НА ПУТИ В ЦИФРОВОЕ ПРОСТРАНСТВО:

РУКОВОДСТВО ПО СБОРУ ЯЗЫКОВЫХ ДАННЫХ

РУКОВОДСТВО ПО ЦИФРОВИЗАЦИИ ЯЗЫКА

TRANSLATION COMMONS

2019 | INTERNATIONAL YEAR OF Indigenous Languages

РУКОВОДСТВО ПО СБОРУ ЯЗЫКОВЫХ ДАННЫХ

Авторы: Julie Anderson и Tex Texin
Соавторы: Grigory Sapunov, Jeannette Stewart, Kirti Vashee, Debbie Anderson
Редакторы: Andrew Owen, Akil Iyer, Shuto Kato, Paula Cirilo
Графика и маркетинг: Leonidas Pappas

Мы открыты для предложений по улучшению руководств.
Напишите нам письмо по электронной почте krista@translationcommons.org

Содержание

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.

Данный документ входит в серию методических руководств под названием *«На пути в цифровое пространство»*, в котором комплексно рассматривается деятельность по цифровизации языков. Авторами этих руководств выступили специалисты по вычислительной технике и лингвисты. 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).

1.1 Об этом документе

Настоящий документ создан для того, чтобы:

- Перечислить типы языковых данных, которые нужно собрать, чтобы перевести язык в цифровое пространство.
- Описать процесс сбора данных.
- Соотнести различные виды языковых данных с их технологическим использованием.

Для цифровизации языка требуется большое разнообразие примеров использования языка. Лингвисты и специалисты по вычислительной технике используют эти примеры для изучения и анализа при разработке правил и компонентов для поддержки языка в цифровых системах. Некоторые примеры

включают список символов в алфавите, различные способы написания символов, а также списки слов языка и их значения. В дополнение к этому, для надежной цифровой поддержки языка также требуется множество других типов языковых данных.

Цифровые системы поддерживают множество видов приложений. Для работы одних приложений требуется соблюдение простых требований к языковым данным. Другие же – предъявляют более высокие требования. Например, приложению «блокнот» может потребоваться только поддержка простого ввода и отображения текста. А приложение для обработки текстов может предлагать сложные варианты компоновки и типографии, проверку орфографии и грамматики, лингвистическую сортировку, выделение и другие функции, зависящие от языка. В этом документе будут описаны различные требования к данным для целого ряда приложений.

2. ОБЗОР ПРОЦЕССА СБОРА ЯЗЫКОВЫХ ДАННЫХ

В данном руководстве описываются этапы по сбору элементов данных, которые носят типовой характер и предоставляют методическую информацию для цифровизации языка, а также шаги по обеспечению доступности таких элементов для просмотра в цифровом репозитории. К этим этапам относятся:

- Привлечение к сбору языковых данных заслуживающих доверия источников.
- Получение языковых данных.
- Регистрация источника, а также прав на использование и публикацию каждого элемента.
- Преобразование нецифровых материалов в цифровой формат.
- Загрузка элементов языковых данных и соответствующих примечаний в репозитории языковых данных.
- Проверка (внесение исправлений и комментариев, классификация и аутентификация данных).
- Публикация (обеспечение доступа к проверенным и одобренным материалам).

2.1. Определение источников языковых данных

Во-первых, языковые сообщества должны определить потенциальные надежные источники языковых данных. Они могут существовать в различных формах. Это могут быть как живые носители языка, так и исторические документы, архивы, произведения искусства и т. д. Полезно также включать примеры современной непринужденной беседы (устной или письменной), а не только формальное, историческое или литературное использование.

2.2. Сбор материалов и цифровое форматирование

На следующем этапе необходимо собрать фактические материалы. Если материалы еще не представлены в цифровой форме, они должны быть преобразованы в подходящую форму для цифрового хранения и передачи (например, текст, изображение, аудио, видео и т. д.). В частности, устные истории могут быть записаны в виде аудио- или видеофайлов или переведены в текст. Произведения искусства, книги, документы и даже рукописные списки слов могут быть отсканированы и внесены в архив в виде файлов изображений.

- Идентификация подходящих материалов.
- Сбор письменных материалов.
- Получение лингвистических знаний от носителей языка.
- Запись аудио или видео.
- Транскрибирование устных материалов.
- Сканирование или фотосъемка письменных материалов.

Члены сообщества могут начинать сбор различных типов языковых данных в любом порядке. Важность заключается в том, чтобы начать процесс сбора данных и внесения примечаний, а также инициировать хранение репозитория. К типовым материалам относятся:

- Рукописная корреспонденция и печатные документы.
- Тексты и книги.
- Одноязычные и переводные словари, а также грамматические системы языка.
- Веб-сайты.
- Социальные сети.
- Аудио- и видеозаписи.
- Песни, стихи и выступления.
- Устные традиции.
- Художественные работы, рисунки и фотографии.
- Знания сообщества и использование языка живыми носителями языка.

2.3. Лицензирование, владение и фондообразование

При сборе языковых данных необходимо убедиться, что вы также получаете законное право на распространение контента. Например, если в вашем распоряжении появляются личные письма, документы или запись чьей-либо речи, то перед тем как будет открыт доступ к данным материалам в репозитории вам может потребоваться разрешение автора или носителя языка (или других заинтересованных сторон). Возможно, вам потребуется получить письменное разрешение на публикацию данных, а также предоставление гарантий о том, что автор/докладчик имеет право предоставлять подобное разрешение. Может потребоваться консультация со старейшинами общины или юристами в пределах вашей юрисдикции, по вопросу возможных проблем с приемом данных, публикацией их в репозитории и использованием данных в качестве основы для оцифровки, а также для определения соответствующих формулировок разрешающих документов на использование данных.

Кроме того, может потребоваться пересмотр и адаптация традиционных подходов к обращению с интеллектуальной собственностью, авторским правом и суверенитетом данных с целью защиты прав сообщества на использование родного языка, знаний, наследия и культуры. Существует множество законных систем и способов решения проблем интеллектуальной собственности. Необходимо задействовать такие модели интеллектуальной собственности и суверенитета данных, которые учитывают мнение членов языкового сообщества. Для дальнейшего ознакомления с темой суверенитета данных, принадлежащих коренным народам, особенно в вопросах языка, см.:

- Battiste and Henderson, *"Protecting Indigenous Knowledge and Heritage"*
- Lovett et al., *"Good Practices for Indigenous Data Sovereignty"*
- Te Taka Keegan, *"Māori Sovereignty over Māori Language Data"*
- Christopher Hutton, *"Who Owns Language? Mother Tongues as Intellectual Property and the Conceptualization of Human Linguistic Diversity"*

В дополнение к юридическим аспектам сбора данных полезно документировать происхождение каждого элемента и его путь в репозиторий. Сбор подобной

информации о происхождении полезен для подтверждения принадлежности данных к конкретному языку, что будет учтено на этапе проверки данных.

2.4. Примечания

Полезно записывать информацию о каждом компоненте собранных данных, когда это возможно. Живой язык естественным образом меняется с течением времени. Языки также развивают региональные и диалектные различия. На речевые акты (устные или письменные) могут повлиять социальные факторы, такие как возраст каждого носителя языка, степень владения языком, гендерная принадлежность или статус, а также обстоятельства (церемониальные, формальные, неформальные и т. д.).

Запись информации о носителях языка, обстоятельствах, времени и месте совершения речевого акта и другие примечания создают более точную картину языка.

К примечаниям относятся:

- Дата создания и местонахождение единицы хранения,
- Информация о носителе языка или авторе, а также адресате или целевой аудитории (возраст, гендерная принадлежность, должность и степень владения языка [носитель языка / свободное владение языком]),
- Степень родства или тип отношений между говорящими,
- Обстоятельства,
- Литературный стиль (проза, поэзия, тексты песен, ритуальный текст и т. д.).

2.5. Хранение данных в репозитории

Файлы можно загрузить в репозиторий. В этом случае члены языкового сообщества, лингвисты, специалисты по вычислительной технике и другие третьи лица, которым сообщество предоставляет доступ, могут просматривать и проверять информацию о файле.

Алгоритм загрузки контента в репозиторий может зависеть от конкретной конфигурации и реализации используемого репозитория. Подробная информация по администрированию приведена в документации конкретного репозитория.

2.6. Контроль доступа для сообщества

Доступ к репозиторию контролируют члены языкового сообщества или их представители. Сообщество определяет администраторов репозитория. Администраторы управляют правами пользователей по загрузке, редактуре, просмотру или иным способам взаимодействия с содержимым репозитория.

Управление доступом в хранилище осуществляется преимущественно путем определения прав учетных записей или профилей пользователей. Для каждой учетной записи возможно подключение одной из возможных комбинаций прав и разрешенных действий с содержимым (так называемые «права доступа»). После того, как пользователи добавлены и у них есть возможность использовать репозиторий, им предоставляют права доступа определенного уровня. Полученные права доступа определяют степень взаимодействия с содержимым хранилища. К примеру, возможность просмотра, создания, редактирования или удаления записей. Для записи может быть предусмотрено множество полей, в том числе и поля для ввода метаданных о времени записи, человеке, сделавшем запись, местоположении и т. д. С этим связана необходимость определять права для просмотра или редактирования каждого из таких полей. Особенно это важно в тех случаях, когда некоторые данные представляют собой информацию, в отношении которой необходимо соблюдать конфиденциальность. Могут также существовать права администратора для скрытия или публикации записи, создания или редактирования категорий для организации записей, а также права для управления пользователями (добавление пользователя, удаление пользователя, изменение роли пользователя и т. д.). Особенности управления правами и разрешениями зависят от настроек репозитория. Обратите внимание, что разрешения могут быть привязаны к конкретной записи или подмножеству пользователей. Например, специалисту по индийским языкам могут быть предоставлены права на редактирование записей, относящихся к индийским языкам, но не записей на других

языках. Модератор группы, работающей над нигеро-конголезскими языками, может иметь права для управления пользователями, которые работают с этими языками, но не пользователями, работающими на других языках.

Некоторые разрешения можно использовать для создания процедуры утверждения или другого типа рабочего процесса в отношении содержимого хранилища. Например, когда материал впервые загружается в репозиторий, он может быть доступен для просмотра только группе рецензентов. Это дает группе возможность задавать вопросы и утверждать контент в качестве соответствующего и репрезентативного для данного языка. После одобрения рецензентами контент можно опубликовать или сделать пометку о его доступности для просмотра основной аудиторией. Также можно помечать контент, доступный для просмотра только совершеннолетними пользователями. Контент, нарушающий региональные законодательные документы, также может быть скрыт и помечен соответствующим образом. Наличие подобных функций определяется особенностями репозитория. Для получения более подробной информации обратитесь к документации для используемого репозитория.

Как правило, члены языкового сообщества предоставляют доступ к материалам следующим категориям лиц:

- Членам сообщества,
- Лингвистам,
- Специалистам по вычислительной технике,
- Иным заинтересованным лицам.

Контроль доступа может стать сложным и потребовать сложного набора прав пользователей и разрешений на доступ. В таблице 1 приведен обзор базовых прав и разрешений. Пустая ячейка таблицы указывает на права или привилегию доступа, в которых **отказано** лицам с соответствующими правами. Новые или отредактированные записи будут *скрыты* для общего доступа до тех пор, пока рецензент не установит для записи параметры *публикации*.

Таблица 1. Управление доступом к хранилищу

	Права доступа					
	Просмотр	**Комментарии**	**Создать/ редактировать**	**Скрыть/ опубликовать**	**Удалить**	**Категоризировать**
Учетная запись пользователя						
Гость	Предоставлено					
Утвержденный пользователь	Предоставлено	Предоставлено				
Участник сообщества	Предоставлено	Предоставлено	Предоставлено			
Исследователь или лингвист	Предоставлено	Предоставлено	Предоставлено			
Рецензент	Предоставлено	Предоставлено	Предоставлено	Предоставлено		Предоставлено
Администратор сообщества	Предоставлено	Предоставлено	Предоставлено	Предоставлено	Предоставлено	Предоставлено

- **Гость** – это анонимный пользователь, желающий посмотреть содержимое репозитория.
- **Утвержденный пользователь** – это физическое лицо, возможно предоставившее учетные данные, подтверждающие необходимый уровень компетентности для участия в проекте.
- **Участник сообщества** является носителем языка или утвержденным членом языкового сообщества.
- **Исследователь** или **лингвист** – это эксперт, которого приглашают для изучения материалов или участия в создании контента для репозитория.
- **Рецензент** – это человек, обладающий передовыми навыками модерирования дискуссий по контенту и способный понимать чувства и требования сообщества, а также разбирающийся в правовых и других вопросах.

- **Администратор сообщества** обладает полным контролем над учетными записями пользователей, привилегиями и данными хранилища.

2.7. Устранение ошибок

Даже при использовании тщательной процедуры утверждения в языковых данных неизбежно будут возникать ошибки. Типографские ошибки, ошибки в транскрипции, ошибочные воспоминания интервьюируемого и т. д. могут привести к ошибкам на этапе сбора данных. Проверять данные следует до того, как они станут доступными и будут задокументированы в качестве репрезентативных для искомого языка. Текущие периодические проверки позволяют исправлять ошибки в языковых данных, которые могут стать частью вашей коллекции.

2.8. Процесс проверки

Процесс проверки важен для обеспечения корректного описания и классификации данных, а также для подтверждения того, что используемые данные допустимы и заслуживают доверия. Подобные проверки могут выявить потенциальные расхождения в языковых данных.

Элементы данных могут быть неполными. Например, элементы могут быть просто фрагментами текста или их происхождение может быть неоднозначным. Такие элементы могут быть важны для проекта в целом. В процессе обзора можно принять решение об актуальности таких элементов.

Данные, внесенные в репозиторий, могут поступать из различных источников. Возможны ситуации, когда информация спекулятивного или надуманного характера может быть предоставлена без злого умысла или же, наоборот, преднамеренно.

Некоторые языки ранее не изучались и не документировались. Словарный состав, фонология, правила грамматики таких языков формально неизвестны. Собранные данные позволят установить структуру и терминологию языка. Обладание большой базой примеров языковых данных, представляющих широкое разнообразие речевых ситуаций, позволяет осуществлять цифровизацию языка с большей

точностью. Однако, если какие-либо из собранных данных не являются действительно репрезентативными для языка, то в этом случае осуществить оцифровку надлежащего качества становится сложнее. Ошибки или упущения могут привести к неверным выводам. Благодаря проверке можно снизить подобную вероятность.

Поэтому важно иметь возможность простого приема материалов в репозиторий с последующей проверкой носителями языка, членами сообщества, лингвистам или другими специалистами. Рецензенты могут оценивать, комментировать и даже оспаривать материалы на предмет подлинности, точности и интерпретации. Они также могут предоставлять дополнительный контекст, консультировать о правильном словоупотреблении и исправлять ошибки. Результатом проверки может быть внесение изменений или рекомендация поиска дополнительных или конкретных примеров данных. На этапе проверки может потребоваться предоставление рецензентом сведений о происхождении или информации о лицензировании. Проверка также может выявить наличие каких-либо нарушений законодательства, таких как запрещенные изображения или высказывания. Проверку можно повторять с определенной цикличностью.

До завершения проверки каждый элемент будет виден только лицам, выполняющим роль рецензента. Таким образом можно гарантировать, что другие пользователи репозитория увидят только утвержденную информацию. В случае, если у рецензентов возникают вопросы касательно какого-либо материала, они решают их с автором подобного материала. Затем элементу можно присвоить статус *публикации* и сделать его видимым для всех.

2.9. Разработка процесса по сбору данных языка

Первостепенной задачей является создание команды, обладающей необходимыми навыками и инструментами для каждой из задач по сбору языковых данных. Руководства, конкретные обязанности членов команды и спланированный рабочий процесс помогут осуществить сбор языковых данных. Обратите внимание на следующие вопросы:

- Сбор материалов
  - Есть ли у вас спланированный рабочий процесс или алгоритм приглашения участников, приема и оцифровки материалов?
  - Кто является потенциальным источником языковых данных?
  - Какие типы языковых материалов вам доступны?
  - Есть ли у вас способ сохранять и архивировать материальные объекты с языковыми данными?
  - Есть ли у вас инструменты и навыки для перевода материалов в цифровой формат?
  - Знакомы ли вы с правовыми аспектами на использование и публикацию языковых данных для вашего региона?
- Размещение языковых данных в цифровом формате в репозитории
  - Проведена ли установка и настройка репозитория языковых данных?
  - Обладаете ли вы навыками использования информационных технологий для управления репозиторием?
- Утверждение языковых данных
  - Создана ли команда квалифицированных специалистов по анализу языковых данных?
  - Созданы ли руководства по аннотированию и утверждению языковых данных?
  - Созданы ли руководства по допуску языковых элементов данных к публикации?
  - Есть ли у вас рекомендации о том, как обращаться к участникам за дополнительной информацией или как тактично опрашивать об уместности, подлинности, правах на публикацию, а также по иным вопросам о предоставленных материалах?
  - Есть ли у вас рекомендации по разрешению споров по вопросам, связанным с предоставленными материалами?
  - Есть ли у вас запланированный рабочий процесс или алгоритм для загрузки, просмотра, аннотирования и внесения исправлений в цифровые языковые данные, которые хранятся в репозитории?

➢ Управление командой

- Понимают ли члены вашей команды свои задачи и обязанности?

2.10. Краткое содержание раздела

Сбор языковых данных – это процесс, повторяющийся с определенной цикличностью. Элементы данных добавляются и анализируются по мере их поступления. В общих чертах процесс сбора языковых данных выглядит следующим образом:

- Привлечение заслуживающих доверия источников языковых данных.
- Получение языковых данных.
- Перевод материалов в цифровой формат.
- Загрузка материалов в репозиторий.
- Проверка прав на использование и публикацию.
- Аннотация.
- Проверка (внесение исправлений, комментариев и аутентификация данных).
- Публикация.

3. СПОСОБЫ ИСПОЛЬЗОВАНИЯ ЯЗЫКОВЫХ ДАННЫХ В КОМПЬЮТЕРНЫХ СИСТЕМАХ

Для поддержки цифровизации языка языковые данные используются разными способами. Их подробное описание не входит в задачи настоящего документа. Тем не менее приведем несколько основных способов использования языковых данных лингвистами и специалистами по вычислительной технике для поддержки цифровизации языка.

3.1. Образцы письма

Образцы письма первоначально используются для определения символов письма (букв, цифр, знаков ударения, знаков тона, знаков препинания и других символов), используемых в языке. По мере определения этого набора символов образцы письма можно использовать для определения того, как каждый символ нарисован от руки, и для создания шрифтов с этими символами. Набор символов также необходим для определения раскладки клавиатуры или метода ввода, используемого для ввода символов в цифровую систему.

Помните, что символы могут быть написаны несколькими способами. Также, символы могут быть стилизованы с помощью засечек и без них, выделены курсивом и иметь множество вариаций. В некоторых языках символы меняют форму в зависимости от их положения в слове или в зависимости от символа, находящегося рядом.

Кроме того, некоторые символы используются реже остальных. Они могут использоваться только для определенных церемоний или могут сохраняться лишь в старых версиях языка. Именно поэтому для большей надежности, полноты картины и приемлемости результатов цифровизации языка важно собрать наибольшее количество образцов.

3.2. Предложения и абзацы

Благодаря данным, содержащим полные предложения и абзацы, можно установить фонологические, орфографические, типографские, грамматические или другие лингвистические особенности языка, необходимые для обработки текста. К примеру:

- Правила грамматики.
- Расстановку переносов, деление слова, употребление заглавных букв, постановку ударения и правила пунктуации.
- Выравнивание и направление письма.
- Произношение.
- Формы обращения (например, почтительные обращения и порядок употребления фамилии и имени).

Данные также могут содержать уникальные правила написания и форматирования, используемые для представления дат, времени, эпох, чисел, процентов и т. д.

3.3. Списки терминов

Списки терминологии можно сформировать из образцов письма, аудиозаписей и других материалов. Такие списки слов и фраз используются средствами проверки орфографии, автозамены, распознавания текста, оптического распознавания символов (OCR) и другими цифровыми функциями.

3.4. Словари

Одноязычные и двуязычные словари содержат определения, указание части речи, произношение, этимологию, переводы и другую информацию. Эта информация может быть полезна для обработки текстов, расстановки переносов, проверки

орфографии и грамматики, автозамены, машинного перевода и других аспектов цифровизации.

Разработка и форматирование словаря могут быть осложнены. Лексикографы внимательно рассматривают, например, варианты формирования *заголовков* в полисинтетических языках, где сложные слова или предложения строятся из многих частей, а также к размещению слов в алфавитном порядке. Вот лишь некоторые из доступных ресурсов по составлению словарей для языков коренных народов:

- Nick Thieberger, "The lexicography of Indigenous languages in Australia and the Pacific",
- Antonia Cristinoi and François Nemo, "Challenges in endangered language lexicography",
- Paul V. Kroskrity, "Designing a Dictionary for an Endangered Language Community",
- Frawley, Hill, and Munro, *Making Dictionaries: Preserving Indigenous Languages of the Americas*,
- Sarah Ogilvie, "Linguistics, Lexicography, and the Revitalization of Endangered Language".

3.5. Аудиовизуальные материалы

Аудио- и видеозаписи могут быть использованы для определения норм произношения. Информация такого рода позволяет преобразовывать текст в речь и речь в текст. Функция преобразования текста в речь является актуальной как для людей с нарушениями зрения, так и для людей, испытывающих затруднения при чтении, а также для малограмотных представителей сообщества. Благодаря распознаванию речи возможно использование голосовых команд. Оно также полезно для людей с ограниченными возможностями, которые не могут печатать на клавиатуре или сенсорном экране.

3.6. Двуязычные данные

Двуязычные данные служат многим целям. Например, переводные словари, видео с субтитрами и переведенные документы позволяют создавать онлайн-словари для поиска слов, инструменты голосового перевода, машинного перевода и другие инструменты. Кроме того, различия в терминологии между языками позволяют выявить структурную специфику языков.

3.7. Объем

Как правило, чем больше собранных данных, тем выше качество цифровизации языка. Грамматика и определения слов становятся более точными и детализированными. Идиомы и редко используемые термины могут быть задокументированы.

Некоторая терминология используется только в связи с определенными темами или доменами. Это свойственно таким сферам как здравоохранение, сельское хозяйство, регулирование и т. д. Чем больше объем собранных данных, тем выше вероятность того, что будет охвачено больше доменов.

Кроме того, некоторые языковые приложения, например машинный перевод, эффективно работают только при наличии значительного объема языковых данных, доступных для обучения системы перевода.

3.8. Разнообразие

Для создания надежной поддержки цифровизации языка эффективно использовать различные виды языковых данных. Не исключайте источники языковых данных, которые кажутся архаичными, неформальными, формальными, официальными, преувеличивающими (реклама) или передающими неправдоподобную (легендарные или исторические истории) информацию, направлены на молодежь или необразованных людей. Целесообразно использовать следующие материалы:

- Детские книги с рассказами и картинками.

- Учебные материалы.
- Переписка, личные письма, заметки и сообщения.
- Нормативные документы (свидетельства о рождении, браке, смерти и т. д.).
- Словари (одноязычные и двуязычные).
- Книги.
- Газеты.
- Вывески.
- Плакаты.
- Архаичное письмо.
- Устные истории и традиции.
- Рисунки и другие произведения искусства.
- Повседневный разговор (устный, письменный, формальный и неформальный).

3.9. Создание новых терминов

Для обозначения новых идей, изобретений и видов деятельности закономерно создание новых терминов. Это особенно актуально в условиях подготовки языка к цифровизации. Например, терминология, используемая в пользовательском интерфейсе программного и аппаратного обеспечения, должна иметь определения или быть адаптирована на родном языке. Такие термины, как меню, кнопка, выпадающий список, файл, редактирование, справка, выход, ок, отмена, щелчок, загрузка и т. д.

Эквиваленты на родном языке не всегда легко или очевидно выбрать или изобрести. Буквальный перевод может быть не лучшим выбором. Например, термин *home page* («главная страница») в некоторых языках переводится как «начальная страница» (*start page*). Кроме того, термин может применяться в разных сферах, поэтому в зависимости от контекста его перевод будет варьироваться. К примеру, иногда термин *cancel* означает «прервать», а в других случаях это может означать «отменить». Также, в английском языке слово *orange* используется для обозначения и цвета, и фрукта, однако в других языках для каждого из них существуют отдельные термины.

Языковому сообществу, возможно, потребуется разработка методики по созданию и согласованию собственной терминологии для нужд цифровизации. Подробная информация о введении новых терминов приведена в документе ***«На пути в цифровое пространство. Руководство по терминологии»***.

3.10. Краткое содержание раздела

Достаточный объем и широкий диапазон типов языковых материалов коллекции позволяют осуществить цифровизацию языка надлежащего качества. В следующей таблице приведены некоторые основные способы использования языковых данных в компьютерных системах.

Таблица 2. Использование языковых данных в компьютерных системах

Тип языковых материалов	Лингвистическая ценность	Использование программным обеспечением
Образцы письма	Определение символов письменности и устоявшихся правил языка	Определение набора символов, шрифта, раскладки клавиатуры, метода ввода и т. д.
Данные о предложениях, абзацы	Выявление фонологических, орфографических, типографских, грамматических и других лингвистических правил	Обработка слов и речи
Списки терминов	Определение словарного состава языка	Проверка орфографии, автозамена, интуитивный набор текста и оптическое распознавание символов
Словари	Формулирование определений и указание частей речи	Проверка орфографии и грамматики, обработка текстовых данных, преобразование текста в

		речь и машинный перевод
Аудиовизуальные материалы	Выявление нормы произношения и образцов разговорной речи	Преобразование текста в речь и распознавание речи
Двуязычные данные	Получение вариантов перевода, межъязыковых сопоставлений и систем письма	Двуязычные словари, видео с субтитрами, переведенные документы и программное обеспечение, средства голосового перевода и машинный перевод
Созданные термины	Определение терминов, необходимых для работы цифровых систем	Программные меню, команды, диалоговые окна и т. д.

4. СПЕЦИАЛИЗИРОВАННЫЕ ЯЗЫКОВЫЕ УТИЛИТЫ

4.1. Стандартный репозиторий локальных данных Юникод

Стандартный репозиторий локальных данных Юникод (CLDR) предоставляет ключевые строительные блоки для процессов по интернационализации и локализации программного обеспечения, которые направлены на поддержку языков мира. Как следует из названия, CLDR представляет собой большой набор данных о локали. При адаптации собственного программного обеспечения многие компании используют информацию репозитория для изучения лингвистических особенностей разных языков.

Например, CLDR содержит принятые переводы многих имен собственных, необходимых для локализации программных продуктов (например, месяцы, дни недели, страны и их структурные единицы, названия языков, единицы измерения и валюты). CLDR также предоставляет выражения, которые можно использовать в исходном коде, чтобы с помощью программного обеспечения можно было автоматизировано форматировать данные в соответствии с местными особенностями презентации даты, времени, числа, системы мер, валюты и других.

Если ваш язык еще не представлен в CLDR, то включить его в данный репозиторий можно, предоставив необходимые данные. Это позволит разработчикам программных приложений беспрепятственно добавлять ваш язык в свои системы. Знакомство с материалами CLDR поможет определить информацию и терминологию, которые необходимы для работы цифровых систем и которые необходимо обеспечить для вашего языка.

4.2. Машинный перевод

Машинный перевод становится все более важным языковым приложением, поскольку он позволяет быстро и эффективно переводить большие объемы текста.

Однако для создания систем машинного перевода требуются большие объемы двуязычных данных. В идеале данные должны быть выровнены и составлены в

параллельные корпуса, где каждый сегмент (или предложение) исходного языка сопряжен с соответствующим сегментом целевого языка. Дополнительная информация о важности машинного перевода и его конкретных требованиях приведена в *приложении C «Требования к машинному переводу»*. В нем содержится обзор особенностей машинного перевода и требований к данным для создания системы машинного перевода.

4.3. Инструменты обработки естественного языка

В этом разделе вы узнаете об инструментах, которые помогают анализировать ваш язык и генерировать языковые данные. Примером может служить инструмент, который сканирует документ и извлекает список слов на вашем языке. Вначале вам следует определить инструменты, которые помогут в решении неотложных задач. Для работы с вашим языком может потребоваться дополнительная настройка подобных инструментов.

При обработке языковых данных многие программные приложения полагаются на компоненты обработки естественного языка (NLP). По завершении основных этапов по сбору материалов языка у вас может появиться желание инвестировать в обучение языковым моделям и обновление библиотек программного обеспечения, используемых в отрасли. Предоставление предварительно подготовленных моделей языка может ускорить внедрение вашего языка многими приложениями.

Несмотря на то, что языки разнообразны, многие из них можно сгруппировать по схожим шаблонам синтаксиса, грамматики и т. д. Инструменты, с помощью которых можно определить шаблоны языка, могут быть полезны при обработке языковых данных. Например, существуют инструменты, с помощью которых возможно анализировать текст для определенных типов языков и создавать списки слов. Существуют и более сложные инструменты, которые используют модели языка для обработки языковых данных. Языковые данные также можно использовать для обучения и создания все более совершенных моделей языка. Существуют некоторые инструменты, специально разработанные для работы с недостаточно документированными языками.

Существует несколько веб-сайтов с открытым исходным кодом, на которых размещены инструменты для обработки естественного языка и языкового моделирования. Выберите такой инструмент, который поддерживает языки, грамматически похожие на ваш, а затем настройте и обучите его с помощью языковых материалов на вашем языке. При достижении желаемого результата, затем можно добавить обученную модель вашего языка на эти веб-сайты. Разработчики программного обеспечения смогут беспрепятственно поддерживать ваш язык, и в дальнейшем у вас появится возможность предложить крупным компаниям-разработчикам программного обеспечения добавить ваш язык в их библиотеки.

К подобным инструментам относятся:

- Векторная визуализация слов, например, word2vec (модель для определения ассоциаций слов из большого массива текста);
- Обученные модели, такие как BERT и GPT (сбор данных грамматических функций, значений и ассоциаций слов);
- Модели для выделения именованных сущностей (NER) (имена, географические единицы, даты и т. д.);
- Модели для маркировки частей речи (существительное, глагол и т. д.);
- Модели для синтаксического анализа / анализа зависимостей (субъект / объект /..., глагольные и именные фразы и т. д.);
- «Трансформаторы» от Huggingface (извлечение информации, перевод и другие функции NLP);
- Библиотека SpaCy (извлечение информации и другие функции NLP);
- Библиотека для обработки естественного языка (NLTK) (библиотеки обработки текстовых данных).

Для настройки на используемый язык многие приложения полагаются на библиотеки идентификации языка. Такие приложения не могут поддерживать языки, которые не распознаются такими библиотеками. Предоставив информацию в такие

библиотеки, чтобы они распознавали ваш язык, вы можете ускорить процесс принятия и поддержки вашего языка в целом.

К примеру:

- CLD3 от Google (компактный языковой определитель v3).
- Библиотека fastText от Facebook.

4.4. Краткое содержание раздела

Эти утилиты помогут при анализе и поддержке процесса цифровизации вашего языка, позволят разработчикам программного обеспечения внедрить ваш язык и ускорить перевод материалов на ваш язык и с вашего языка.

- Стандартный репозиторий локальных данных Юникод.
- Машинный перевод.
- Инструменты обработки естественного языка.

5. ГЛОССАРИЙ

Термин	Описание
Символ	Буква, логограф, знак, помета или символ, используемые в письменной форме.
Корпус	Большое или полное собрание письменных материалов.
Диахрония	Рассмотрение каких-либо явлений, особенно языковых, в процессе развития.
Диалектный	Принадлежность к диалекту языка или отношение к нему.
Цифровизация	Преобразование материала в цифровую форму, которая может быть обработана компьютером.
Формат файла	Способ хранения информации в компьютерном файле. Способ зависит от типа хранимых данных и, как правило, может быть идентифицирован по расширению файла (например, .html для веб-страницы).
Шрифт	Графическое представление текста. Набор символов письменности с аналогичным графическим оформлением.
Глоссарий	Алфавитный список слов с определениями, относящихся к конкретному предмету.
Коренной народ	Обитатели на определенных землях.
Лексикография	Практика составления словарей.
Данные о локали	Информация, используемая для настройки пользовательских интерфейсов для данного языка и культуры региона.
Информационное пространство	1.Средства массовой коммуникации (вещание, издательская деятельность и Интернет), рассматриваемые в совокупности. 2.Устройства хранения данных.

МП	Машинный перевод.
NER	Выделение именованных сущностей.
NLP	Обработка естественного языка.
NLTK	Библиотека для обработки естественного языка.
Орфография	Набор правил передачи слов языка на письме. Включает нормы правописания, расстановки переносов, употребления заглавных букв, деления слова, постановки ударения и правила пунктуации.
Фонология	Раздел лингвистики, изучающий звуковой строй языка.
Полисинтетический	Язык, характеризующийся сложными словами, состоящими из нескольких морфем, в котором одно слово может функционировать как целое предложение.
Репозиторий	Центральное место, в котором хранятся и обрабатываются данные.
Сканирование	Копирование и хранение информации в цифровом формате.
Сегмент	Дискретная значимая единица текста или разговорного языка. Сегментация текста – это процесс разделения письменного текста на значимые единицы, такие как слова, предложения или темы. Сегментация речи – это процесс определения границ между словами, слогами или фонемами в естественных (разговорных) языках.
Засечка	Небольшой выступ, завершающий штрих буквы в определенных шрифтах.
Преобразование текста в речь	Вспомогательная технология, которая синтезирует человеческую речь из текстового материала.
Типографика	Искусство оформления шрифта, призванное придать разборчивость, читаемость и привлекательность письменности.
Unicode	Международный стандарт кодирования, включающий в себя символы разных языков и письменностей. Каждой букве, цифре, знаку или символу присваивается уникальное числовое

	значение, которое можно применять на разных платформах и программах.
Загрузка	Передача (данных) с одного компьютера на другой, обычно на тот, который больше, или удален от пользователя, или работает в качестве сервера.
URL	Универсальный указатель ресурсов, адрес страницы или другая информация в Интернете.
UTF-8	Кодировка символов на основе Юникода переменной ширины, используемая для электронной передачи текста.
Преобразование речи в текст	Программа распознавания речи, которая преобразует разговорную речь в письменный текст.
XLIFF	Формат обмена файлами локализации XML – формат файла на основе XML для обмена локализуемыми данными.

Приложение A. Примеры цифровых приложений

Указанные типовые виды деятельности могут быть доступны в цифровых системах на вашем родном языке. Также возможно создание новых приложений, соответствующих потребностям вашего сообщества.

Информационное взаимодействие

- Отправка и получение текстовых сообщений.
- Отправка и получение электронных писем.
- Отправка и получение мультимедийных материалов (изображения, аудио и видео).
- Автоматический перевод текста, машинный перевод.
- Автоматическое преобразование голосовых сообщений для отображения в виде текста и наоборот.

Публикация, документирование и обработка текстов

- Публикация информации и получение доступа к ней на веб-сайтах.
- Создание и распространение документов, книг, средств массовой информации, вывесок, плакатов и учебных материалов.
- Создание печатных и онлайн-словарей.
- Сканирование документов для преобразования в цифровой текст.
- Создание шрифта для письменности родного языка.
- Покупка и продажа товаров онлайн.
- Локализация веб-сайтов и приложений на родной язык.
- Создание приложений на родном языке.
- Проверка орфографии.
- Проверка грамматики и автозамена.

Пользовательские интерфейсы и поддержка пользователей с ограниченными возможностями

- Распознавание речи (актуальная функция для людей с ограниченными физическими возможностями).
- Использование голосовых команд для управления устройствами.
- Программы преобразования текста в речь и чтения с экрана (актуальная функция для людей с нарушениями зрения или малограмотных представителей сообщества).
- Преобразование речи в текст и создание субтитров в реальном времени (актуальная функция для людей с нарушениями слуха).

Приложение Б. Требования к данным для технологических приложений

Таблица «Требования к данным для технологических приложений» представляет собой наглядное руководство, которое поможет определить подходящую отправную точку для сбора языковых данных в соответствии с целями по созданию технологических приложений на родном языке.

Работать с таблицей можно по двум направлениям: искать необходимые компьютерные приложения и находить типы языковых данных, которые требуются для их создания. Или, основываясь на типах языковых данных, которые доступны или могут быть получены, сообщество может определить приложения, которые могут быть реализованы в ближайшей перспективе.

В основу данной таблицы легла информация настоящего документа. Таблица поможет пользователям ставить соответствующие цели в области цифровизации языка.

Таблица 3. Требования к данным для технологических приложений

Типы данных	Символы письменности	Терминология	Перевод	Речь
Примерное содержание	символы, буквы, цифры, знаки препинания, шрифты и др.	списки терминов, одноязычные словари	двуязычные словари, лексические базы данных, корпуса	фонология, правила произношения, аудио- и видеозаписи
Цифровые приложения				
Информационное взаимодействие				
Отправка/получение текстовых сообщений	x			
Отправка/получение электронных писем	x			
Автоматический перевод текста, машинный перевод	x	x	x	
Типы данных	Символы письменности	Терминология	Перевод	Речь

Автоматическое преобразование голосовых сообщений для отображения в виде текста и наоборот	x	x		x
Публикация, документирование, обработка текстов				
Публикация информации и получение доступа к ней на веб-сайтах	x			
Создание и распространение документов, книг, средств массовой информации, вывесок, плакатов и учебных материалов	x			
Создание печатных и онлайн-словарей	x	x	x	
Сканирование документов для преобразования в цифровой текст (OCR)	x			
Создание шрифта для письменности своего языка	x			
Покупка и продажа товаров онлайн	x			
Локализация веб-сайтов и приложений на родной язык	x	x	x	
Создание приложений на родном языке	x			
Проверка орфографии	x	x		
Проверка грамматики, автозамена	x	x		

Пользовательские интерфейсы и поддержка пользователей с ограниченными возможностями				
Типы данных	Символы письменности	Терминология	Перевод	Речь
Распознавание речи (актуальная функция для людей с ограниченными физическими возможностями, а также для взаимодействия с виртуальным помощником, например, Алисой)		x		x
Преобразование текста в речь и программы для чтения с экрана (актуальная функция для людей с нарушениями зрения или малограмотных представителей сообщества)	x	x		x
Преобразование речи в текст и создание субтитров в реальном времени (актуальная функция для людей с нарушениями слуха)	x	x		x
Легенда	**x**	Отмеченные данные с наибольшей вероятностью будут использоваться для создания приложений данного типа.		
Примечание		В таблице приведен далеко не полный список типов данных или приложений.		

Типы данных	Язык и правила разметки страницы		Не рассортированные языковые данные
Примерное содержание	орфография, грамматические правила, расстановка переносов, заглавные буквы, пунктуация, направление написания, выравнивание	правила написания: дат, времени, эпох, чисел, процентов и т. д.	предложения, абзацы, корпуса одноязычных текстов
Цифровые приложения			
Информационное взаимодействие			
Отправка/получение текстовых сообщений			
Отправка/получение электронных писем			
Автоматический перевод текста, машинный перевод	x	x	x
Типы данных	Язык и правила разметки страницы		Не рассортированные языковые данные
Автоматическое преобразование голосовых сообщений для отображения в виде текста и наоборот	x	x	x
Публикация, документирование, обработка текстов			
Публикация информации и получение доступа к ней на веб-сайтах	x	x	
Создание и распространение документов, книг, средств массовой информации, вывесок, плакатов и учебных материалов	x	x	
Создание печатных и онлайн-словарей	x	x	
Сканирование документов для преобразования в цифровой текст (OCR)			x
Создание шрифта для письменности своего языка	x		x

Покупка и продажа товаров онлайн	x	x	
Локализация веб-сайтов и приложений на родной язык	x	x	
Создание приложений на родном языке	x	x	
Проверка орфографии			
Проверка грамматики, автозамена	x	x	

Пользовательские интерфейсы и поддержка пользователей с ограниченными возможностями			
Типы данных	Язык и правила разметки страницы		Не рассортированные языковые данные
Распознавание речи (актуальная функция для людей с ограниченными физическими возможностями, а также для взаимодействия с виртуальным помощником, например, Алисой)			
Преобразование текста в речь и программы для чтения с экрана (актуальная функция для людей с нарушениями зрения или малограмотных представителей сообщества)	x	x	x
Преобразование речи в текст и создание субтитров в реальном времени (актуальная функция для людей с нарушениями слуха)	x	x	x
Легенда		Отмеченные данные с наибольшей вероятностью будут использоваться для создания приложений данного типа	
Примечание		В таблице приведен далеко не полный список типов данных или приложений.	

Приложение C. Требования к данным машинного перевода

Что такое машинный перевод?

Машинный перевод (МП) – это перевод текста или речи с одного языка на другой с использованием программного обеспечения.

Для достижения высокого качества переводов системам МП недостаточно механически заменять одно слово на другое. Для этого необходимо использовать передовые алгоритмы и большое количество языковых данных.

Качество переводов, выполненных системами МП, не сравнимо с качеством переводов, выполненных человеком. Для повышения качества системы МП часто настраивают в зависимости от сферы деятельности или отрасли, чтобы ограничить объем контента.

МП полезен как инструмент для оказания помощи переводчикам-людям, а в определенных ситуациях полученный результат можно использовать без изменений.

Создание систем машинного перевода

Чтобы автоматизировать перевод между двумя языками, система МП должна быть создана специально для этой пары. Для настройки системы необходимо подобрать соответствующую технологию МП и использовать языковые данные для обоих языков.

Создание системы МП для новых языковых комбинаций возможно только при условии обеспечения надежной цифровой основы для каждого из языков. Одним из условий также должно быть естественное и растущее распространение

лингвистических ресурсов на языке, только что прошедшем через процесс цифровизации.

Существуют языки, насчитывающие десятки миллионов носителей, но для которых, на настоящий момент, не существует подходящих систем МП. Это связано с отсутствием достаточного количества ресурсов с данными, или же проведен недостаточный объем работ. В то время как основная технология по разработке МП продолжает совершенствоваться, создавать новые системы с меньшим количеством исходных данных становится все проще. Несмотря на это, уже на начальном этапе важно понимать, что для создания *хороших* систем МП требуются значительные объемы данных.

Данные, используемые для построения систем МП, называются обучающими данными.

В то время как базовую систему можно разработать быстро, если в распоряжении имеются основные обучающие данные, для внесения улучшений и повышения производительности достаточно добавлять дополнительные данные в уже существующую систему МП.

Системы МП постоянно развиваются и совершенствуются по мере внесения новых данных, применения новых методов и благодаря существованию постоянной обратной связи, направленной на внесение поправок. Следует учитывать периодическую оценку и обновление системы.

В каких сферах целесообразно использовать МП?

МП применим для предоставления результатов обработки больших объемов информации и ресурсов знаний при относительно низких финансовых затратах и в сжатые сроки. Однако следует понимать, что в настоящее время результат перевода средствами МП не сравним с результатами перевода, выполненного квалифицированным переводчиком.

Тем не менее, МП имеет высокую скорость обработки текстов, зачастую обладает достаточно хорошим аппроксимирующим алгоритмом и может быть доступен в любое время для миллионов пользователей в Интернете после создания системы МП. Существование МП позволяет миллионам людей получать доступ к информации, которую они в противном случае не смогли бы получить. Хотя историй о неудачах МП и неправильных переводах предостаточно, для многих становится все более очевидным, что научиться успешно использовать и расширять возможности этой технологии становится крайне важно. Хотя МП вряд ли сможет заменить людей в любом приложении, где качество имеет первостепенное значение, растет число случаев, которые показывают, что МП подходит для:

- Контента с высокой долей повторов.
- Контента, который в противном случае просто не был бы переведен.
- Контента, для которого недоступен перевод, выполняемый людьми.
- Ценного контента, который меняется каждый час и каждый день.
- Знаний, которые облегчают и расширяют международное распространение важнейшей информации.
- Контента, созданного для улучшения и ускорения коммуникации с клиентами по всему миру, которые предпочитают модель самообслуживания.
- Контента, который не обязательно должен быть высокого качества, но на достаточном для понимания уровне качества.

Какие данные необходимы для разработки системы МП?

Все современные технологии по разработке МП основаны на данных, то есть компьютеры анализируют большие объемы накопленных переводческих данных, чтобы *научиться* переводить с одного языка на другой. Лингвистические данные, используемые для разработки систем МП, называются обучающими данными. В настоящее время наибольшее распространение имеет нейронный МП, который постепенно заменяет технологию прошлого поколения, называемую статистическим МП. В основе обеих технологий лежит принцип разработки систем МП с использованием следующих типов данных:

- Двуязычный текст.
- Глоссарии для перевода.
- Одноязычные данные на целевом языке.
- Одноязычные данные на исходном языке или близкородственных языках**.**

Двуязычные данные на исходном и целевом языках

Большие коллекции текстов, переведенных предложение за предложением, называются параллельными корпусами. Создать исходный механизм перевода возможно на основе минимум 100 000 двуязычных переведенных сегментов (предложений). Сегмент перевода может быть полным предложением или группой слов, которые дают перевод важным терминам и фразам. Предпочтительный объем составляет не менее 1 000 000 сегментов. Некоторые системы МП построены на основе миллиардов сегментов. Как правило, более высокое качество перевода МП можно обеспечить за счет больших объемов двуязычных сегментов высокого качества.

Многие языковые сообщества не располагают большими объемами таких данных. На этапе сбора данных часто необходима согласованная и долгосрочная работа, а также сотрудничество между государственными учреждениями, учебными заведениями и сообществом в целом. Тем временем существует технология, позволяющая использовать меньшее количество сегментов, а обратная связь с людьми обеспечивает постепенные улучшения по мере перевода данных.

Корпуса, используемые в качестве обучающих наборов данных для алгоритмов МП, обычно извлекаются из больших массивов аналогичных источников, таких как базы данных новостных статей, написанных на исходном и целевом языках, описывающих аналогичные события.

Однако извлеченные фрагменты могут содержать искажения информации, дополнительные элементы, привнесенные в каждый корпус. Существующие методы извлечения данных позволяют различать двуязычные элементы, представленные в обоих корпусах, и одноязычные элементы, представленные только в одном корпусе, для извлечения более точных параллельных фрагментов двуязычных элементов.

Сопоставимые корпуса используются для непосредственного получения информации, необходимой для переводческой деятельности. Однако высококачественные параллельные данные трудно получить, особенно для языков с ограниченными ресурсами.

В качестве обучающих данных, используемых для создания системы МП, чаще всего выступают данные памяти переводов (архив прошлых переводов) или другие ресурсы, полученные в ходе перевода более ранних проектов. Такая информация определит сферу, в которой система МП научится переводить лучше всего. Зачастую объем доступных данных ограничен. В таких случаях необходимо приложить особые усилия, чтобы научить систему усваивать материал, на переводе которого она, скорее всего, будет сосредоточена. Помните, что специфика тренировочных данных определит сферу, в которой система будет переводить лучше всего. Таким образом, систему МП, которая будет использоваться для перевода медицинского контента, лучше всего обучать с использованием памяти переводов и глоссариев по медицинской тематике.

Необходимым условием использования двуязычных данных для обучения системы является их **выравнивание** перед использованием: исходный текст и перевод должны быть прямыми переводами друг друга. Не подходят для использования переведенные тексты, которые являются сводными обзорами, аннотациями или комментариями к оригинальному тексту. Предварительный анализ данных необходимо проводить, чтобы определить их применимость для обучения системы.

Большей точности перевода можно добиться благодаря использованию глоссариев и словарей с ключевыми терминами.

Разработка комплексной стратегии метаданных для собираемых лингвистических данных является ценным ресурсом, который можно использовать и в дальнейшем. На начальных этапах, чтобы получить минимально необходимое количество данных, основное внимание обычно уделяется поиску данных в любых возможных источниках. Вместе с тем, по мере развития движков МП, при условии использования подходящих данных для их создания, возможно достижение высоких показателей производительности. Таким образом, вместо неспециализированной

единой системы с помощью оптимизации можно создать систему МП для работы с контентом, связанным с медицинской тематикой, или контентом, связанным с компьютерными технологиями. Такая специализация часто приводит к созданию более эффективных систем МП.

Форматы двуязычных данных

Двуязычные данные обладают тремя важными характеристиками, которые могут быть полезны при использовании в качестве обучающих данных для машинного перевода. Они должны быть в доступном для импорта в систему МП формате. Желательная кодировка символов текстовых данных – это Unicode UTF-8. Параллельные данные должны быть выровнены таким образом, чтобы каждый сегмент исходного языка был сопряжен с соответствующим сегментом целевого языка.

Данные могут быть доставлены в следующих форматах файлов; список приведен в порядке убывания приоритетности:

- Память переводов (TMX, TBX, XLIFF, CSV): предпочтительный формат.
- Обычный текст (TXT).
- Веб-контент (HTML).
- Структурированный (XML).
- Microsoft Office (DOC, DOCX, PPT, PPTX, XLS, XLSX).
- Форматы публикации или DTP (TTX, PDF, FrameMaker).
- Оптическое распознавание символов (OCR), (TIFF, PNG, JPEG и т.д.).

Выравнивание и сопоставление данных между исходным и целевым языками уже может быть выполнено или данные могут быть доставлены в необработанных формах, таких как, например, документы Microsoft Word или HTML. Если данные не выровнены, необходимо будет использовать инструменты, которые позволят выровнять данные с высоким уровнем точности.

Языковые данные на одном языке

Несмотря на то, что переводы, имеющиеся в архиве, и двуязычный текст, возможно, являются наиболее важными данными для создания систем МП, достаточно важно также иметь качественные одноязычные данные на целевом языке. На этапе перевода эти данные используются для изучения правильной грамматической структуры и статистически влияют на результат стилистического оформления получаемого текста. Это особенно актуально при построении статистических систем МП.

Гораздо легче получить одноязычные данные, чем данные на двух языках. Достаточно эффективным может быть подход по сбору URL-адресов веб-сайтов с похожим доменом или грамматическим стилем. Языковые данные этих ресурсов могут быть извлечены, и лингвистические знания, содержащиеся в этих данных, можно использовать.

Как правило, собрать языковые данные для языков коренных народов достаточно сложно. Для большинства языков существует множество источников одноязычных языковых данных. Однако примером ситуации, когда важны одноязычные языковые данные, является создание системы МП для работы с текстами медицинской тематики. Сканирование веб-сайтов медицинской и смежной тематики, богатых подобным специфическим контентом, может выявить языковую информацию, важную для создания глоссариев и памяти переводов.

Желательная кодировка символов текстовых данных – это Unicode UTF-8. Данные могут быть доставлены в следующих форматах; список приведен в порядке убывания приоритетности:

- Обычный текст (TXT).
- Память переводов (TMX, TBX, XLIFF, CSV).
- URL-адреса веб-контента.
- Веб-контент (HTML).
- Структурированный (XML).
- Microsoft Office (DOC, DOCX, PPT, PPTX, XLS, XLSX).

- Форматы публикации или DTP (TTX, PDF, FrameMaker).

Приложение Г. Преимущества цифровизации языка

Мы надеемся, что в серии руководств «На пути в цифровое пространство» приведена вся необходимая информация, которая может понадобиться заинтересованным членам языковых сообществ для цифровизации их языка. Цифровизация языка позволяет использовать компьютерные технологии на родном языке.

Преимущества перевода языка в цифровую форму будут зависеть от целей сообщества его носителей. К таким целям могут относиться почитание красоты языка, поддержание систем знаний, распространение ценностей, создание приложений и продуктов, обмен преданиями и историческими знаниями, содействие охране окружающей среды и новаторским идеям, а также расширение торговых связей, распространение образования, обеспечение трудовой занятости, обогащение сферы развлечений, а также поддержка здравоохранения и безопасности. Цифровизация позволяет членам сообщества пользоваться постоянно расширяющимся набором компьютерных инструментов для поддержания языка, его возрождения и просветительской деятельности. Достаточно выраженное цифровое присутствие в Интернете и, следовательно, повышенная видимость языкового сообщества способны повлиять на государственную политику в сфере поддержки общин коренных народов и подтолкнуть бизнес к интеграции. Учитывая повсеместное распространение смартфонов среди молодежи, естественным способом привлечения этой части общества к изучению родного языка может стать цифровизация. Более широкое распространение языка с помощью цифровых платформ может расширить возможности носителей языка, а также повлиять на их востребованность в разных сферах.

Присутствие языковых сообществ на мировой арене посредством цифровых платформ оказывает благоприятное воздействие на мир в целом. Их опыт, знания и уникальное мировоззрение внесут значительный вклад в развитие остального

мира. В результате подобного сотрудничества возможно появление новых решений общемировых проблем. Благодаря цифровизации обеспечивается сохранность подобной информации и возможность ее издания. Таким образом, многообразие и содержание естественных языков становятся наследием всего мира, и сохраненные языки станут достоянием человечества, раскрывая свою ценность с неизвестной в настоящее время стороны.

Подводя итог, можно сказать, что основные преимущества цифровизации языка заключаются в следующем:

- У носителей одного языка появляется возможность легко получать доступ к материалам на родном языке, создавать такие материалы и обмениваться ими, в том числе на больших расстояниях, с отдельным лицам или большим группам.
- Расширяется доступность медицинской информации и информации, связанной с предоставлением медицинских услуг.
- Обеспечивается поддержка экстренной связи в чрезвычайных ситуациях и при стихийных бедствиях.
- Расширяется как местная, так и электронная торговля.
- Создаются новые возможности для обмена результатами творческой деятельности, инновационными и философскими идеями.
- Разрабатываются учебные материалы на родном языке.
- Улучшаются отношения и взаимодействие с соседями.
- Улучшаются механизмы урегулирования споров.
- Обеспечивается защита интересов, а также доступ к правовым и правительственным процедурам на родном языке.

- Расширяется доступ к информации в сети Интернет для образовательных и коммерческих целей, а также для привлечения общественности, как на родном, так и на других языках по мере развития средств перевода.
- Появляется возможность познакомить представителей других культур с языками, культурой и мудростью коренных народов.
- Расширяется роль и значимость общин коренных народов во всем мире.
- Появляется возможность поддержки и оживления родного языка для маргинализированных групп и меньшинств, даже если язык поглощен доминирующими группами.
- Сохраняются исконные системы знаний, культура, история, искусство, медицина, мудрость, ценности и мировоззрение.