УДК 004.42

РАЗРАБОТКА СИСТЕМЫ МАШИННОГО ПЕРЕВОДА С ЧУВАШСКОГО НА РУССКИЙ ЯЗЫК

[1]Желтов В.П., [2]Желтов П.В.

[1]ФГБОУ ВПО «ЧГУ им. И.Н. Ульянова», Чебоксары, e-mail: chnk@mail.ru;
[2]АУ «"НИИ экологии" Минприроды Чувашии», Чебоксары, e-mail: tchouvachie@yandex.ru

Основной проблемой перевода малых языков является то, что для большинства из них нет систем машинного перевода. Согласно классификации ЮНЕСКО, чувашский язык входит в список вымирающих. В данной ситуации весьма актуальна задача разработки системы машинного перевода для чувашского языка. Объектом исследования является машинный перевод. Предметом исследования служат создание русско-чувашской языковой пары, а также семантический, синтаксический и морфологический синтез. Рассмотрены этапы разработки чувашско-русской языковой пары, а следовательно, и системы перевода. Исследованы актуальные данные, предоставляемые платформой Apertium, технологии разработки на платформе Apertium. Рассмотрены языковые словари, а также словари правил. Приведены методы их заполнения и технология по созданию, модификации и совершенствованию языковых пар: создание одноязычных словарей; создание двуязычных словарей; создание трансферных правил; решение проблем со множественными словами; склонения по падежу и роду. Рассмотренная в работе языковая пара содержит актуальные данные, которые были получены путем их поиска в открытых источниках, в том числе в сети «ранжировать Интернет ранжировать». Разработана система машинного перевода, которая может стать базисом для создания более продвинутых переводчиков на национальный язык.

Ключевые слова: система Apertium, чувашский язык, русский язык, машинный перевод

DEVELOPMENT OF A MACHINE TRANSLATION SYSTEM FROM CHUVASH TO RUSSIAN LANGUAGE

[1]Zheltov V.P., [2]Zheltov P.V.

[1]FGBOU VPO «ChGU of I.N. Ulyanov», Cheboksary, e-mail: chnk@mail.ru;
*[2]Autonomous Institution «Research Institute of Ecology» of the Ministry of Natural Resources of Chuvashia, Cheboksary, e-mail: tchouvachie@yandex.ru_*

The main problem with translating small languages is that there are no machine translation systems for most of them. According to UNESCO, the Chuvash language is on the endangered list. In this situation, the task of developing a machine translation system for the Chuvash language is very relevant. The object of the research is machine translation. The subject of the research is the creation of a Russian-Chuvash language pair, as well as semantic, syntactic and morphological synthesis. The stages of development of the Chuvash-Russian language pair, and, consequently, the translation system are considered. The actual data provided by the Apertium platform, development technologies on the Apertium platform have been investigated. Language dictionaries and rules dictionaries are considered. Methods for their filling and technology for the creation, modification and improvement of language pairs are given: creation of monolingual dictionaries; creation of bilingual dictionaries; creation of transfer rules; solving problems with multiple words; declension by case and gender. The language pair developed in the work contains actual data that was obtained by searching for them in open sources, including in the «rank the Internet to rank» network. A machine translation system has been developed, which can become the basis for creating more advanced translators into the national language.

Keywords: Apertium system, Chuvash language, Russian language, machine translation

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 [1].

Известны системы машинного перевода: Apertium, PC-KIMMO и Trados и др. [2].

Разработка на платформе Apertium системы чувашско-русского машинного перевода является актуальной задачей. В чувашском языке важную роль при морфологическом анализе выполняет не словарь основ, а словарь морфем. Сочетаемость аффиксов друг с другом, шаблоны этих пар основаны на схемах следования [3]. Общее число аффиксов – 170–200.

Материалы и методы исследования

Описание языка в PC-KIMMO состоит из двух файлов, которые предоставляет пользователь (рис. 1).

Программа распространяется бесплатно, написана на языке программирования C++ и имеет открытый исходный код. Недостатком программы является достаточно сложная для неподготовленного пользователя система записи фонологических и морфотактических правил [4].

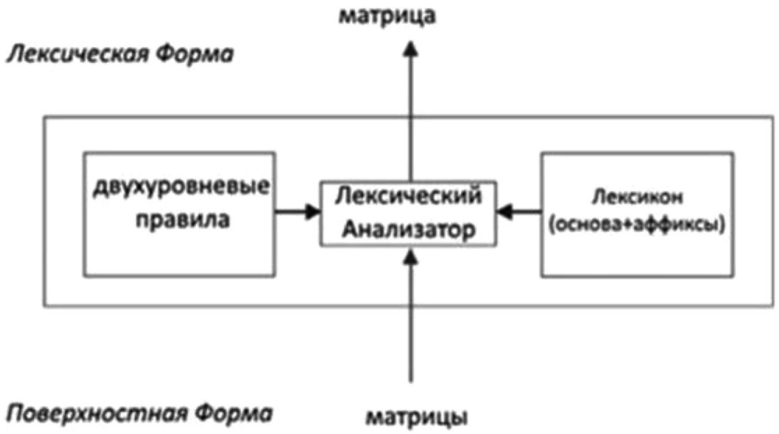

Рис. 1. Структурная схема PC-KIMMO

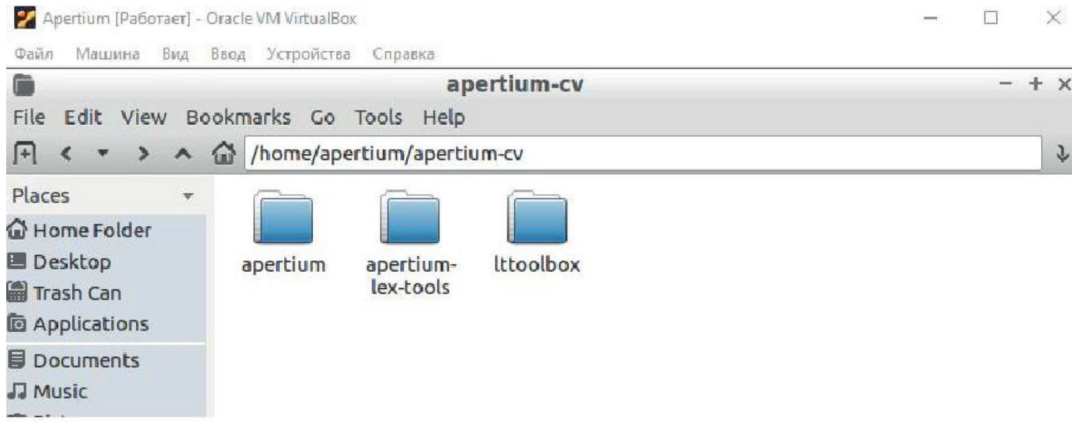

Рис. 2. Папка с готовым инструментарием

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.

Apertium не работает в Windows, поэтому необходимо установить систему Linux. Это в принципе является существенным недостатком, препятствующим ее использованию 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 в 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. 2). Они находятся в папке apertium-cv.

Apertium – это система машинного перевода поверхностно-трансферного типа. Это значит, что он имеет дело с формальной передачей грамматических правил. По существу, поверхностный трансфер представляет собой операции с некоторыми группами лексических единиц. Таких словарей три [5].

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

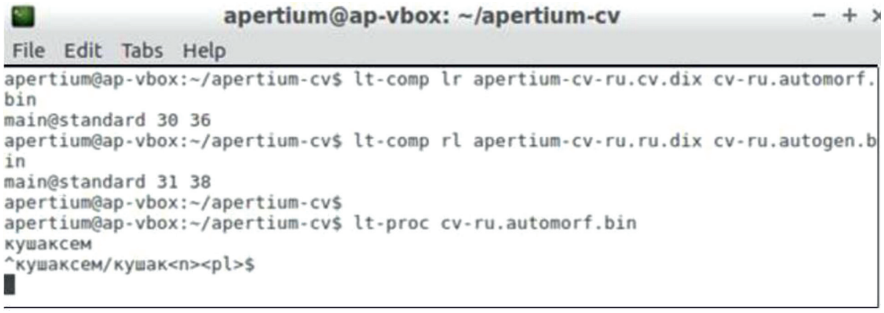

Рис. 3. Результат компиляции и тестирования словаря

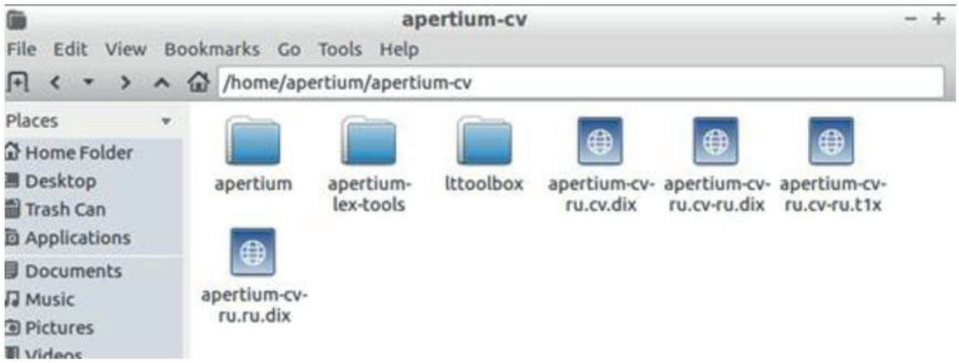

Рис. 4. Весь инструментарий и словари готовы к работе

Протестировать его можно, введя в командной строке какое-либо слово из словаря с парадигмой (т.е. в форме, отличной от леммы), например «кушаксем» – «кошки».

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.

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 – 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. 4).

Остается лишь скомпилировать словари для создания морфологических анализаторов, морфологических генераторов и поисковиков слов.

lt-comp lr apertium-cv-ru.cv.dix cv-ru.automorf.bin
lt-comp rl apertium-cv-ru.ru.dix cv-ru.autogen.bin
lt-comp lr apertium-cv-ru.ru.dix ru.cv.automorf.bin
lt-comp rl apertium-cv-ru.cv.dix ru-cv.autogen.bin
lt-comp lr apertium-cv-ru.cv-ru.dix cv-ru.autobil.bin
lt-comp rl apertium-cv-ru.cv-ru.dix ru-cv.autobil.bin

Теперь имеются два морфологических словаря и двуязычный словарь. Все, что сейчас необходимо, – это правила трансфера существительных.

Результаты исследования и их обсуждение

Откроем файл apertium-cv-ru.cv-ru.t1x и вставим в него базовый скелет.

<? xml version = "1.0" encoding = "UTF-8"?>
<перевод>
</ перевод>

Добавим необходимые разделы:

<section-def-cats>
</ section-def-cats>
<section-def-attrs>
</ section-def-attrs>

Так как слова у нас изменяются не только по числам, лицу и роду, но и по падежам, нам необходимо добавить все эти атрибуты. Но сначала добавим необходимую категорию:

<def-cat n="nom">
 <cat-item tags="n.*"/>
</def-cat>

Она "покрывает" все существительные (леммы, за которыми следует <n> и за ним еще что-нибудь) и ссылается на них как "nom".

В раздел атрибутов добавляем число, лицо, время и падежи.

Атрибуты числа:

<def-attr n="nbr">
<attr-item tags="sg"/>
<attr-item tags="pl"/>
</def-attr>

Атрибуты времени:

<def-attr n="temps">
<attr-item tags="pres"/>
<attr-item tags="past"/>
<attr-item tags="fut"/>
</def-attr>

Атрибуты падежей:

<def-attr n="case">
<attr-item tags="im"/>
<attr-item tags="ro"/>
<attr-item tags="da"/>
<attr-item tags="vi"/>
<attr-item tags="tv"/>
<attr-item tags="pr"/>
</def-attr>

Атрибуты лица:

<def-attr n="person">
<attr-item tags="p1"/>
<attr-item tags="p2"/>
<attr-item tags="p3"/>
</def-attr>

Далее нам необходимо добавить раздел для глобальных переменных.

<section-def-vars>
</section-def-vars>

Эти переменные используются для сохранения атрибутов или их передачи между несколькими правилами. Пока нам нужна только одна:

<def-var n="number"/>

Наконец, вам нужно добавить само правило, которое позволяет вам взять существительное и затем отобразить его в правильной форме.

Рассмотрим добавление глаголов. Наличие двуязычного словаря для системы машинного перевода чувашского языка позволяет переводить существительные. Однако на данный момент пользы от этого немного, ибо нам необходимо переводить и глаголы, и местоимения, и даже предложения. Начнем с глагола «видеть». В чувашском языке его эквивалентом является слово «курма». Следовательно, порядок преобразования будет таким:

куратăп.

Видеть<p1><sg> (Словоформа «видеть» первого лица единственного числа)

Вижу.

Переведем чувашское «кушаксене куратăп» в русское «вижу кошек»; в правилах нет шаблонов для глаголов, поэтому необходимо их добавить.

Для начала необходимо добавить символ для глагола, который будет иметь название «vblex» (verb lexical). Также вместе с числом у глаголов есть атрибуты лица и времени. Добавляем их:

<sdef n = "vblex" />
<sdef n = "p1" />
<sdef n = "pres" />

Как и с существительными, добавим парадигму спряжения глаголов. Первой строкой будет:

<pardef n="кур/ма__vblex">

Знаком «/» разграничивается слово на основную часть и часть, к которой будет добавляться содержимое из «l».

Затем добавим изменяющееся при склонении или спряжении окончание слова. Так как у нас первое лицо и единственное число, то результат будет таким:

<e> <l>атăп</l>
<r>ма<s n="vblex"/><s n="pri"/><s n="p1"/><s n="sg"/></r> </e>

Далее в основной раздел добавляем словоформу и коррелирующую с ней парадигму. Скомпилируем и проверим полученный результат (рис. 5).

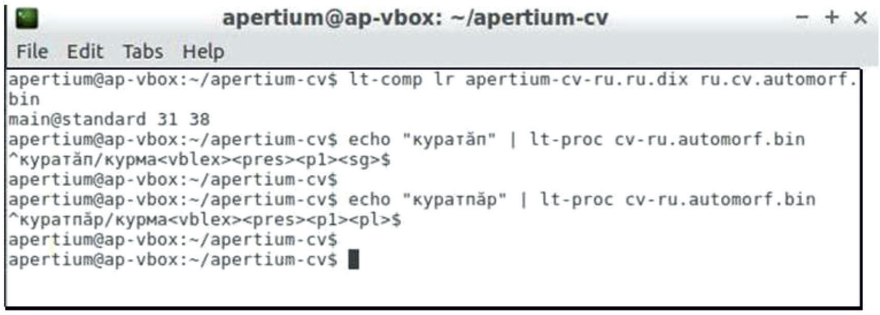

Рис. 5. Проверка корректности анализа глаголов

Также заполним и проверим русский словарь (рис. 6).

Рис. 6. Проверка корректности анализа глаголов в русском словаре

Осталось добавить обязательную запись в двуязычный словарь, скомпилировать и протестировать (рис. 7).

<e> <l>курма<s n="vblex"/></l><r>видеть<s n="vblex"/></r> </e>

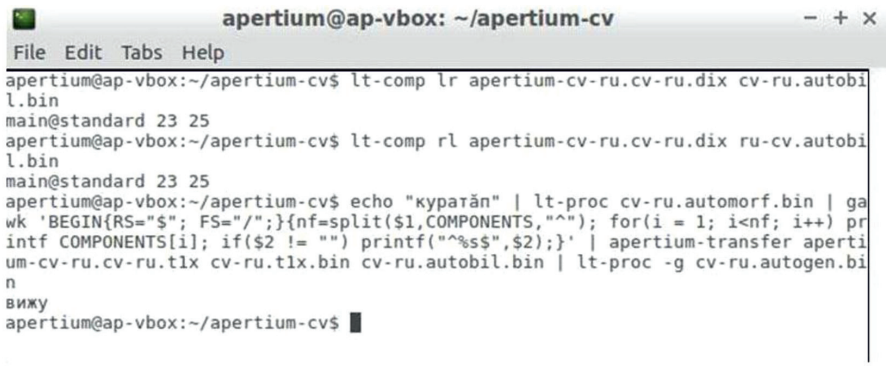

Рис. 7. Корректная генерация слова в конечном языке

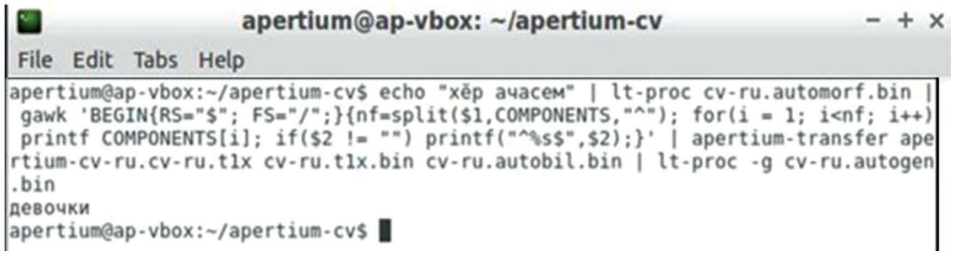

Рис. 8. Перевод слов во множественном числе и идиоматических выражений

Возникает проблема с идиоматическими выражениями. На данном этапе система будет переводить их дословно. Например, на чувашское «хĕр ача» переводчик будет выводить «девушка ребенок». А если подобное словосочетание стоит во множественном числе «хĕр ачасем», то правильный перевод должен быть «девочки». Чтобы получать корректный результат, добавим лемму, которая будет разрешать данный нюанс.

```
<e lm="хĕр ача"><i>хĕр<b/>ача</i><par n="вăрман__n"/></e>
```

Как можно заметить, нет необходимости создавать новую парадигму, а можно использовать, например, уже созданную у слова «вăрман» ‘лес’, которая есть в словаре. Результат вполне удовлетворительный (рис. 8).

Добавляем новые слова, чтобы переводчик получал все больше языковых данных и развивался.

Выводы

Рассмотрена разработка системы машинного перевода с чувашского на русский язык. Необходимо провести сравнение качества переводов относительно других систем и определить временные затраты на создание систем машинного перевода.

Список литературы

1. UNESCO Atlas of the World’s Languages in Danger. [Electronic resource]. URL: http://www.unesco.org/languages-atlas/en/atlasmap/language-id-338.html (date of access: 02.12.2020).

2. Семенов А.Л. Современные информационные технологии и перевод. М.: Академия, 2008. 224 с.

3. Желтов П.В. Национальный корпус чувашского языка: концепция и архитектура. Чебоксары: Изд-во Чувашского ун-та, 2017. 159 с.

4. PC-Kimmo. [Electronic resource]. URL: https://software.sil.org/pc-kimmo (date of access: 02.12.2020).

5. Apertium Documentation. [Electronic resource]. URL: http://wiki.apertium.org/wiki/Documentation (date of access: 02.12.2020).