для проведения консультаций, а также активных партнеров.

2. ОБЗОР ПРОЦЕССА

В документе представлено два рабочих процесса. Задача первого процесса состоит в том, чтобы определить актуальный статус присутствия языка в цифровом пространстве. Второй рабочий процесс используется, чтобы разработать технические решения для перевода языка в цифровой формат. Каждый шаг на рисунке отражает соответствующий раздел документа, в котором приведена более подробная информация. Эти шаги носят исключительно рекомендательный характер, а некоторые из них можно осуществлять параллельно друг с другом.

2.1. Рабочий процесс определения статуса языка

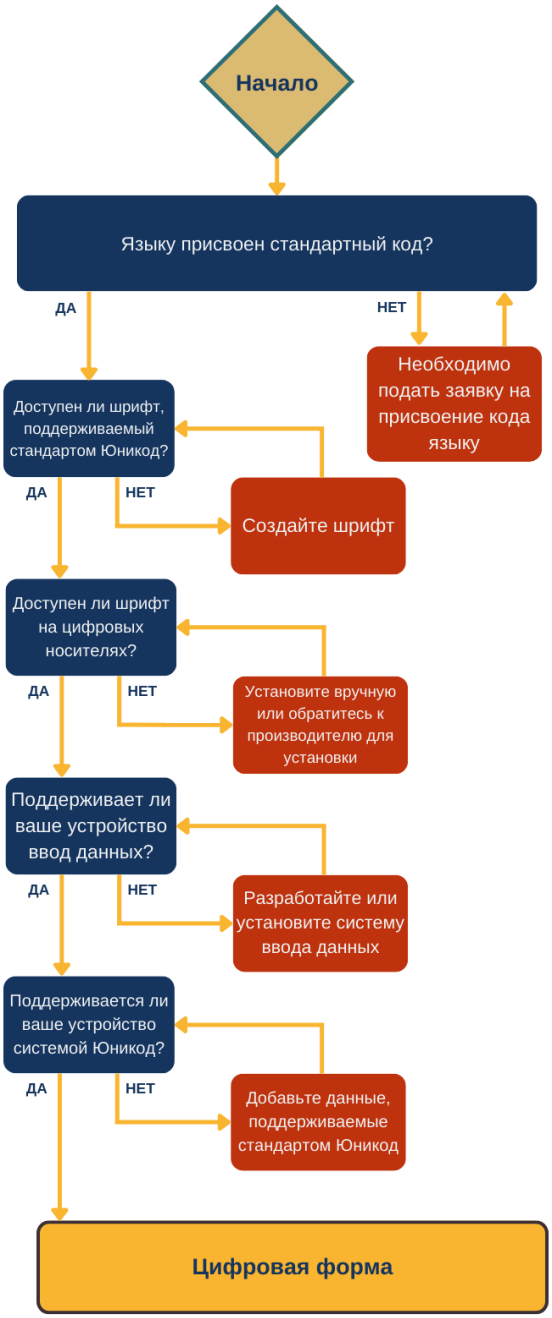
Данный рабочий процесс (рисунок 2) описывает шаги для определения актуального статуса языка при подготовке его перевода в цифровой формат.

Рисунок 2. Определение статуса языка

2.2. Рабочий процесс по внедрению технических решений

Зная статус вашего языка, можно приступить к данному рабочему процессу (рисунок 3) для определения доступных технических решений по цифровизации языка.

Рисунок 3. Внедрение технических решений



3. СТАТУС ЯЗЫКА

В данном разделе приведено пошаговое описание определения статуса языка. Целью процесса является определение текущего уровня поддержки вашего языка на компьютерах и мобильных устройствах. Также предлагаются и последующие шаги, необходимые для вывода вашего языка в сеть Интернет. Для ответа на эти вопросы вам не потребуются специальные технические знания.

3.1. Говорит ли на этом языке какое-либо сообщество?

Да: Есть ли у этого языка письменная форма?
Нет: Предназначен ли язык для активного употребления внутри сообщества?

3.2. Предназначен ли язык для активного употребления внутри сообщества?

Да: Возродить язык
Нет: Есть ли у этого языка письменная форма?

3.2.1. Возродить язык

При наличии общественной инициативы и соответствующей организации возможно возродить даже те языки, носителей которых больше нет в живых. Для этого потребуются ресурсы по документации языка, а также технические материалы, преподавательский состав и значительные усилия общественности для поддержания этой деятельности на протяжении нескольких лет, а возможно десятилетий.

При возрождении языка маори в Новой Зеландии, также известного как те рео, частично применялся подход, легший в основу возрождения валлийского языка в Уэльсе. Использование языка коренных народов в средствах массовой информации, специализированных школах и т.п. Для возрождения ирландского языка использовался подобный подход. Другие примеры подобных проектов по всему миру:

- **Иврит**. Пример возрождения языка, на котором не говорили столетиями.
- **Чероки**. Общественность и органы власти ведут активную деятельность для возрождения языка, например, открывают школы с языковым погружением для детей на протяжении всего школьного обучения, а также занятия для взрослых.
- **Муцун**. Программа возрождения языка племени амах муцун, живущего в Калифорнии.

- **Чакма**. Один из языков, на котором говорят в Бангладеш и Восточной Индии. При общественной поддержке письменность начали использовать в сфере образования и в борьбе с неграмотностью.
- **Туника.** Праязык племени туника-билокси ушел в забвение в 1948 году. Его возрождением занимаются потомки носителей языка, которые обучают новых носителей и призывают других изучать этот язык методом погружения.
- **Корнский.** Первые попытки по возрождению корнского языка были предприняты в начале 1900-х годов, но настоящий размах обрели в 2000-х, когда носители корнского языка обратились к интернет-форумам, чтобы найти друг друга и привнести язык в ежедневное общение.

3.3. Входит ли язык в Реестр языков мира?

Если язык входит в реестр или список языков.

Да: Есть ли у этого языка письменная форма?
Нет: Языку присвоен стандартный код

3.4. Есть ли у этого языка письменная форма?

Есть ли у вашего языка письменная форма?

Да: Документировать язык
Нет: Разработать письменную форму

3.4.1. Разработать письменную форму

Бесписьменные языки настоящим документом не охвачены. Несмотря на это, у вас есть возможность применять язык, используя аудио и видео ресурсы Существуют организации, предлагающие поддержку и инструменты для преимущественно устных языков. Также полезными ресурсами могут стать лингвистические факультеты учебных заведений.

Аудио и видео ресурсы принесут гораздо больше пользы, если они находятся в открытом доступе для носителей языка и научных работников. Для распознавания записей при автоматическом индексировании необходимо использовать языковые тэги или коды. В этом вам помогут стандартные коды языков, например, IETF BCP-47.

3.4.2. Документировать язык

Ваш язык задокументирован с помощью грамматики, словаря или лингвистического исследования?

Да: Язык задокументирован
Нет: Язык не задокументирован

3.4.2.1. Язык задокументирован

Доступны ли словари, своды грамматических правил и другие лингвистические данные носителям языка? Каким образом можно расширить доступ к этой информации для общественности? Для этих целей может понадобиться разработка интернет-ресурсов и образовательных материалов для начальной школы, а также для школ с погружением в языковую среду. Важно, чтобы все права на интеллектуальную собственность по данным материалам принадлежали языковому сообществу. Необходимо приобрести или создать цифровые форматы для хранения данных материалов, чтобы при желании языкового сообщества можно было сделать их доступными в сети Интернет.

При отсутствии подобных ресурсов постарайтесь сделать доступными для языковых сообществ книги, фонды электронных материалов и другую информацию на языке.

3.4.2.2. Язык не задокументирован

Даже если язык еще не был задокументирован, возможно использование его в письменной или устной форме. Однако в этом случае будет ограничена автоматическая проверка правописания, поиск, а также интуитивный ввод текста.

3.5. Есть ли у языка единая письменная форма?

Письменная форма — это ряд правил по применению одного или более алфавитов для того, чтобы писать на определенном языке. Для некоторых языков возможно использование более одного алфавита для письма. Например, в сербо-хорватском языке допустимо как применение кириллического, так и латинского алфавита. Кроме того, большинство алфавитов используются для письма в более чем одном языке. К примеру, бирманский алфавит используется для письма в бирманском, шанском и монском языках.

Использует ли язык стабильно хотя бы один алфавит, в том числе для орфографии?

Да: Соответствует ли письменная форма языка стандарту?
Нет: Соответствуют ли буквы и символы существующему стандарту?

3.5.1. Соответствуют ли буквы и символы существующему стандарту?

При отсутствии единой системы письменность может носить неформальный характер, пренебрегать орфографическими или грамматическими правилами (вплоть до использования более чем одного алфавита). Несмотря на то, что современные инструменты делают создание и чтение такого текста возможным, ограниченным будет применение грамматических и орфографических ресурсов. Необходимо понимать, что в случае конкурирующих письменностей, каждый вид правописания отражает интересы различных групп внутри сообщества.

Вам может понадобиться помощь лингвистов и технических специалистов:

- Во многих сообществах используют различные виды правописания и алфавиты для письма, включая диалектные особенности. Это необходимо учитывать при разработке таких языковых инструментов, как клавиатура, чтобы было учтено использование более чем одного алфавита, диакритических знаков и вариантов написания, отвечающих нуждам всех сообществ.
- С технической точки зрения нужно найти способы сделать идентификацию и поиск подобных вариантов в сети Интернет удобнее, а возможно даже наладить способ переключения между этими вариантами, если это необходимо членам языкового сообщества.
- Если поддержка алфавита уже налажена, то задача технических специалистов заключается в том, чтобы совместно с сообществом идентифицировать и разработать шрифты и клавиатуру.
- В случае, если отсутствует единый вариант языка, местные деятели образования, политики, лидеры языкового общества и лингвисты могут помочь в выборе наиболее часто употребляемого письменного варианта языка для использования в цифровом пространстве.

3.6. Соответствует ли письменная форма языка стандарту?

Включает ли письменная форма языка общепринятые орфографические и грамматические правила (формальные или неформальные)?

Неформальный стиль общения и письма возможны даже в случае, если они не соответствуют стандарту. И все же нестандартная орфография, грамматика, словарный запас или сильные региональные различия могут затруднить использование социальных сетей на более высоком уровне, обмен документами, поиск веб-сайтов и информации с помощью онлайн-поиска, а также других инструментов.

Да: Подайте заявку на добавление новых символов
Нет: Разработайте стандарт

3.6.1. Подайте заявку на добавление новых символов

Единые правила правописания, пунктуации и прочие аспекты письменной формы языка расширяют возможности членов языкового сообщества по использованию существующих интернет-сервисов, например, автоматической проверки правописания, интуитивного ввода текста, поиска информации и приложений для письма.

Стандарт Юникод выделяет алфавиты, а не языки. Сразу несколько языков могут использовать один и тот же алфавит. К примеру, английский, индонезийский языки и суахили используют латиницу. Если символы вашего языка поддерживаются стандартом, то можно переходить к следующему рабочему процессу по внедрению языковых технологий.

Если же символы существующей письменной формы языка не входят в стандарт Юникод, целесообразно унифицировать такие символы, чтобы их можно было использовать на устройствах с выходом в сеть Интернет, мобильных устройствах, персональных компьютерах и ноутбуках. Даже без унифицированного алфавита возможен обмен файлами между пользователями, чьи шрифты и способы ввода текста одинаковы. Интернет-инструменты и сервисы, однако, будут ограничены.

Для соответствия стандарту Юникод необходимо:

1. Установить, включены ли ваши символы в стандарт Юникод.
2. Если поддерживаются не все символы, то нужно подготовить и отправить заявку на их добавление.

3.6.2. Разработать Стандарт

Как известно, у языка может быть несколько письменных форм, и разработка стандарта не является обязательным шагом. В случае, если создание единой письменной формы отражает интересы сообщества, то разработка правил правописания, пунктуации и грамматики даст членам данного сообщества возможность обмениваться идеями в цифровом пространстве. Это также улучшит функцию клавиатур по проверке правописания и позволит использовать интуитивный ввод текста. Помимо этого, поисковики и прочие интернет-сервисы смогут предлагать более подходящие результаты поиска.

3.7. Переход к следующему шагу

Стандарт Юникод поддерживает символы языка. Поддерживаемый алфавит включает в себя все символы письменной формы языка. Теперь можно переходить к рабочему процессу по внедрению языковых технологий.

4. РАБОЧИЙ ПРОЦЕСС ПО ВНЕДРЕНИЮ ЯЗЫКОВЫХ ТЕХНОЛОГИЙ

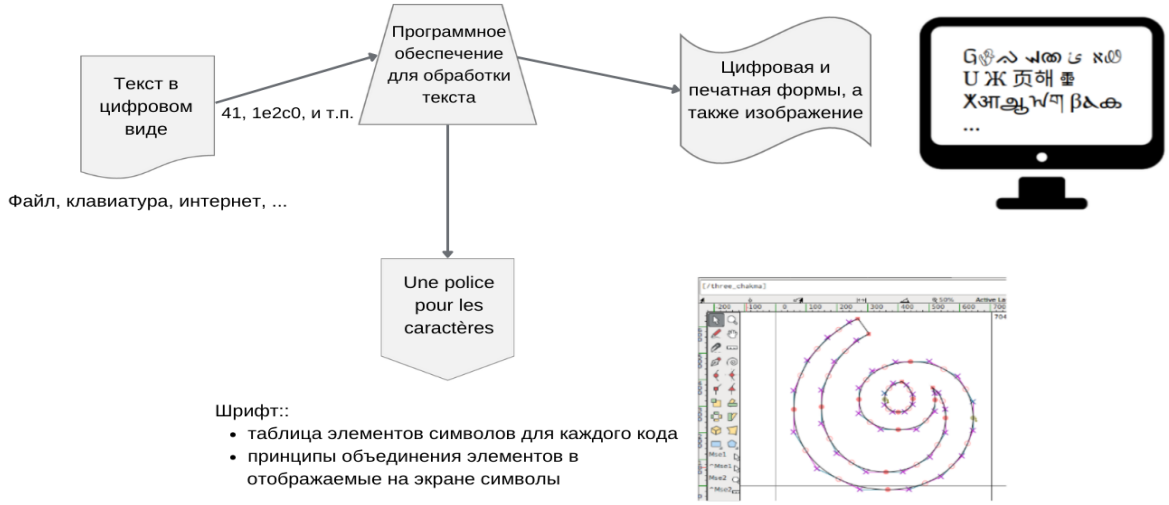
4.1 Примечания о технических решениях для текстов в цифровых системах

Текст в цифровых устройствах хранится в виде наборов битов, используемых в качестве *кодовой точки*. К примеру, кодовая точка U+0041 может обозначать букву «A». Этот код может быть создан с помощью клавиатуры, храниться в файле или пересылаться в другое приложение, отображаться или *обрабатываться* системой, которая понимает систему кодирования символов.

Стандарт кодирования присваивает значение каждому возможному коду. К примеру, стандарт ISO-8859-1 характеризует символы от 0 до 255 с указанием значения для каждой кодовой точки. Юникод — это стандарт кодирования символов, включающий в себя знаки из разных систем письменности. Каждому символу из более чем 150 различных письменностей, включая латиницу, кириллицу, иврит, деванагари, китайский, арабский, тамильский, мьянманский алфавиты и многие другие, присваивается уникальное значение.

Текст, который был написан с помощью определенного кодирования, может быть набран с использованием совместимого способ ввода и показан на экране с использованием совместимого шрифта. Нижеприведенная схема (рисунок 4) демонстрирует, каким образом система, подобная Юникод, может отображать символы разных языков.

Рисунок 4. С помощью Юникод возможно отображать символы разных языков



Большинство алфавитов состоят из базовых символов с добавлением диакритических знаков или модификаторов, таких как «e», который с добавлением

восходящего акцента образует «é». Кодировка позволяет сочетать базовый символ с диакритическим знаком для образования комбинированного символа. В некоторых случаях в кодировании могут использоваться составные знаки, состоящие из кодовых точек, в которые инкорпорирована не только база, но и комбинированные кодовые точки.

Для использования письменной формы любого языка на цифровом носителе необходимо следующее:

- Стандартная система кодирования символов, например, Юникод,
- Графическая поддержка символов языка, в том числе и шрифтов, а также
- Способы или компьютерные приложения для воспроизведения кодовых точек в различных носителях информации.

В этом документе мы делаем акцент на стандарте Юникод, поскольку он используется во всех современных цифровых устройствах. Также он является стандартным способом кодирования для многих источников текста. С помощью этого стандарта процесс цифровизации языка значительно упрощается.

4.2. Терминология, связанная с цифровой поддержкой

В этом разделе приведена справочная информация по внедрению цифровой поддержки письменной формы языка. Во многих языках применяются нестандартные шрифты и способы написания. Это не мешает использованию языка в сети Интернет, но ограничивает уровень доступной для языка поддержки. Следующие шаги описывают аспекты цифровой поддержки письменных языков в сети Интернет.

Основные требования к письменной форме языка в сети Интернет:

- **Символы**: необходимо определить используемый в орфографии языка набор букв, диакритических знаков, лигатур, знаков пунктуации, цифр, иероглифов и т. п.
- **Идентификатор ISO**: для языка и его письменности. Сюда могут входить региональные идентификаторы, например, *es-MX* для испанского языка, на котором говорят в Мексике.
- **Стандарт кодирования**: может использоваться как традиционный стандарт (Юникод с сортировкой, упорядочением и сочетанием символов), так и неформальный стандарт (кодирование шрифтов).
- **Поддержка шрифтов**: вся необходимая информация для создания типографского шрифта, охватывающего все требуемые элементы символа:
    - Элементы воспроизведения текста, такие как лигатуры и комбинированные элементы (эти элементы могут быть предоставлены системой Юникод).
    - Образцы текста для тестирования шрифтов дизайнерами и разработчиками.

Может возникнуть необходимость обновления браузерного движка и прочего программного обеспечения.

- **Методы ввода**: способ ввода символов на всех цифровых устройствах. Это может быть как физическая клавиатура, так и цифровая.

Также доступны дополнительные языковые ресурсы по теме:

- Общий репозиторий данных локали, включающий дополнительную информацию о языке, например, календари.
- Стандартные коды языков, например, IETF BCP-47.
- Поддержка компьютерными приложениями сегментации слов:
  - Словари, написанные без явных пробелов.
  - Правила, в том числе употребления интервалов и пунктуации.

4.3. Языку присвоен стандартный код?

Существует ли доступный для языка стандартный код, применение которого не вызывает разногласий?

Да: Доступен ли шрифт, поддерживаемый стандартом Юникод?
Нет: Необходимо подать заявку на присвоение кода языку

4.3.1. Необходимо подать заявку на присвоение кода языку

При отсутствии доступного языкового кода вам потребуется разработать стандартный языковой код с возможностью добавления региональных вариантов и алфавитов. Применение нестандартных языковых кодов и тэгов может вызвать путаницу и привести к ошибкам в идентификации языкового источника. Следуйте методическим рекомендациям, приведенным в стандарте ISO-639 и IETF BCP-47.

4.4. Доступен ли шрифт, поддерживаемый стандартом Юникод?

Доступен ли к шрифт, поддерживаемый стандартом Юникод, членам языкового сообщества для работы на компьютере или ноутбуке?

Да: Доступен ли шрифт на цифровых носителях?
Нет: Создайте шрифт

4.4.1. Создайте шрифт

При первоначальном добавлении алфавита в Юникод, шрифтов, поддерживающих его символы, может еще не существовать. В этом случае будет полезным разработать поддерживаемые шрифты совместно с графическими дизайнерами и другими техническими специалистами.

4.5. Доступен ли шрифт на цифровых носителях?

Доступен ли шрифт на цифровых носителях, которыми пользуются члены сообщества? В этом случае они смогут читать информацию на сайтах, в социальных сетях и прочих приложениях.

Важно понимать, что большинство компьютеров и мобильных устройств эффективнее работают с текстом или способами ввода текста, которые совместимы с системой Юникод.
Удостоверьтесь, доступен ли шрифт в коллекции Noto или на других сайтах, предоставляющих шрифты, поддерживаемые Юникод. Имейте в виду, что на многих мобильных устройствах скачивание и установка шрифтов напрямую может не поддерживаться. Настольные компьютеры и ноутбуки предоставляют возможность скачивать и устанавливать шрифты.

В некоторых текстовых редакторах или браузерах вам может потребоваться совершить дополнительные настройки приложения, чтобы шрифт отображался в документах и на интернет-страницах.

Web-шрифты используются сайтами для изображения текста с помощью шрифта, предоставленного самим сайтом. Авторы информации на таких сайтах могут выбрать определенный шрифт текста, даже в случае, если он не поддерживается устройством напрямую. Следует помнить, что веб-шрифт работает только на тех страницах, где он настроен, и его использование не гарантирует его постоянную доступность на вашем устройстве. Последняя версия шрифтов коллекции Noto доступна в виде веб-шрифтов, что может быть удобным, если ваше устройство не поддерживает актуальную версию шрифта или его установка невозможна.

Да: Поддерживает ли ваше устройство ввод данных?
Нет: Установите вручную или обратитесь к производителю для установки

4.5.1. Установите вручную или обратитесь к производителю для установки

Системы письменности, стандартизированные относительно недавно, могут еще отсутствовать на мобильных устройствах, даже если шрифт есть в открытом доступе. Обратитесь к производителю устройств для добавления необходимых для языка шрифтов.

В качестве альтернативы можно вручную установить в устройстве шрифт, поддерживаемый стандартом Юникод. Этот способ достаточно эффективен, однако требует специальных технических знаний.

Внимание! Установка на мобильное устройство шрифта из сети Интернет может нанести вред конфиденциальности и безопасности данных, а также повлиять на гарантийное обслуживание и поддержку вашего устройства.

4.6. Поддерживает ли ваше устройство ввод данных?

Носителям языка нужен способ ввода данных для эффективного создания сообщений, электронных писем, интернет-публикаций и другой деятельности на своем языке. Предусмотрена ли на устройстве (мобильном или ином) встроенная поддержка ввода данных?

Да: Поддерживают ли приложения и устройства сторонних разработчиков систему ввода данных на вашем языке?
Нет: Разработайте систему ввода данных

4.7. Поддерживают ли приложения и устройства сторонних разработчиков систему ввода данных на вашем языке?

Входит ли система ввода данных в стандартный набор клавиатур? Если она поддерживается, то важно научиться ее включать. Несмотря на то, что этот процесс происходит по-разному на разных устройствах, в сети Интернет можно найти необходимую информацию по подключению поддержки клавиатуры.

Производители мобильных устройств предлагают множество клавиатур для операционных систем iOS и Android. Например:

- Gboard, клавиатура для мобильных устройств от Google (https://support.google.com/gboard/answer/6380730?hl=en&co=GENIE.Platform=Android) поддерживает многочисленные языки как на iOS, так и Android.
- Сервис Инструменты ввода данных от Google (https://www.google.com/inputtools/) предлагает виртуальные клавиатуры, которыми можно пользоваться на сайтах через браузер Chrome на компьютерах (мобильными устройствами этот сервис не поддерживается).
- В магазинах App Store и Google Play представлены многочисленные приложения для ввода данных от сторонних производителей.

Для многих языков уже существуют общедоступные раскладки клавиатуры, например в хранилище клавиатурных раскладок в общем репозитории данных локали (CLDR).

4.7.1. Разработайте систему ввода данных

Существует ряд инструментов для установки способов ввода данных для различных языков, а также для их разработки. Вот некоторые из них:

- Приложения для клавиатуры в магазинах App Store, Google Play и других источниках.
- Keyman — инструмент для языковых клавиатур (https://keyman.com/).

- Microsoft Keyboard Layout Creator (https://www.microsoft.com/en-us/download/details.aspx?id=102134).
- SIL Ukelele (https://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=ukelele).

Внимание! Установка клавиатуры или других способов ввода из сети Интернет может нанести вред конфиденциальности и безопасности данных.

4.8. Поддерживает ли ваше устройство систему Юникод?

Доступны ли данные, поддерживаемые стандартом Юникод?

Да: продолжайте использование вашего языка.
Нет: при необходимости добавьте необходимую информацию о вашем языке в общий репозиторий данных локали (CLDR). Сюда входит языковой тэг и название языка. В разделе Дополнительная поддержка языка вы найдете более подробную информацию об общем репозитории данных локали (CLDR).

5. ДОПОЛНИТЕЛЬНАЯ ПОДДЕРЖКА ЯЗЫКА

В процесс стандартизации для кодировки Юникод входят многие элементы обработки текста, созданного той или иной письменностью и языком. Благодаря поддержке шрифта и системы ввода данных будет возможна работа с языком. Сюда входят многие элементы обработки текста, работы с электронными таблицами и электронной почтой.

Шаги, приведенные в данном документе, являются лишь начальным этапом процесса. Общий репозиторий данных локали (CLDR) предоставляет «ключевые элементы, из которых строится поддержка программным обеспечением мировых языков», являясь коллекцией полезной информации о разных локалях (сочетании языка и региона). К этим данным относится информация о названиях языков, стран, месяцев, дней недели и иная информация. Эти данные позволяют также определить необходимый формат даты, времени, чисел и другую подобную информацию.

И хотя наличие информации о языке в Общем репозитории данных локали (CLDR) не является необходимым условием для повседневного письменного общения на языке коренных народов, ее наличие может значительно расширить функциональность языка. Наличие возможности ввода текста на клавиатуре и доступность шрифтов делают возможным эффективное использование электронной почты, сообщений, социальных сетей и прочих сервисов.

Такая информация необходима разработчикам программного обеспечения при создании интернет-приложений для локализации элементов пользовательского интерфейса, например, для надлежащего отражения календаря, электронных таблиц, чисел, пунктов меню и других элементов.

В Общем репозитории данных локали (CLDR) также содержится дополнительная информация о языке, например, используемые в письме символы и раскладка клавиатуры для набора текста. В разделе репозитория, посвященном клавиатурам, приведена более подробная информация.

Однако, для лучшей поддержки языка в сети Интернет и в мобильных приложениях могут понадобиться дополнительные данные и инструменты по обработке текста. Для того, чтобы интернет-поддержка языков коренных народов не уступала поддержке более распространенных языков, следует учитывать некоторые особенности:

- **Сегментация текста на слова и предложения.** Сегментация необходима для корректного отображения текста и его макета на странице, а также для деления на кластеры графем, слов и предложенй. Во многих языках нет четко выраженных указателей словораздела, таких как пробелов и знаков пунктуации. В таких случаях за необходимой информацией следует обращаться к словарям и алгоритмам. Подробнее по ссылке: userguide.icu-project.org/boundaryanalysis.
- **Правила переноса текста.** В языках существует множество правил, по которым текст может быть прерван для перехода на новую строку. К примеру, в разных языках цитирование текста подчиняется разным правилам, учитывающим характерные для языка символы и правила использования пробелов между словами и предложениями. Также свойства самих символов Юникод, применяемых в письме, влияют на место разбивки текста при переходе на новую строку. Кроме того, сочетания числа и знака валюты (например, $10) может потребоваться сделать слитными, чтобы их значение было понятным. В разных языках и регионах правила пунктуации тоже различаются. Правила переноса текста являются частью отдельного раздела Юникод по анализу словораздела (userguide.icu-project.org/boundaryanalysis#TOC-Line-break-Boundary).
- **Идентификация языка текста.** Документы должны быть помечены языковым кодом или другим идентификатором, указывающим на язык этого документа. Наличие этой информации в тексте позволяет цифровым инструментам предлагать его пользователям более подходящую информацию. В документе, который написан с использованием нескольких языков, возможно указывать язык текста для отдельных частей или даже параграфов. Важно использовать специализированные тэги, например, IETF BCP-47 (https://tools.ietf.org/html/bcp47):
    - В связи с тем, что в разных приложениях механизмы подобной идентификации могут отличаться, может потребоваться дополнительное обучение. Обратите внимание, что такая идентификация может быть осуществлена с помощью предоставленного списка опций, а не присвоением любого возможного тэга.
    - В языке HTML для указания языка любого компонента предусмотрен атрибут *lang* (https://www.w3schools.com/tags/att_global_lang.asp).

Значение этого атрибута должно входить в перечень стандартных наборов идентификаторов языка, а не создаваться пользователем. Например, для сербского языка следует использовать *rs*, для немецкого – *de*, для упрощенного китайского языка – *zh-Hans* либо *zh-CN* (или даже *zh*).

- **Распознавание языка.** Интернет-сервисы и прочие цифровые приложения предлагают более релевантные результаты в случаях, когда язык текста им известен. В случае отсутствия языковых тэгов вам помогут определители языка, например, cld2 (https://github.com/optimaize/language-detector). Эти инструменты предлагают возможный язык текста, основываясь на статистическом анализе его символов. Такой способ облегчает задачу определения языка, т. к. большинство алфавитов используются многочисленными языками. Например, латиница используется в суахили, вальбири, лакотском, финском языках; кириллица – в русском, украинском, казахском; бирманское письмо – в бирманском, шанском и монском языках.
- **Словари для текстовых редакторов.** Большинство текстовых редакторов позволяют создавать, редактировать текст, а также распечатывать и делиться им с другими. Интуитивный ввод текста, проверка правописания и грамматики и другие подобные инструменты за основу своей работы берут словарные базы с указанием частотности употребления, словари и прочие лингвистические данные. Для инструментов для поиска в интернете также будет полезной информация о синонимах и часто употребляемых выражениях.
- **Символы за пределами диапазона ASCII (Американский стандартный код для обмена информацией).** Во многих письменностях присутствуют цифры, написание которых отличается от западного. В качестве примера можно привести мьянманскую и арабскую письменность, а также аджам и фарси. В таких приложениях, как электронные таблицы, такие цифры распознаются не как числовая величина, а как текст. Разработчики такого рода приложений могут учитывать свойства Юникод подобных символов, чтобы обрабатывать их как цифры, однако такая поддержка не реализована последовательно (https://en.wikipedia.org/wiki/Numerals_in_Unicode).
- **Перевод пользовательского интерфейса.** Для ряда приложений, особенно в сфере образования и информации, будет полезным перевести текст, являющийся частью пользовательского интерфейса (UI), на язык пользователя. В качестве примера можно привести текст пунктов меню функций операционной системы, такие как «Пуск» или «Открыть файл». Относительно немного разработчиков приложения в состоянии перевести эту информацию на малораспространенные языки. Если пользовательский интерфейс доступен хотя бы на одном из языков, понятных пользователю, то наличие перевода пользовательского интерфейса менее критично.
- **Оптическое распознавание текста (OCR).** Во многих языках накоплены значительные фонды письменной литературы, хранящиеся в виде книг и других документов. Благодаря оптическому распознаванию текста возможно преобразовать этот текст в цифровой формат. В открытом доступе можно найти инструменты для оптического распознавания текста (https://pdf.iskysoft.com/ocr-pdf/open-source-ocr.html), которые можно обучить

работе с новыми письменностями. Следует обратить внимание на то, что наличие языковой модели с базой часто употребляемых слов значительно повышает точность распознавания текста.

5.1. Общедоступные языковые ресурсы

В распоряжении языковых сообществ есть множество языковых ресурсов, как бесплатных, так и доступных по подписке.

5.1.1. Языковые ресурсы

- Panlex (https://panlex.org/)
- Wikitongues (https://wikitongues.org/)

5.1.2. Инструменты

- SIL - Keyman
- Инструменты для работы со шрифтами
- Инструменты с открытым кодом для составления словарей

5.2. Продвинутые языковые технологии

Для настройки функций, описываемых ниже, требуется значительное количество данных для обучения алгоритмов. Несмотря на появление общедоступного программного обеспечения с открытым кодом, значительная часть работы все же сосредоточена в академических исследованиях и в секторе разработки коммерческих продуктов. Кажется маловероятным, что многие из этих функций станут доступны для большинства языков.

- **Преобразование речи в текст.** Распознавание человеческой речи и преобразование ее в текстовый формат. Подобный текст возможно преобразовать в документ или использовать его для управления приложениями и устройствами.
- **Преобразование текста в речь**. Создание звучащей речи из текста. Эта функция может использоваться для бесконтактных интерфейсов и для устройств, зачитывающих текст с письменного источника.
- **Транскрипция аудиоматериалов в академических целях.** Важна для документации языков, особенно для академических исследований в лингвистике. Некоторые проекты с открытым кодом нацелены на решение этой задачи. К примеру:
- **Ускоренная транскрипция для лингвистов**: https://github.com/CoEDL/elpis
- **Машинный перевод.** Задача преобразования текста с одного языка на другой является одной из самых сложных для компьютера. Существующие системы способны производить перевод лишь для ограниченного перечня языков.

Однако, подобные системы не воспринимают весь контекст текста и не способны сравниться с человеком по качеству перевода. Технологии машинного перевода увеличивают качество подобного перевода при наличии большого корпуса языка. Поддержка машинного перевода для языков коренных народов не является широко доступной, но начинают появляться проекты с открытым исходным кодом и академические инициативы. К примеру, www.apertium.org это инструмент в сети Интернет, поддерживающий машинный перевод для малораспространенных языков.

6. ГЛОССАРИЙ

ASCII – американский стандартный код для обмена информацией – кодирование символов для электронного общения.

BCP-47 – *IETF* тэг для идентификации языков.

CLDR – общий репозиторий данных локали. Содержит дополнительную информацию о языке.

Символ – наименьший компонент письменной формы языка, обладающий семантическим значением; характеризуется образным значением, а не только графическим изображением. Смотрите также: Глоссарий Юникод

Кодовая точка – число, представляющее определенный символ или способ форматирования.

Диакритический знак – символ, добавляемый к букве или знаку с целью модифицировать его звучание или семантическое значение.

Мертвый язык – язык, носителей которого не осталось в живых.

Шрифт – группа знаков для визуального изображения символов.

Знак – единица набора символов, представляющая собой читабельный знак с целью письма.

Грамматика – правила, регулирующие строение *естественного языка*.

IETF – инженерная рабочая группа интернета.

Язык коренных народов – родной для определенного региона язык.

IYIL2019 – 2019: Международный год языков коренных народов.

Возрождение языка – предотвращение исчезновения языка или возвращение к жизни *мертвого языка*.

Лигатура – сочетание двух или более знаков для создания единого знака.

Естественный язык – язык, который получил естественное развитие с течением времени в отличие от формальных языков, применяемых в вычислительных

устройствах. Среди естественных языков выделяют устные, визуальные, визуально-мануальные (знаковые) и письменные языки.

Noto – коллекция шрифтов, состоящая из более 100 отдельных шрифтов, которые в совокупности соответствуют всем алфавитам, закодированным в системе Юникод (алфавиты Юникод версии 6.0 и более ранних версий).

Орфография – набор правил правописания.

PUA – область личного пользования — диапазон *кодовых точек* в системе *Юникод*, которые не закреплены за определенными символами.

Пунктуация – пробелы и символы, которые не обозначают звуки, но помогают пониманию текста.

Алфавит – набор букв и прочих символов, используемых в письме одной или несколькими письменностями.

Translation Commons – интернет-сообщество и платформа, целью которой является свободный обмен лингвистическими знаниями.

ЮНЕСКО – учреждение Организации Объединенных Наций по вопросам образования, науки и культуры.

Юникод – наиболее распространенный стандарт для цифрового кодирования символов языков мира.

Письменность – это ряд правил по применению одного или более алфавитов для того, чтобы писать на определенном языке.

7. СПРАВОЧНЫЕ МАТЕРИАЛЫ

7.1. Возрождение языка

Руководство по возрождению языка от издательства «Рутледж» (Routledge Handbook of Language Revitalization. Hinton, Huss, & Roche 2018)
Зеленая книга о возрождении языков на практике (The Green Book of Language Revitalization in Practice. Hinton & Hale 2001)
Документация и возрождение языков в контексте стран Латинской Америки (Language Documentation and Revitalization in Latin American Contexts. Perez-Baez, Rogers, & Roses Labrada 2016)
Создание орфографии для бесписьменных языков (Developing Orthographies for Unwritten Languages. Cahill & Rice 2014)
http://cherokeepreservation.org/what-we-do/cultural-preservation/cherokee-language/
https://language.cherokee.org/
http://amahmutsun.org/language
https://rising.globalvoices.org/blog/2011/11/29/languages-online-activism-to-save-chakma-language/
https://www.languageconservancy.org/programs/indigenous-language-program-support/

7.2. Классификация языков

https://www.ethnologue.com/
https://glottolog.org/

7.3. Юникод и кодирование шрифтов

https://unicode.org/main.html
https://unicode.org/standard/supported.html
https://unicode.org/standard/where/
https://unicode.org/pending/proposals.htm
https://unicode.org/glossary/
https://linguistics.berkeley.edu/sei/

Для создания нестандартных шрифтов необходим нетрадиционный подход к созданию символов, например, области личного пользования (PUAs) и другие диапазоны символов (ASCII или арабские знаки) с настраиваемыми символами для кодовых точек. Этот процесс называется кодирование шрифтов. Благодаря такому подходу пользователям видны символы во время набора текста, хотя для всех остальных их не видно, если они не пользуются теми же шрифтами. При такой кодировке шрифтов сервисы и инструменты в сети Интернет не смогут обработать текст надлежащим образом в силу отсутствия информации о значении символа.

Хотя текст, основанный на кодировке Юникод, имеет преимущества перед текстом, закодированным шрифтом, может потребоваться использование специализированных шрифтов до тех пор, пока символы в системе письма не будут включены в стандарт Юникод. В таких случаях следует использовать только знаки из диапазона области личного пользования (PUAs) в Юникоде, а не кодовые значения, предназначенные для других алфавитов. Этот способ поможет избежать наложения кодовых значений и упрощает использование существующих алфавитов. Дальнейший перевод такого шрифта в Юникод будет осуществляться гораздо легче при условии последовательного использования кода из области личного пользования (PUAs).

7.4. Коды языков

https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

7.5. Шрифты

https://www.google.com/get/noto/

Инструменты для создания шрифтов:
- FontForge
- FontLab
- Glyphs
- BirdFont (https://birdfont.org/)

Университ г. Рединг предлагает программу магистратуры по дизайну шрифтов (typefacedesign.net/), и вы можете обратиться к его студентам за помощью в создании нового шрифта Юникод.

Создатели шрифтов могут разработать новый шрифт Юникод на коммерческой основе.

8. ЗАМЕТКИ

Настоящий документ охватывает лишь языки с существующей письменностью. Помимо этого, существуют другие формы общения:

- Эмоджи
- Устные языки
- Визуально-мануальные (знаковые) языки
- Визуальные языки

Языки со своим алфавитом могут быть также быть написаны с использованием других алфавитов. Например:

- Изначально письменность турецкого языка использовала арабский алфавит, но в ее современной форме применяет латиницу.
- Для текстов на китайском языке можно использовать латиницу (пиньинь).

Этот документ не рассматривает диалекты языков, хотя при подаче заявки на добавление языка в стандарт их не следует обходить вниманием.

Для продвинутого поиска на языке следует учитывать некоторые особенности:

- Идентификация языка (на основании текста)
- Сегментация – дробление текста на слова
- Определение *основы* слова в языках с различными формами слов, например, *домашний*, *дома*, *дом*.

И хотя устные языки настоящим документом не охвачены, следующие ресурсы могут быть полезны:

- Google Планета Земля (https://docs.google.com/forms/d/e/1FAIpQLSdphaDaz33syPoUDyTOTwwkaLWZx90zopUklha4uadfkUKG8A/viewform)
- https://www.blog.google/products/earth/indigenous-speakers-share-their-languages-google-earth/
- https://www.gerlingo.com/
- XTrans (https://www.ldc.upenn.edu/language-resources/tools/xtrans)—это многоплатформенный, многоязыковой и многоканальный инструмент для

транскрипции, позволяющий вручную транскрибировать и аннотировать аудиозаписи.