Разработка системы машинного перевода с чувашского на русский язык.

В настоящее время обучение чувашскому языку в учебных заведениях испытывает определенные трудности. Чтобы он был конкурентоспособен с такими востребованными обществом языками как русский и английский, необходимо использовать современные методы. Одним из таких является модернизация обучения чувашскому языку в школе за счет привлечения компьютерных технологий. В данной ситуации, нам кажется весьма актуальной задача разработки системы машинного перевода для чувашского языка Рассмотрим возможность разработки на уроках чувашского языка компьютерных переводчиков с чувашского на русский и английский. Это позволит сделать обучение чувашскому языку более интересным и востребованным [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.

Существуют различные платформы машинного перевода: Apertium, PC-KIMMO и Trados и другие[1].

Apertium – платформа машинного перевода, которая разрабатывается при финансировании со стороны правительств Испании и Каталонии в Университете Аликанте (Universitat d'Alacant). Это свободное программное обеспечение, которое бесплатно издаётся разработчиками в соответствии с условиями GNU GPL [3].

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.

Apertium использует конечные преобразователи для всех своих лексических трансформаций, а также скрытые модели Маркова для выделения частей речи или устранения противоречий в категориях слов. К сожалению, более-менее стабильная реализация есть только на Unix-системах, поэтому крайне не рекомендуется использовать её на компьютерах с Windows и Macintosh.

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.

Следует отметить, что существует задел для чувашского языка в ситеме Apertium и чувашский язык имеется в списке языков перевода на сайте https://turkic.apertium.org. Русско-чувашский и чувашско-русский переводчики на базе Apertium были разработаны уроженцем Каталонии, проживающим в Чебоксарах, Эктором Алос-и-Фонтом.

Однако это дело не получило дальнейшего продолжения, поэтому весьма актуально продолжить создание системы чувашско-русского машинного перевода с помощью Apertium. Ввиду отсутствия финансирования этого проекта представляется целесообразным общественная инициатива. Большую помощь могли бы оказать учителя чувашского языка и преподаватели филологических факультетов местных вузов совместно с учениками школ и учащимися вузов. Более того, представляется интересным создание подобной системы на уроках чувашского языка в русскоязычных (городских) школах. Это позволило бы сделать их более интересными для учащихся.

Подробно в научной прессе применение платформы Apertium для создания переводчиков для различных тюркских языков была описана в трудах конференции TurkLang, см., например, в [4]. К сожалению, большинство работ выполнено на английском языке и

написано научным языком, в специальных терминах и потому малопонятны большинству учителей национальных языков. В публикуемой серии статей нашей задачей было дать элементарные инструкции по работе с ней простым и доступным языком для учителей чувашского языка и учащихся общеобразовательных школ. Поэтому данная статья, как и ряд других, публикуемых в данном сборнике и посвященных работе с Apertium, носит популяризаторский характер и предназначена для учителей чувашского языка и их учеников, а также для всех учителей национальных языков русскоязычных школ, которые с легкостью смогут применить изложенные рекомендации для своих языков.

Рассмотрим применение платформы Apertium для разработки системы чувашско-русского машинного перевода.

Основные используемые термины:

1) лемма – это каноническая форма слова, слово без грамматической информации. В русском и чувашском языках лемма существительного будет в именительном падеже и единственном числе;

2) символ В – контексте Apertium'а символ означает грамматический знак. Например, слово «стулья» – это существительное множественного числа, значит, у него будет символ существительного и символ множественного числа;

3) парадигма; так как записывать все окончания невероятно трудозатратно, то для показа словоформ используется парадигма – пример или точнее шаблон склонения или спряжения определенной группы слов.

При разработке чувашско-русского машинного перевода с использованием Apertium необходимо разработка языковой пары русского и чувашского языков.

Сначала необходимо создать электронный словарь исходного языка, где будут определены: алфавит, символы (в частности для существительных в единственном и во множественных числах именительного падежа) и словарь с двумя разделами – один стандартный, со словами, а второй – «безусловный», содержащий в себе знаки препинания.

Следующим шагом будет создание такого же электронного словаря, но уже для конечного языка.

После заполнения обоих словарей, необходимо перейти к следующему шагу – создания двуязычного электронного словаря – словаря, описывающего соответствия слов.

После всех предыдущих действий имеются два морфологических словаря и один двуязычный словарь. Все, что еще необходимо сделать – это написать правила трансфера существительных. По сути, это объявление категорий и атрибутов, которые, в свою очередь, позволяют объединять грамматические символы.

Эта система уже способна переводить существительные. Но до полноценного переводчика этого недостаточно, т.к. кроме существительных предложения содержат еще другие части речи. Необходимо дополнить словари и перекомпилировать переводчик.

Чувашско-русский машинный перевод, как и любой другой, реализуется с помощью Apertium следующим образом.

1. Текст на исходном языке передается в Apertium для перевода.

2. Деформатор убирает язык разметки (HTML, RTF и т.д.) который должен храниться на месте, но не должен быть переведен.

3. Морфологический анализатор сегментирует текст и ищет данные сегменты в языковых словарях, а затем возвращает основную форму и теги со всех совпадений.

4. Морфологический определитель контекста разрешает спорные сегменты (сегменты со множественными совпадениями) выбором наиболее подходящего.

5. Лексический трансфер ищет неоднозначные слова исходного языка, чтобы найти их эквиваленты в конечном языке.

6. Происходит лексический отбор между альтернативными переводами, когда слово исходного текста имеет разные значения.

7. Структурный трансфер изменяет текст исходного языка и подгоняет его под грамматически корректную форму конечного языка.

8. Происходят необходимые грамматические изменения из-за связок слов.

9. Возвращается форматирование, убранное в пункте 2.

10. Apertium выводит перевод на конечный язык.

1. 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:

курайăн.

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="курăма__vblex">

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

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

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

Далее в основной раздел добавляем словоформу и коррелирующую с ней парадигму. Скомпилируем и проверим полученный результат (рис.2).

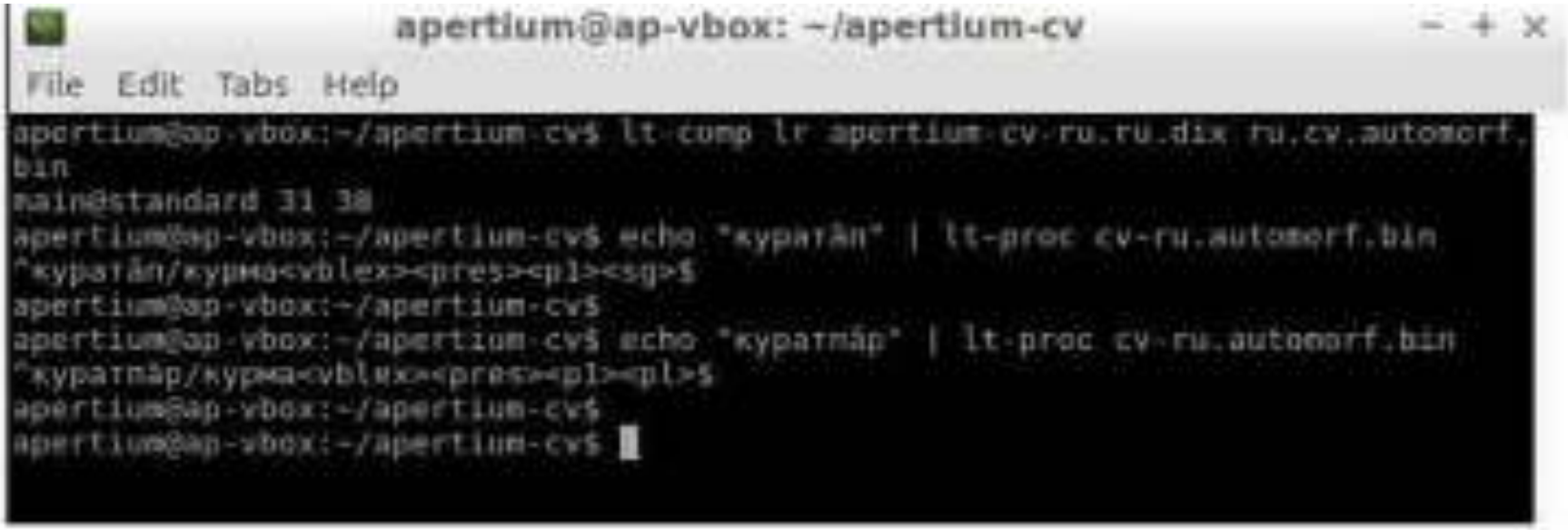

Рис. 2. Проверка корректности анализа глаголов

Также заполним и проверим русский словарь (рис.3.).

Рис.3. Проверка корректности анализа глаголов в русском словаре

Осталось добавить обязательную запись в двуязычный словарь, скомпилировать и протестировать (рис.4).

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

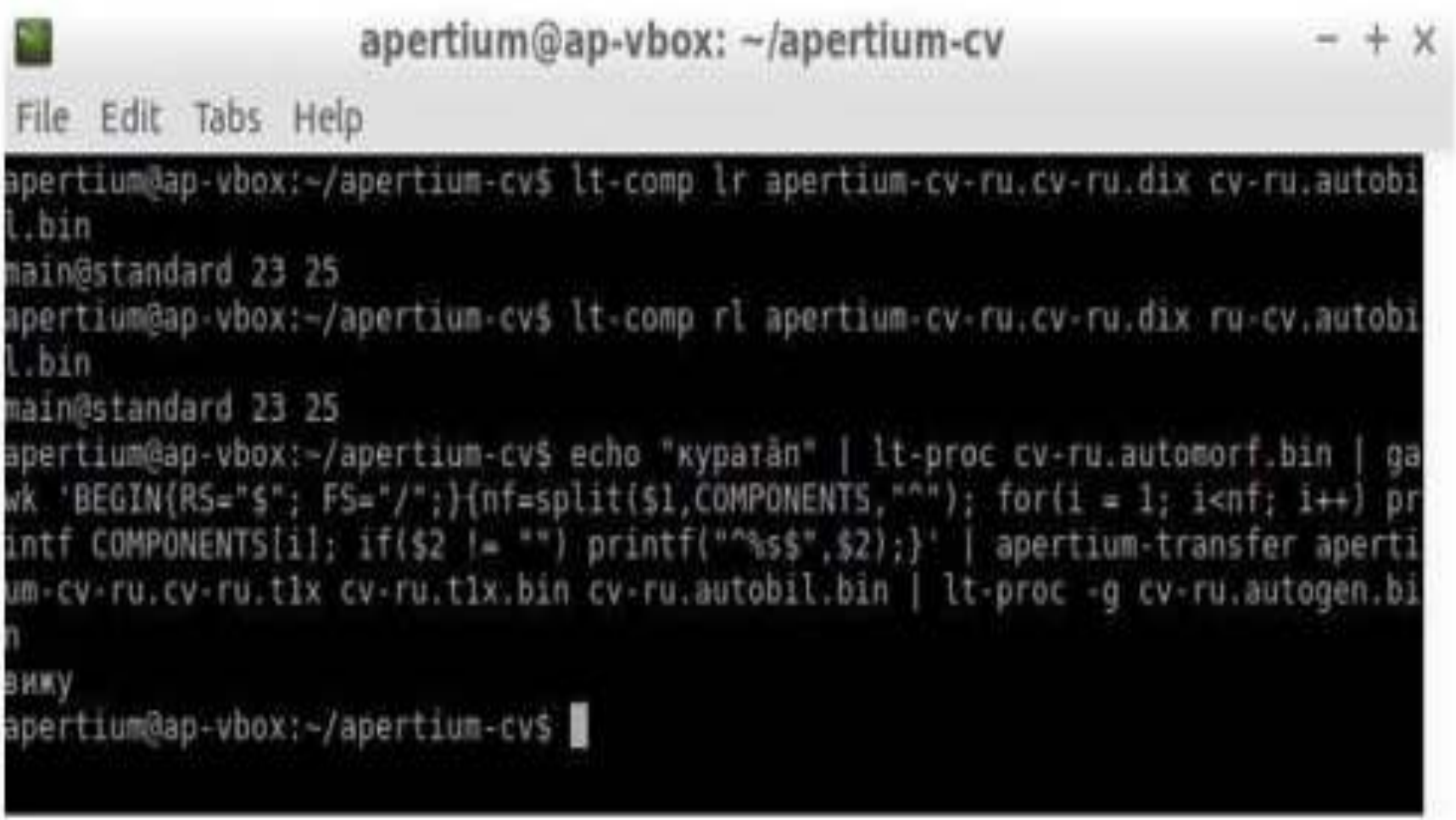

Рис. 4. Корректная генерация слова в конечном языке.

2. Слова во множественном числе и идиоматические выражения

Возникает проблема с идиоматическими выражениями. На данном этапе он будет переводить их дословно. Например, на чувашское «хĕр ача» переводчик будет выводить «д ушка 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">хĕр<b/>ача<par n="вăрман_n"/></e>

Как можно заметить, нет необходимости создавать новую парадигму, а можно использовать, например, уже созданную у слова «вăрман» ‘лес’, которое есть в словаре. Результат вполне удовлетворительный (рис.5).

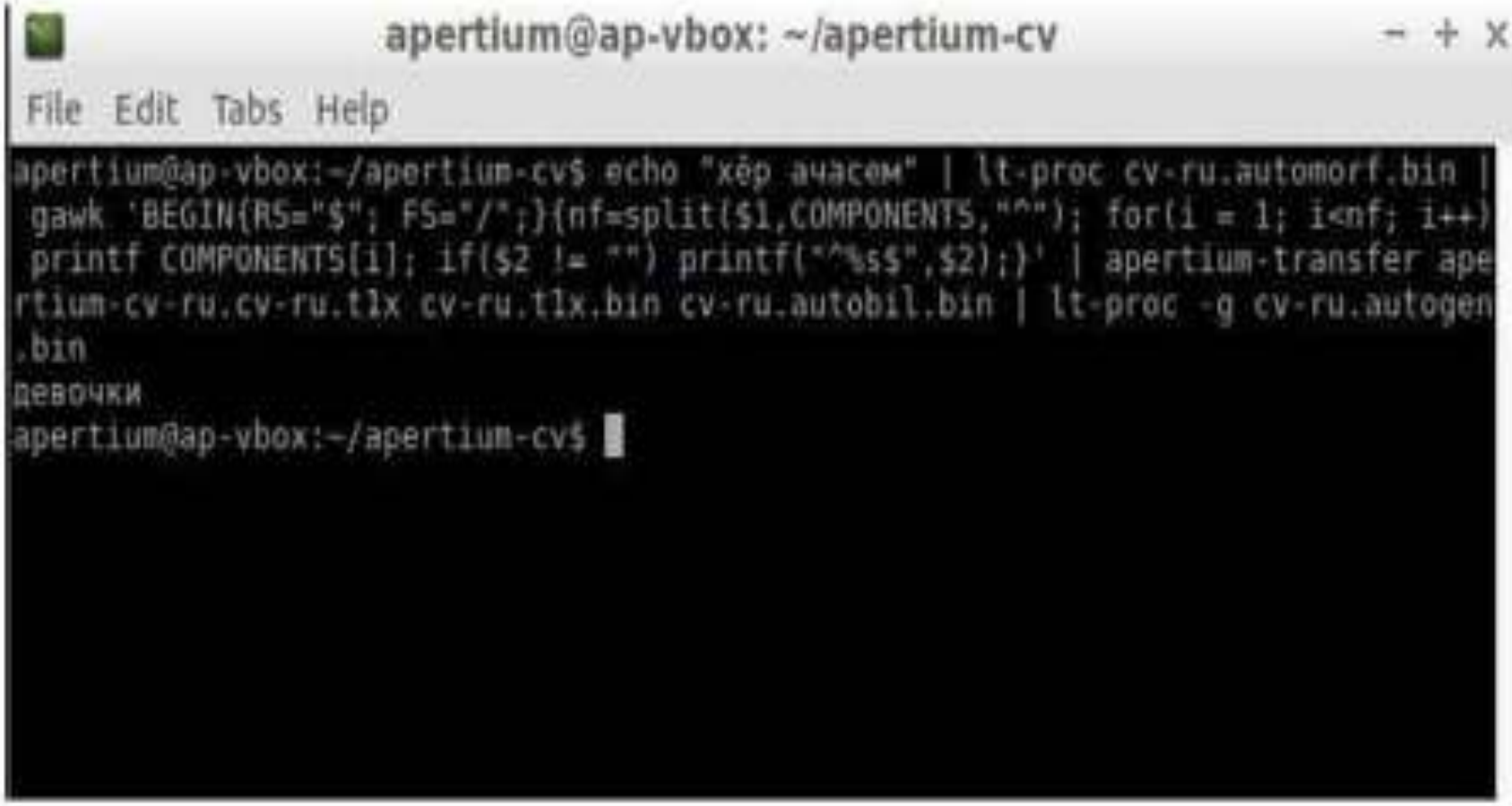

Рис.5. Перевод слов во множественном числе и идиоматических выражений

Остается лишь добавлять новые слова, чтобы переводчик получал все больше языковых данных и развивался.

Выводы

PC-KIMMO и Apertium являются наиболее перспективными платформами для реализации чувашско-русского машинного перевода. Для окончательного выбора между этими системами необходимо провести сравнительный анализ качества переводов, реализованных системами PC-KIMMO и Apertium, и затрат по времени на создание систем машинного перевода с их использованием.

Рассмотрена разработка системы машинного перевода с чувашского на русский язык на представленных в статье элементарных примерах.

Литература

1. UNESCO Atlas of the World's Languages in Danger (англ.) — http://www.unesco.org/languages-atlas/en/atlasmap/language-id-338.html.
2. PC-Kimmo (англ) — Режим доступа: https://software.sil.org/pc-kimmo/.
3. Apertium Documentation (англ.) — Режим доступа: http://wiki.apertium.org/wiki/Documentation.
4. Balzhan A. et al. Study of the problem of creating structural transfer rules and lexical selection for the Kazakh-Russian machine translation system on Apertium platform // Proceedings of the International Conference on Computer processing of Turkic Languages "TurkLang-2015". Казань, 2015. PP. 5-9.
5. Ахо А., Ульман Дж. Теория синтаксического анализа, перевода и компиляции. М.: Мир, 1978. Т.1. 612 с. Т.2. 487 с.
6. Семенов А.Л. Современные информационные технологии и перевод. М.: Академия, 2008. – 224 с.
7. Чăвашла-вырăсла словарь: 40000 сăмаха янăн / ред. М. Н. Скворцов ; Чăваш АССР Министрсен Совечĕ çумĕнчи "Хисеп Паллă" орденлă чĕлхе, литература, истори тата экономика наука тĕпчев институчĕ. — Мускав : Изд-во "Русский язык", 1982. — 712 с.