Желтов Валериан Павлович
канд. техн. наук, профессор
ФГБОУ ВО «Чувашский государственный университет
им. И.Н. Ульянова»
Желтов Павел Валерианович
канд. техн. наук, ведущий инженер
АУ Чувашской Республики «Научно-исследовательский институт
экологии и природопользования»
Министерства природных ресурсов и экологии
Чувашской Республики
г. Чебоксары, Чувашская Республика

СОЗДАНИЕ ОДНОЯЗЫЧНЫХ СЛОВАРЕЙ ДЛЯ СИСТЕМЫ МАШИННОГО ПЕРЕВОДА ЧУВАШСКОГО ЯЗЫКА

Аннотация*: в статье рассмотрена система Apertium, а также её функции, способ работы и создание одноязычных словарей для системы машинного перевода чувашского языка.*

Ключевые слова*: система Apertium, чувашский язык, одноязычные словари.*

Введение

В современном мире перевод текстов стал обыденностью. Любой человек, даже не знающий какой-либо язык, кроме родного, может открыть Google Translate или Яндекс.Переводчик и без каких-либо проблем перевести необходимый ему текст. Но даже нынешнюю эпоху – эпоху глобализации есть один нюанс. Это миноритарные языки. Основной проблемой перевода для этих языков является то, что для большинства из них нет систем машинного перевода. А те языки, которым повезло иметь переводчик, используют язык-посредник (чаще всего английский). Другими словами, текст переводится не напрямую – сначала происходит перевод оригинального языка на язык посредник, и только затем на необходимый язык, что заметно сказывается на качестве перевода [1].

В настоящее время чувашский язык добавлен в список языков Яндекса, однако статистический перевод, используемый Яндексом и основанный на корпусах текстов, для языков с небольшими корпусами, к каковым относится и чувашский желает лучшего [2].

Большой проблемой является обучение чувашскому языку в городских школах, где для того, чтобы он был конкурентоспособен с такими востребованными обществом языками как русский и английский, необходимо придавать обучению какие-то новые свойства. Одним из таких является модернизация обучения чувашскому языку в школе за счет привлечения цифровых технологий. На наш взгляд, было бы актуальным на уроках информатики и чувашского языка вовлекать учащихся в разработку компьютерных переводчиков с чувашского на русский и английский. Это позволит сделать обучение чувашскому языку более интересным и востребованным.

OpenTrad Apertium (Apertium для краткости) – это платформа с открытым исходным кодом, которая включает в себя инструменты и программы, необходимые для создания и запуска систем машинного перевода на основе правил, хотя некоторые из их модулей являются статистическими или гибридными; переводчики, построенные с помощью Apertium, могут переводить десятки тысяч слов в секунду [3]. Первоначально Apertium был разработан как поверхностная система передачи, ориентированная на родственные языки (например, каталонский-испанский, португальский-испанский, чешский-словацкий и т. д.), Но с момента выпуска второй версии, опубликованной в декабре 2006 года, механизм перевода использует более продвинутую систему передачи, которая позволяет обрабатывать лингвистические черты, присутствующие на неродных языках (например, испанский-английский).

Apertium является результатом различных проектов государственных грантов, в которых до сих пор участвовали различные университеты (Университет Аликанте, Университет Виго, Политехнический университет Каталонии, Университет страны Басков и Университет Помпеу Фабра) и такие компании как

Eleka Ingeniaritza Linguistikoa и т. п. Платформа Apertium основана на опыте и знаниях, приобретенных группой Transducens Universitat d'Alacant в развитии каталонско-испанской системы interNOSTRUM и португальско-испанского переводчика Tradutor Universia, которые, однако, имеют закрытый исходный код.

Apertium в основе своей написан на C++ и может быть скомпилирован и запущен на операционной системе Linux, хотя его адаптация к другим операционным системам в теории не должна вызывать много проблем. В любом случае программа может быть легко установлена на сервере интернет-приложений для удаленного доступа. На сегодняшний день разработаны функциональные языковые данные для следующих языковых пар (в обоих направлениях перевода): каталонский-испанский, галисийский-испанский, португальский-испанский, аранский-каталонский, каталонский-французский и каталонский-английский. Для некоторых из упомянутых языковых пар частота ошибок составляет от 5% до 10% с журналистскими текстами, но эти результаты можно легко улучшить, увеличив набор словарей или правил трансфера.

В настоящее время на основе платформы Apertium также создано несколько подобных переводчиков для тюркских языков (для башкирского, казахского, каракалпакского, татарского, турецкого, узбекского и чувашского). Перевод с этих языков доступен на сайте https://turkic.apertium.org.

В публикуемой серии статей нашей задачей было описать инструкции по работе с ней простым и доступным языком для учителей и учащихся общеобразовательных школ. Русско-чувашский и чувашско-русский переводчики на базе Apertium начали создаваться в Чувашской Республики уроженцем Каталонии, проживающим в Чебоксарах, Эктором Алос-и-Фонтом.

Следует отметить, что требуется значительная доработка чувашского словаря (его расширение). Так, если ввести на сайте https://turkic.apertium.org чувашское слово «япала» «вещь» и перевести его на татарский язык, то получим татарское слово «милек» «веник».

Механизм перевода Apertium, вспомогательные инструменты, соответствующая документация и большинство лингвистических данных, разработанных на сегодняшний день для Apertium, могут быть загружены с веб-сайта проекта в *https://www.apertium.org*, а также с сайта https://turkic.apertium.org.

Система Apertium – это платформа машинного перевода. Фактически, она обеспечивает пользователя необходимыми инструментами, с помощью которых можно строить свои собственные системы машинного перевода. Необходимо лишь определить исходные данные, которые и обеспечат работоспособность системы. Если упростить всю схему, то нам необходимо создать три словаря и определенный набор правил, которые обеспечат грамматически корректные трансформации, логические перестановки слов, и т. д.

Apertium не работает в Windows, поэтому необходимо установить систему Linux. Это в принципе является существенным недостатком, препятствующим ее использование учителями миноритарных языков в школах. Поэтому она должна запускаться на предварительно установленной виртуальной машине, например Oracle VM VirtualBox (Oracle Virtual Machine VirtualBox, виртуальной машине базы данных). Загрузить ее на компьютер можно с официального сайта компании Oracle,по адресу *https://www.oracle.com/ru/virtualization/virtualbox/*.

Для начала работы нам понадобятся сама платформа Apertium и lttoolbox – набор инструментов для лексической обработки, морфологического анализа и генерации слов (рис.1). Они находятся в папке apertium-cv.

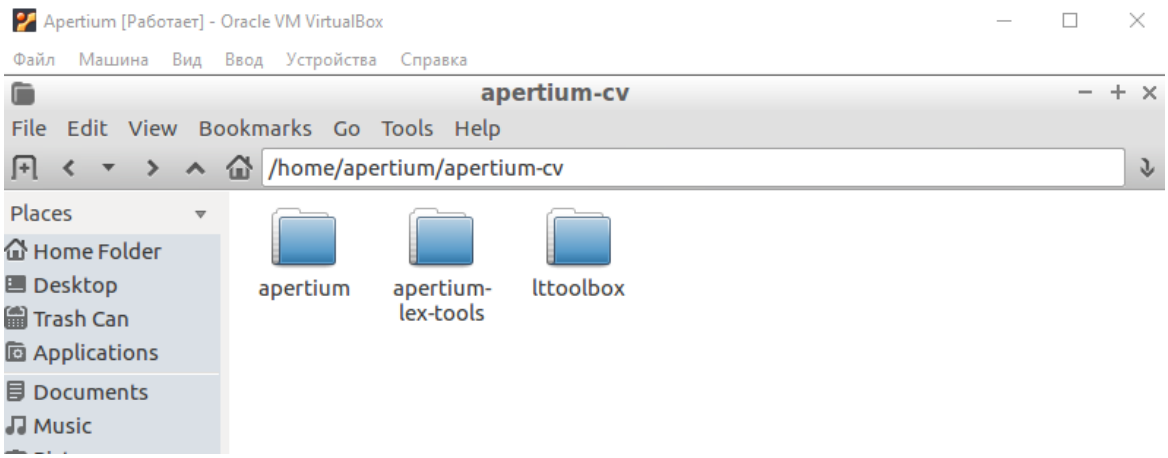

Рис. 1. Папка с готовым инструментарием

Apertium – это система машинного перевода поверхностно-трансферного типа. Это значит, что он имеет дело со словарями и правилами поверхностного трансфера, т.е. формальной передачей грамматических правил. По существу, поверхностный трансфер отличается от глубокого тем, что в нем не выполняется полный синтаксический разбор предложений, а правила, в отличие от операций на дереве синтаксического разбора, представляют из себя операции с некоторыми группами лексических единиц. Таких словарей три.

Морфологический словарь для первого языка: он содержит правила о том, как видоизменяются слова в этом языке. В нашем случае мы назовем его: apertium-cv-ru.cv.dix. Здесь аббревиатура «cv» означает Chuvash «чувашский», «ru» означает Russian «русский».

Морфологический словарь для второго языка: в нем содержится та же информация что и в первом словаре, только уже для данного языка. Называться он будет так: apertium-cv-ru.ru.dix.

Двуязычный словарь. Он содержит в себе соответствия слов и символов в обоих языках. У нас он будет называться apertium-cv-ru.cv-ru.dix.

В этой паре любой язык может быть как исходным, так и целевым.

Остается лишь добавить файл с правилами трансфера. Это такие правила, которые определяют расположение слов в предложениях, согласуют род (для русского языка), число, а также могут использоваться для удаления и вставки лексических единиц, например: вышел на улицу – тухрăм урама – урама тухрăм. Его названием будет apertium-cv-ru.cv-ru.t1x (рис. 2).

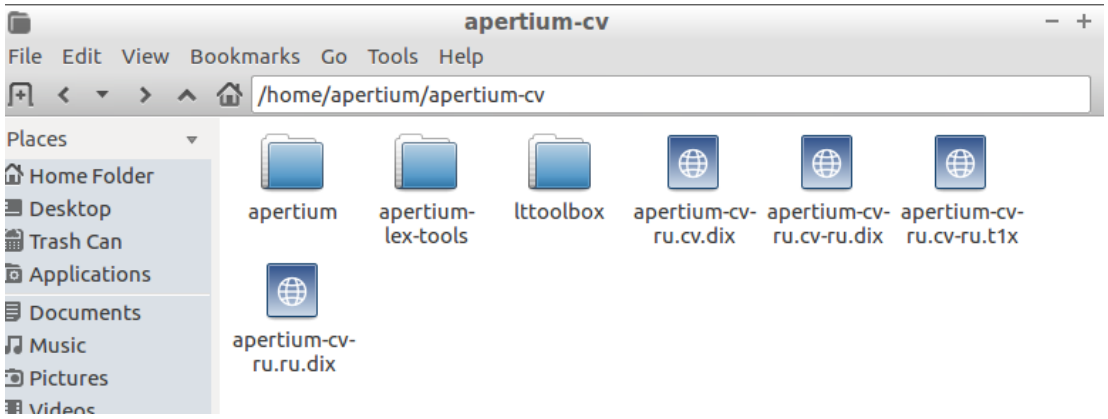

Рис. 2. Весь инструментарий и словари готовы к работе

1. Создание одноязычных словарей.

Создадим первый словарь. По сути своей словари являются XML-файлами. Поэтому для создания основы словаря откроем файл apertium-cv-ru.cv.dix, и впишем в него следующее:

<?xml version=«1.0» encoding=«UTF-8»?>

<dictionary>

</dictionary>

Теперь файл определяет, что начинаем создание словаря. Чтобы данный файл был более полезным, необходимо добавить в него еще несколько записей и первый из них – алфавит. Он определит набор букв, которые могут быть использованы в чувашском словаре. Он выглядит так:

<alphabet>АĂБВГДЕЁЁЖЗИЙКЛМНОПРСҪТУЎФХЦЧШЩЪЫЬЭЮЯаăбвгдеёёжзийклмнопрсҫтуўфхцчшщъыьэюя</alphabet>

Далее необходимо определить некоторые символы. Начнем с более простых – существительного (n) в единственном (sg) и множественном (pl) числах.

<sdefs>

<sdef n=«n»/>

<sdef n=«sg»/>

<sdef n=«pl»/>

</sdefs>

Названия символов не обязаны быть такими короткими, их можно писать полностью, но так как делать это придётся много раз, есть смысл их сокращать.

Следующим шагом определим раздел для парадигм,

<pardefs>

</pardefs>

и раздел для словаря:

<section id=«main» type=«standard»>

</section>

Теперь появляется скелет словаря, и переходим к добавлению существительного. Им будет, например, слово «кушак» «кошка’.

Так как ранее определённых парадигм у нас нет, первым делом нужно определить парадигму.

Напомним, что имеется ввиду парадигмы для леммы, т.е. для основы в именительного падеже, когда речь идет о существительных (этот падеж в чувашской грамматической традиции принято именовать основным). Формой единственного числа является «кушак», формой множественного числа – «кушаксем». Таким образом:

<pardef n=«кушак__n»>

<e> <l/><r><s n=«n»/><s n=«sg»/></r> </e>

<e> <l>сем</l><r><s n=«n»/><s n=«pl»/></r> </e>

</pardef>

Следует пояснить обозначения символов: <e> означает запись (entry), соответствует словарной статье, означает пару (pair), <l> означает влево (left), r означает вправо (right)

Таким образом, определяется парадигма. Теперь нужно связать её с леммой – чувашским словом «кушак». Это действие выполняется в разделе «section».

<e lm=«кушак»>кушак<par n=«кушак__n»/></e>

Кратко поясним сокращения: <lm> означает лемму (lemma), <i> означает идентичность (identity), т.е. одно и тоже как при синтезе, так и при анлизе, <par> означает парадигму (paradigm).

Эта запись определяет лемму слова, «кушак», основу слова, «кушак», и парадигму склонения этого слова «кушак_n». Разница между леммой и основой в том, что лемма есть «цитируемая форма» слова, в то время как основа есть часть леммы, к которой присоединяются аффиксы. Теперь можно проверить словарь. Для этого необходимо скомпилировать файл. Делается это в командной строке при помощи команды:

lt-comp lr apertium-sh-en.sh.dix sh-en.automorf.bin

После компиляции у нас появился анализатор. Кроме анализатора нам необходим еще и генератор. Он создается при помощи команды:

lt-comp rl apertium-sh-en.sh.dix sh-en.autogen.bin

Таким образом, создается чувашский словарь. Протестировать его можно введя в командной строке какое-либо слова из словаря с парадигмой (т.е. форме отличной от леммы), например, «кушаксем» «кошки».

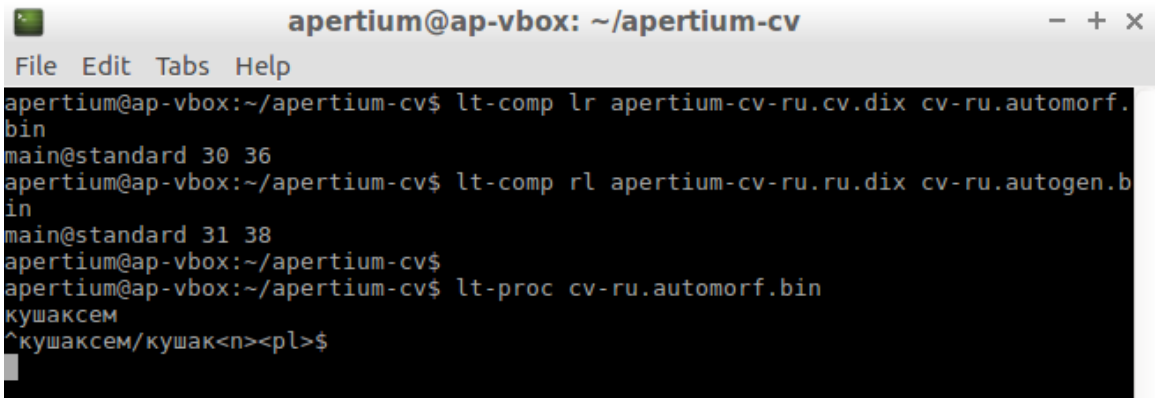

Рис. 3. Результат компиляции и тестирования словаря

Как видно на рис. 3, после анализа слова «кушаксем» получили лемму «кушак», а также информацию о том, что это существительное во множественном числе.

Таким же образом необходимо заполнить и скомпилировать словарь для русского языка или воспользоваться готовым словарем.

Выводы

Рассмотрена система Apertium, а также её функции и способ работы по созданию одноязычных словарей на примере создания словарей для чувашского языка. Системы с открытым исходным кодом имеют огромное преимущество по сравнению с системами с закрытым исходным кодом. Создание одноязычных словарей для системы машинного перевода в школах представляется одним из способов повышения привлекательности информатики и чувашского языка как предметов.

Список литературы

1. PC-Kimmo [Электронный ресурс]. – Режим доступа: https://software.sil.org/pc-kimmo

2. UNESCO Atlas of the World's Languages in Danger [Электронный ресурс]. – Режим доступа: http://www.unesco.org/languages-atlas/en/atlasmap/language-id-338.html

3. Apertium Documentation [Электронный ресурс]. – Режим доступа: http://wiki.apertium.org/wiki/Documentation