ИНТЕЛЛЕКТ
ЯЗЫК
КОМПЬЮТЕР

Выпуск
17

ТРУДЫ

МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ ПО КОМПЬЮТЕРНОЙ И КОГНИТИВНОЙ ЛИНГВИСТИКЕ

TEL–2016

Казань, 21–24 апреля 2016 г.

ИНТЕЛЛЕКТ. ЯЗЫК. КОМПЬЮТЕР

Выпуск 17

ТРУДЫ МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ ПО КОМПЬЮТЕРНОЙ И КОГНИТИВНОЙ ЛИНГВИСТИКЕ

TEL-2016

Казань, 21–24 апреля 2016

КАЗАНЬ
2016

УДК 801+681.3
ББК 81.1
Т78

Академия наук Республики Татарстан
Институт прикладной семиотики АН РТ

Казанский федеральный университет
Институт филологии и межкультурной коммуникации им. Льва Толстого
Высшая школа информационных технологий и информационных систем
Институт вычислительной математики и информационных технологий

Российский фонд фундаментальных исследований
Российская ассоциация искусственного интеллекта

Издание осуществлено при финансовой поддержке
Казанского федерального университета,
Академии наук Республики Татарстан
и Российского фонда фундаментальных исследований
(проект № 16-07-20112г).

Печатается по постановлению
Редакционно-издательского совета
Казанского федерального университета

Научные редакторы:
академик АН РТ, профессор **Д.Ш. Сулейманов**;
доцент **О.А. Невзорова**

Т78 Труды международной конференции по компьютерной и когнитивной лингвистике TEL-2016. – Казань: Изд-во Казан. ун-та, 2016. – 392 с.

ISBN 978-5-00019-650-2

Сборник содержит материалы Международной конференции по компьютерной и когнитивной лингвистике TEL-2016 (Казань, 21–24 апреля 2016).

Для научных работников, преподавателей, аспирантов и студентов, специализирующихся в области компьютерной и когнитивной лингвистики и ее приложений.

УДК 801+681.3
ББК 81.1

ISBN 978-5-00019-650-2

Предисловие

Компьютерная лингвистика – современное научное направление, ориентированное на разработку компьютерных моделей, методов и технологий в лингвистике и смежных областях. В списке актуальных направлений компьютерной лингвистики можно указать распознавание и синтез речи, анализ и генерация текстов, машинный перевод, компьютерный анализ документов (реферирование, классификация, поиск), вопросно-ответные системы, извлечение знаний из текстов, лингвистические технологии в Интернете, онтологии и лингвистические базы данных, компьютерная лексикография, корпусная лингвистика и когнитивное моделирование языка. Фактически, перечисленные выше направления в настоящее время являются самостоятельными науками со своими методами и технологиями лингвистического моделирования. Практически каждое направление имеет свои организационные структуры в форме ассоциаций и специальных исследовательских групп, которые осуществляют издательскую деятельность и проводят международные научные конференции.

В сборнике представлены статьи по актуальным проблемам когнитивной и компьютерной лингвистики, включая формальные модели синтаксиса и семантики, системы поиска и классификации, онтологии, корпуса и лингвистические базы данных, программные системы обработки ЕЯ и др.

Тематика конференции находится в постоянном развитии. В 2014 году возникло новое направление, связанное с задачами тюркской компьютерной и корпусной лингвистики, в том числе с задачей унификации систем грамматической разметки в корпусах тюркских языков. Это направление было активно представлено в программе конференции TEL-2016. В работе конференции приняли участие ведущие разработчики основных тюркских корпусов из Казахстана, Якутии, Кыргызстана, Башкортостана, России.

Организаторы конференции выражают благодарность директору Института филологии и искусств КФУ Замалетдинову Р.Р., директору Высшей школы информационных технологий и информационных систем КФУ Хасьянову А.Ф., директору Института вычислительной математики и информационных технологий КФУ Мосину С.Г., а также сотрудникам института «Прикладная семиотика» АН РТ за их вклад в организацию и успешное проведение конференции TEL-2016.

Научные редакторы:
Сулейманов Д.Ш., Невзорова О.А.

РАЗДЕЛ I | СЕМАНТИЧЕСКИЕ ТЕХНОЛОГИИ

УДК 004.91

АВТОМАТИЗИРОВАННАЯ ОБРАБОТКА АРХИВНОЙ КОЛЛЕКЦИИ НАУЧНОГО ЖУРНАЛА «ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ»

Д.Ю. Ахметов
Казанский (Приволжский) федеральный университет
akhmetov.dy@gmail.com

Представлены разработанные и апробированные способы автоматического извлечения метаданных из архивов электронного научного журнала «Электронные библиотеки». Разработан программный комплекс выделения и обработки метаданных статей журнала, реализованный на языке PHP с использованием технологий CURL, html dom и htmlsql.

Ключевые слова: *автоматическое извлечение метаданных, семантический веб, xml*

С развитием современных технологий Семантического Веба и библиометрии стало возможным сравнительно легко автоматически выделять метаданные из статей, подаваемых в редакции научных журналов. Кроме того, использование инструментов автоматизации редакционно-издательских процессов позволило максимально полно определять все метаданные научной публикации (название, аффилиацию, аннотацию, ключевые слова и др.). Одновременно электронные журналы, использующие автоматизированные информационные системы организации редакционно-издательских процессов (отказавшиеся, например, от традиционной формы приёма статей по электронной почте) и имеющие архивные выпуски, размещенные, в частности, на своих сайтах, вынуждены переводить свои архивы в новые форматы для поисковой оптимизации и улучшения извлечения библиометрических данных соответствующих статей и журнала в целом [1–3].

В настоящей работе предложены и апробированы методы управления электронным контентом, в том числе выделение метаданных

из архивных коллекций журнала «Электронные библиотеки», размещенных на сайте http://elbib.ru, и автоматического преобразования их в формат .xml для обеспечения возможности семантической обработки, в частности, нахождения «похожих» публикаций по ключевым словам [4], работы с библиографией [5], сбора наукометрических данных об авторах [6]. Все архивы журнала «Электронные библиотеки» являются набором html-страниц (все выпуски, каждая статья – на отдельной странице, без структурного разделения метаданных на блоки и семантической разметки). Выпуски разбиты по годам (см. рис. 1), на странице каждого выпуска размещены названия статей и авторов (рис. 2).

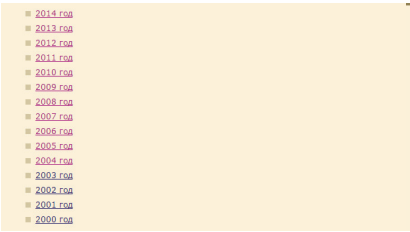


Рис. 1

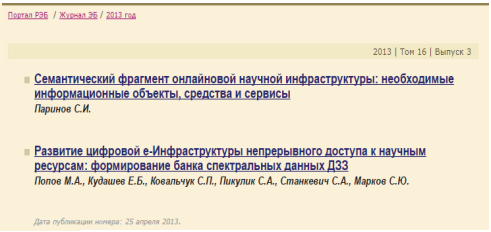


Рис. 2

Разработанный программный комплекс выделения и обработки метаданных состоит из трех модулей. Первый модуль отвечает за выделение метаданных из архивов журнала «Электронные библиотеки», второй модуль генерирует блок авторов и ключевых слов из полученных метаданных, третий модуль размещен на новом сайте журнала (http://elbib.kpfu.ru) и обрабатывает базы данных авторов, ключевых слов и выпусков журнала, формирует соответствующие веб-страницы. Модули реализованы на языке PHP с исполь-зованием технологий CURL[https://curl.haxx.se/], html dom, htmlsql. Результатом работы алгоритмов являются xml-файлы. Статьи из архивов журнала сконвертированы в формат .pdf при помощи технологий tcpdf, mpdf и онлайн-сервисов по конвертации данных. Выбор способа хранения метаданных статей журнала «Электронные библиотеки» в xml-файлах обусловлен тем, что этот формат позволяет структурированно хранить информацию в виде текстовых файлов без применения какой-либо базы данных. Кроме того, хранимые данные представляет собой однородную неизменяемую информацию, исчисляемую килобайтами.

К некоторым выявленным недостаткам можно отнести то, что полученные данные дублируют друг друга в различных частях

разработанной системы, однако в силу того, что метаданные опубликованных работ представляют собой неизменяемую информацию, имеющаяся избыточность метаданных не оказывает негативного воздействия на процесс реализации построенных алгоритмов.

Каждый выпуск журнала переведен в xml-представление с сохранением статей внутри выпуска (см. рис. 3). Хранение информации обо всех авторах публикации внутри одного тега <authors> связано с тем, что разделение списка на отдельных авторов происходит программно.

```
<?xml version="1.0" encoding="utf-8"?>
<issue>
  <volume>18</volume>
  <title>№1-2</title>
  <articles>
    <article>
    <article>
    <article>
        <id>3</id>
        <title>АВТОМАТИЗАЦИЯ РЕДАКЦИОННЫХ ПРОЦЕССОВ В ИНФОРМАЦИОННОЙ СИСТЕМЕ УПРАВЛЕНИЯ
        ЭЛЕКТРОННЫМИ НАУЧНЫМИ ЖУРНАЛАМИ</title>
        <authors>Д.Ю. Ахметов, А.М. Елизаров, Е.К. Липачев</authors>
        <annotation>Исследованы особенности использования информационных систем в процессе
        издания электронных научных журналов и проведено их сравнение с точки зрения
        автоматизации редакционных процессов. Описаны программные модули, созданные для
        расширения функционала платформы Open Journal Systems в целях автоматизации ряда
        редакционных процессов электронного научного журнала. Приведены алгоритмы автоматической
        стилевой валидации текстов на этапе регистрации автором статьи в информационной системе
        электронного научного журнала, автоматического подбора рецензентов, рассылки уведомлений
        и контроля сроков рецензирования.</annotation>
        <keywords>издательские системы; электронный научный журнал; интеграция электронных
        ресурсов; данных научного цитирования; экстракция метаданных; Open Journal Systems
        </keywords>
        <pages>32-45</pages>
```

Рис. 3

Кроме того, для каждого автора создан индивидуальный xml-файл (рис. 4), содержащий информацию обо всех работах, опубликованных им в этом журнале (статьи каждого автора легко находятся за весь период).

```
<?xml version="1.0" encoding="utf-8"?>
<articles>
    <article>
        <id>2015;1;3</id>
        <title>АВТОМАТИЗАЦИЯ РЕДАКЦИОННЫХ ПРОЦЕССОВ В ИНФОРМАЦИОННОЙ СИСТЕМЕ УПРАВЛЕНИЯ
        ЭЛЕКТРОННЫМИ НАУЧНЫМИ ЖУРНАЛАМИ</title>
        <authors>Д.Ю. Ахметов, А.М. Елизаров, Е.К. Липачев</authors>
        <volume>18</volume>
        <number>№1-2</number>
    </article>
</articles>
```

Рис. 4

Идентификатор <id> содержит полный относительный путь (примененный при реализации) к xml-файлу статьи (папка: «2015», файл: «1.xml», порядковый номер статьи: 3).

Каждое ключевое слово также хранится в отдельном xml-файле (рис. 5) для последующей обработки (определение частоты использования, организация быстрого поиска).

```
<?xml version="1.0" encoding="utf-8"?>
<keyword>
    <name>издательские системы</name>
    <articles>
        <article>
            <id>2015;1;2</id>
            <year>2015</year>
            <title>Формирование метаданных для международных баз цитирования в системе управления
            электронными научными журналами</title>
            <authors>А.Н. Герасимов, А.М. Елизаров, Е.К. Липачев</authors>
            <volume>18</volume>
            <number>№1-2</number>
        </article>
        <article>
            <id>2015;1;3</id>
            <year>2015</year>
            <title>АВТОМАТИЗАЦИЯ РЕДАКЦИОННЫХ ПРОЦЕССОВ В ИНФОРМАЦИОННОЙ СИСТЕМЕ УПРАВЛЕНИЯ
            ЭЛЕКТРОННЫМИ НАУЧНЫМИ ЖУРНАЛАМИ</title>
            <authors>Д.Ю. Ахметов, А.М. Елизаров, Е.К. Липачев</authors>
            <volume>18</volume>
            <number>№1-2</number>
        </article>
        <article>
            <id>2015;1;5</id>
```

Рис. 5

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проекты №№ 15-07-08522, 15-47-02472).

Литература

1. Бездушный А.А. Управление личными каталогами научных публикаций с использованием технологий Semantic Web // Вестник НГУ. Серия: Информационные технологии. 2015. Т. 13, вып. 1. С. 16–23.

2. Ахметов Д.Ю., Елизаров А.М., Липачев Е.К. Автоматизация редакционных процессов в информационной системе управления электронными научными журналами // Электронные Библиотеки. 2015. Т.18 (1-2). С. 32–45.

3. Ахметов Д.Ю., Елизаров А.М., Липачев Е.К. Информационные системы и сервисы комплексной поддержки периодических научных изданий // Научный сервис в сети Интернет: труды XVII Всероссийской научной конференции (21–26 сентября 2015 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2015. С. 16–25.

4. Елизаров А.М., Жижченко А.Б., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К. Онтологии математического знания и рекомендательная система для коллекций физико-математических документов// Докл. РАН. 2016. Т. 467, №4. С. 392–395.

5. Герасимов А.Н., Елизаров А.М., Липачёв Е.К. Формирование метаданных для международных баз цитирования в системе управления электронными научными журналами // Электронные библиотеки. 2015. Т. 18, №1-2. С. 6–31.

6. Когаловский М.Р., Паринов С.И. Наукометрические измерения в электронных библиотеках на основе рубрикаторов научной информации //Электронные библиотеки. 2012. Т. 15, № 6.

УДК 519.683.8

РАЗРАБОТКА ПРИЛОЖЕНИЯ ПОДДЕРЖКИ ПРИНЯТИЯ УПРАВЛЕНЧЕСКОГО РЕШЕНИЯ

А.Ф. Бакунина, К.С. Цыбенко
Казанский федеральный университет, Казань
bakuninaa@gmail.com, ktsybenko@mail.ru

В статье описывается способ разработки приложения поддержки принятия управленческого решения средствами Visual Basic for Application и Microsoft Office Excel, основанного на методе анализа иерархий Саати.

Ключевые слова: *Visual Basic for Application, Microsoft Office Excel, метод Саати.*

Введение

Проблема выбора оптимального товара или услуги из множества альтернативных вариантов в современном индустриальном обществе стоит очень остро. Рациональный выбор оборудования, стратегии развития предприятия, способа распределения пакета инвестиций обеспечивает самое выгодное вложение денежных средств и получение максимальной пользы. В данной работе описывается способ создания приложения, помогающего принять управленческое решение в выборе модели автомобиля. Приложение будет создаваться средствами Visual Basic for Application и Microsoft Office Excel. Оптимальное решение будет определяться на основе одного из самых популярных и эффективных способов - метода анализа иерархий Томаса Саати (Analytic Hierarchy Process).

Краткая характеристика Visual Basic for Application, Microsoft Office Excel и метода Саати. Microsoft Excel (также иногда называется Microsoft Office Excel) – программа для работы с электронными таблицами, созданная корпорацией Microsoft для Microsoft Windows,

Windows NT и Mac OS, а также Android, iOS и Windows Phone. Она предоставляет возможности экономико-статистических расчетов, графические инструменты и язык макропрограммирования VBA (Visual Basic for Application). Microsoft Excel входит в состав Microsoft Office и на сегодняшний день Excel является одним из наиболее популярных приложений в мире [2].

Visual Basic (VB) является языком программирования третьего поколения (событийный язык программирования) и средой разработки от Microsoft для модели программирования COM. Этот язык был получен из BASIC и допускает быструю прикладную разработку (RAD) графического интерфейса пользователя (GUI), доступ к базам данных при помощи DAO, RDO, ADO, создание элементов управления Active X и объектов. Языки сценариев (VBA, VB Script) синтаксически подобны Visual Basic [3].

Visual Basic for Applications (VBA) – немного упрощённая реализация языка программирования Visual Basic, встроенная в линейку продуктов Microsoft Office (включая версии для Mac OS), а также во многие другие программные пакеты, такие как Auto CAD, Solid Works, Corel DRAW, Word Perfect и ESRI Arc GIS. VBA покрывает и расширяет функциональность ранее использовавшихся специализированных макроязыков, таких как Word Basic. VBA является интерпретируемым языком. VBA, будучи языком, построенным на COM, позволяет использовать все доступные в операционной системе COM объекты и компоненты Active X. К достоинствам языка можно отнести сравнительную лёгкость освоения, благодаря которой приложения могут создавать даже пользователи, не программирующие профессионально. К особенностям VBA можно отнести выполнение скрипта именно в среде офисных приложений.

Одним из вариантов ситуации принятия управленческого решения является критериальная постановка. Для получения обоснованного «лучшего» решения применяют методы критериального анализа иерархий или метод Саати. Основа метода Саати – попарные сравнения альтернатив по каждому из критериев и попарное сравнение критериев с точки зрения важности для поставленной цели. Его суть заключается в следующем: в ходе парных сравнений двух альтернатив по каждому из критериев выставляются оценки превосходства одной альтернативы над другой по шкале Томаса Саати; полученные оценки пересчитываются по одному из четырех алгоритмов, которые позволяют получить суммарную оценку по каждому претенденту[1]. Шкала показана на рисунке.

Качественное сравнение	Количественный аналог	Качественное сравнение	Количественный аналог
равно, одинаково, безразлично	1	равно, одинаково, безразлично	1
немного лучше, важнее	3	немного хуже, менее важнее	1/3
лучше, важнее	5	хуже, менее важно	1/5
значительно лучше, важнее	7	значительно хуже, менее важно	1/7
принципиально лучше, важнее	9	принципиально хуже, менее важно	1/9

Рис. 1. Шкала критериев

В случае если лицо, принимающее решение (ЛПР), не может определиться между двумя качественными признаками, при наличии промежуточного мнения, Саати рекомендует использовать промежуточные баллы 2, 4, 6, 8.Этапы применения метода Саати:

1. Выделение проблемы. Определение цели.
2. Выделение основных критериев, обуславливающих достижение цели.
3. Выделение группы альтернатив, представляющих наибольший интерес.
4. Построение иерархии: дерево от цели через критерии к альтернативам.
5. Построение матрицы попарных сравнений критериев по цели.
6. Построение матриц попарных сравнений альтернатив по критериям.
7. Применение методики анализа полученных матриц.
8. Определение весов альтернатив по системе иерархии.

Описание разработки приложения поддержки принятия управленческого решения. Разработка приложения начинается с запуска редактора Visual Basic и базы данных. В конструкторе таблиц создается исходная таблица. Пользователю предлагается выбрать основные интересующие его критерии.

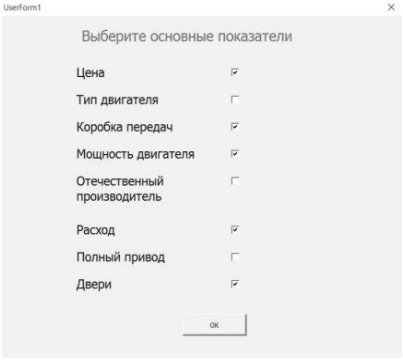


Рис.2. Окно выбора критериев

После подтверждения выбранных критериев пользователю предлагается оценить критерии, выбирая соответствующую характеристику из выпадающего списка. Затем качественные характеристики переводятся в количественные (рис.3.).

Рис. 3. Распределение критериев по приоритету

После нажатия на кнопку введенные данные переносятся на Лист Excel.

```
Private Sub CommandButton1_Click()
    Worksheets("Лист1").Range("B2").Value = TextBox17.Text
    Worksheets("Лист1").Range("B3").Value = TextBox1.Text
    Worksheets("Лист1").Range("B4").Value = TextBox7.Text
    Worksheets("Лист1").Range("B5").Value = TextBox12.Text
    Worksheets("Лист1").Range("B6").Value = TextBox22.Text
    …
EndSub
```

В итоге получим:

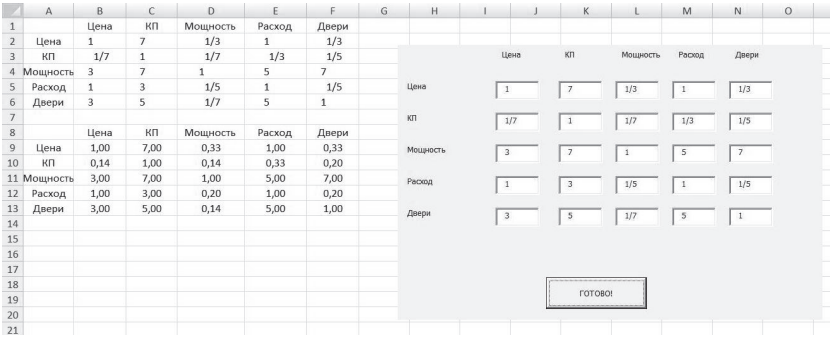


Рис. 4. Заполнение матрицы критериев

Аналогичным образом строятся матрицы сравнения отдельных альтернатив по каждому из критериев.

Согласно методу Саати производится расчет для получения весов альтернатив с точки зрения достижения поставленной цели.

В результате получим пять матриц, каждая из которых нормируется, и определяются веса строк. Также нормируется матрица сравнения критериев, что дает возможность выявить самый важный критерий.

	Цена	КП	Мощность	Расход	Двери	СРЗНАЧ
Цена	0,122805	0,304347826	0,18306762	0,081083	0,038131	0,145887
КП	0,017561	0,043478261	0,07861462	0,027001	0,022902	0,037911
Мощность	0,368415	0,304347826	0,54975261	0,405416	0,801557	0,485898
Расход	0,122805	0,130434783	0,10995052	0,081083	0,022902	0,093435
Двери	0,368415	0,217391304	0,07861462	0,405416	0,114508	0,236869

Например, из получившейся матрицы можно сделать вывод о том, что самым важным критерием является Мощность (49%).

Аналогичным образом получаются веса критериев. После этого полученная матрица умножается на столбец весов критериев по цели. Альтернатива, которой соответствует наибольшее значение, станет управленческим решением.

Машина1	0,355831684	не решение
Машина2	0,279198085	не решение
Машина3	0,364970231	Управленческое решение

Рис. 5. Управленческое решение

Отметим, что полученные оценки отражают исключительно точку зрения конкретного ЛПР.

Литература

1. Красильников В. Выбор мобильного телефона по методу Саати/ Компьютерные вести, №47, 2005. [Электронный ресурс]. URL: http://old.kv.by/index2005471201.htm
2. Общие сведения о Microsoft Excel [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/Microsoft_Excel
3. Visual Basic [Электронный ресурс]. URL: http://progopedia.ru/ dialect/visual-basic/

УДК 681.4

МОДЕЛИ ВЕКТОРНОГО ПРЕДСТАВЛЕНИЯ СЛОВ НА ОСНОВЕ ТЕХНОЛОГИИ WORD2VEC

Ф.М. Гафаров, Э.И. Шайдуллина, В.Р. Гафарова
Казанский федеральный университет, Казань
fgafarov@yandex.ru, lelechka_29@mail.ru

Аннотация

В статье описана технология векторного представления слов – word2vec, анализируются ее преимущества и недостатки. Изучены возможности технологии на корпусах татарского, английского и русского языков.

Ключевые слова: *word2vec, векторное представление слов, CBOW, continuous bag-of-words, skip-gram, softmax.*

В 2013 году исследователь Томас Миколов из компании Google опубликовал статью «Efficient Estimation of Word Representations in Vector Space», после чего выложил код утилиты word2vec. Word2vec – это инструмент для получения векторного представления слов на

естественном языке. Векторные представления слов были известны и ранее, но не получили широкого распространения из-за сложных алгоритмов и долгих вычислений. Метод word2vec превзошел все существующие ранее нейросетевые языковые модели по точности и быстроте обнаружения семантических связей между словами.

Word2vec – это технология статистической обработки больших массивов текстовой информации [1]. Она делает отображение текстового корпуса на множество векторных представлений слов из этого корпуса. Такие векторные представления основаны на том, как часто слова встречаются вместе или при каких похожих ситуациях они используются. Word2vec можно использовать для всевозможных задач обработки естественного языка:

- выявления смыслового сходства между словами;
- анализа тональности;
- классификации текстов
- машинный перевод
- информационный поиск
- фильтрация нежелательного содержимого и т.д.

Основу word2vec составляют два алгоритма: Continuous Bag-of-Words (CBOW) и skip-gram. Данные модели реализуются при помощи двух-либо трехслойной нейронной сети. Основным преимуществом этих моделей являются значительно меньшие затраты на вычисления и обучение, по сравнению с ранее известными подходами. В основном это достигается благодаря использованию иерархической функции softmax, основанной на представлении слов в виде дерева Хаффмана. Тогда как полный softmax, используемый в модели Bengio, существенно замедляет работу соответствующей нейронной сети.

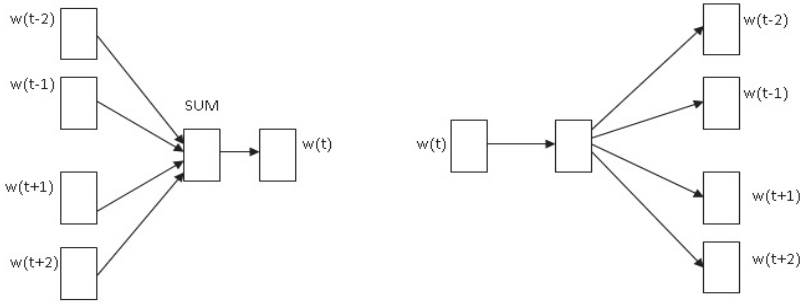


Рис.1. Схематическое представление моделей CBOW и Skip-gram

Continuous bag-of-words – эта модель пытается предсказать центральное слово w(t) на основе окружающего его контекста слов w(t-n),…,w(t-1), w(t+1),…, w(t+n). Особенностью модели является динамический размер окна: число n принимает значение от 1 до N. Обучение нейронной сети заключается в максимизации функции:

$$L = \sum_{t,k} \log P(w(t)|w(t-k), \dots, w(t-1), w(t+1), \dots, w(t+k))$$

Схема работы модели CBOW состоит в следующем. На втором (скрытом) слое осуществляется усреднение распределенных векторов. Число нейронов в нем равно размерности распределенных векторов. На третьем слое реализуется иерархическая функция softmax. Каждой не листовой вершине дерева Хаффмана соответствует один нейрон третьего слоя с синаптическими весами.

Предсказание слова w(t) осуществляется следующим образом. Допустим, имеется некоторое множество X нейронов третьего слоя. Каждый нейрон из этого множества осуществляет скалярное умножение вектора своих синоптических весов на вектор выходных сигналов второго слоя. К получившемуся результату применяют логистическую функцию, например, функцию вида:

$$\varphi(u) = (1 + \exp(-\alpha u))^{-1}, \alpha > 0$$

Совокупность выходных сигналов нейронов множества X сравнивается с кодом Хаффмана для слова w(t) (число выходных сигналов нейрона равно длине кода Хаффмана для слова w(t)). Цель обучения заключается в том, чтобы сделать их как можно ближе. В итоге получается:

$$P\left(w(t)\middle|w(t-k), \dots, w(t-1), w(t+1), \dots, w(t+k)\right) = \prod_i |w(t_i) - w^*(t_i)|,$$

здесь $w(t_i) \in \{0,1\}$ – цифра, стоящая в позиции i в коде Хаффмана слова w(t). После вычисления $w(t_i)$ происходит коррекция синоптических весов нейронов первого слоя.

При этом ошибка равна:

$$E = -\sum_{c=1}^{T} \log p(w(t)|w(t-c), \dots, w(t-1), w(t+1), \dots, w),$$

c – размер контекстного окна. Целью обучения является также и минимизация функции ошибки.

Разработчиками из Google было показано, что вектора обладают некой лингвистической систематичностью. Например: пара слов <Париж, Франция> представляет собой семантическое отношение <столица, страна>. Задача состоит в следующем: имеются пары слов <X1, X2> и <Y1, Y2>, между которыми имеется семантическая связь. При заданных X1,X2, Y1 необходимо найти Y2. Т.е., если имеются слова (мешок слов – bag-of-words) «Берлин, Германия, Париж», результатом работы модели будет «Франция».

В модели skip-gram используется другая целевая функция:

$$L = \sum_{t,k} \sum_{j \in Context(t)} \log P(w(j)|w(t))$$

В данной модели применяется двухслойная нейронная сеть. На втором слое реализуется иерархический softmax.

Skip-gram отличается от CBOW тем, что слово w(t) предсказывается столько раз сколько слов в контексте.

Снова введем некое множество X. Во время предсказания w(t) на основе w(j) каждый нейрон из X производит скалярное умножение вектора своих синоптических весов на распределенный вектор слова w(j), к результату применяется логистическая функция. В итоге имеем:

$$P\left(w(t)\middle|w(j)\right) = \prod_{i} |w(t_i) - w^*(t_i)|$$

После чего производится коррекция векторов слов и весов нейронов. Скорость коррекции синоптических весов (скорость обучения) в процессе обучения уменьшается до нуля. Начальные значения синоптических весов первого слоя сети выбираются случайно.

Для skip-gram функция ошибки имеет вид:

$$E = -\sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} \log p(w(t+j)|w(t))$$

Мы провели экспериментальную оценку моделей векторного представления текстовой информации на трех языках – русском, татарском и английском. В качестве текстовых данных использовались: корпус английского языка [3], корпус татарского языка «Туган тел»

и корпус составленный на основе произведений русской классической литературы.

Во всех экспериментах использовался симметричный контекст с шириной окна 10 слов. Для CBOW и skip-gram были обучены векторные представления размерностью 50, 100, 200, 300, 500 элементов. Для нахождения ближайших слов (most similar) использовались слова «красивый», « матур», «beautiful». Для поиска аналогий (задачи типа «мужчина относится к женщина как король относится к X) был применены следующие: писать - письмо= X-хлеб, китап-укый =X-ашый, king-man=X- woman.

Таблица 1

Модель	Слово на входе			Размер вектора	Язык		
	русский	татарский	английский		Русский	Татарский	Английский
CBOW – ближай-шие слова	красивый	матур	beautiful	50	высокий 0.959333598614	гүзәл 0.785982489586	handsome 0.753469109535
				100	высокий 0.934601545334	гүзәл 0.706184208393	beauty 0.698206484318
				200	высокий 0.921762108803	гүзәл 0.696669995785	lovely 0.699172496796
				300	высокий 0.922815859318	гүзәл 0.672835350037	beauty 0.661501228809
				500	высокий 0.918797075748	гүзәл 0.671297669411	lovely 0.694669961929
CBOW-поиск аналогий	писать-письмо = X-хлеб	китап-укый = X-ашый	king-man= X- woman	50	покупать 0.726285398006	ипи 0.752121031284	queen 0.724276483059
				100	покупать 0.706483125687	йомырка 0.675853133202	queen 0.70689278841
				200	покупать 0.666280150414	йомырка 0.658689677715	queen 0.669976890087
				300	покупать 0.682923197746	йомырка 0.654567122459	queen 0.63525223732
				500	покупать 0.671586811543	йомырка 0.651859819889	queen 0.682458817959

В табл.1 перечислены примеры нахождения семантических и синтаксических аналогий для русского, татарского и английского языков. При обучении использовались слова, представленные в нижнем регистре.

Задача векторного представления слов русского и татарского языков на сегодняшний день исследована недостаточно. Причина в том, что в открытом доступе отсутствуют проработанные корпусы русско- и татароязычных текстов. Анализируя причины таких низких показателей, следует обратить внимание и на особенности русского и татарского языков. В частности, сложность, связанную с наличием отдельных падежей и родов.

Так, например, в английском языке слово «good», может быть и прилагательным, и наречием, мужского и женского рода, единственного

и множественного числа. В русском же языке большое количество словоупотреблений каждой отдельной формы слова согласно лицу, числу и падежу: «хороший», «хорошая», «хорошим», «хорошему», «хороших» и т.д. Данный факт подчеркивает значительность размера корпуса для языка. Ожидаемо, что при наличии трех корпусов данных с одинаковым количеством словоупотреблений, показатели синтаксической точности для русского языка будут ниже ввиду обилия различных словоформ.

Ввиду всего вышесказанного можно вывести основные недостатки технологии word2vec:

- ограниченность словаря не позволяет получить вектора для всех текстовых запросов;
- семантическая близость присутствует лишь для слов. При их суммировании для получения представления текста схожесть может теряться, причем, чем больше слов суммируется, тем заметнее этот эффект;
- в силу краткости запросов невозможно использовать окно достаточной ширины.

Создатель технологии word2vec – Томас Миколов считает, что модель CBOW больше подходит для работы с большими массивами текстовой информации (миллиарды, гигабайты информации). Skip-gram же больше подходит для относительно небольших текстов – не больше ста миллионов слов. И по быстроте обучения модель CBOW в сравнении с skip-gram более эффективна. Несмотря на все недостатки, в данный момент технология word2vec является самым точным подходом к векторному представлению слов естественного языка.

Благодарности. Авторы выражают благодарность НИИ «Прикладная семиотика» АН РТ за предоставление корпуса татарского языка «Туган тел».

Литература

1. Mikolov T., Corrado, G., Chen K., Dean J.: Efficient Estimation of Word Representations in Vector Space// arXiv:1301.3781v3, - 2013.

2. Biemann C., Handschuh S., Freitas A., Meziane F. Metais E.: Natural Language Processing and Information Systems. – 2015. – Springer, 453 c.

3. http://mattmahoney.net/dc/text8.zip

УДК 004.91

СЕРВИС ИНТЕГРАЦИИ НОВОСТНЫХ ЛЕНТ НА ПЛАТФОРМЕ УПРАВЛЕНИЯ ЭЛЕКТРОННЫМИ НАУЧНЫМИ ЖУРНАЛАМИ

А.Н. Герасимов
Казанский (Приволжский) федеральный университет
gerasimov.mailstore@gmail.com

На основе рекомендательных систем предложен метод анализа содержимого новостных лент научных порталов и формирования персонифицированного набора новостей. Реализованы алгоритмы согласования различных форматов новостных лент и фильтрации новостей на основе персонального профиля пользователя. Алгоритм консолидации научно-новостного контента реализован как отдельный модуль платформы OJS.

Ключевые слова: *электронный научный журнал, интеграция электрон-ных ресурсов, новостные ленты.*

В настоящее время новостные ленты являются важным инструментом коммуникации, который активно применяется различными информационными системами, в том числе ориентированными на научное сообщество, для доставки пользователям часто обновляемой информации через Веб. Примерами использования новостных лент являются: порталы научных журналов (например, Math-Net.Ru – http://www.mathnet.ru/rss/rssRecentIssues.phtml?option_lang=rus), издательские платформы (Elsevier и другие – http://www.elsevierscience.ru/rss/), научные объединения и сообщества (см., например, портал Европейского математического общества (European Mathematical Society) www.euro-math-soc.eu/rss-feeds), научные учреждений и университеты (Российская академия наук и ее институты – http://www.ras.ru/news/newslist/newschannellist.aspx; российские национальные, федеральные и национальные исследовательские университеты, см., например, http://www.msu.ru/news/rss/, http://kpfu.ru/news).

Для организации новостных лент могут использоваться различные технологии, такие, например, как RSS (Really Simple Syndication), Atom, OPML (Outline Processor Markup Language) или ставший популярным метод Broadcasting, ориентированный на увеличение пропускной способности интернет-каналов (см. [1]). Наиболее часто применяемой технологией организации новостных лент является RSS, при применении которой контент представляется в привычном html-формате. В настоящее

время используется несколько спецификаций RSS-ленты: Rich Site Summary (RSS 0.9x); RDF Site Summary (RSS 0.9 и 1.0); Really Simple Syndication (RSS 2.x).

Важной задачей при интеграции новостных лент являются согласование различных форматов представления информации и ее агрегация в базу данных. Именно на этом этапе производится семантический анализ получаемой информации, что, в частности, позволяет обеспечить фильтрацию информации, например, отсев ненаучного содержимого. Нами [2] реализован алгоритм консолидации научно-новостного контента в виде сервиса на платформе Open Journal Systems (см. [3]) с использованием языка программирования PHP и облачных технологий (см. [4, 5]). Этот сервис формирует новостную ленту с использованием рекомендательной системы [6] и на основе персонального профиля пользователя, учитывающего тематику проводимых исследований и историю прочитанных публикаций и новостей. Схема функционирования этого сервиса представлена на рисунке ниже.

На первом этапе осуществляется семантический анализ новостных потоков из различных научных порталов. Далее собранная информация приводится к единому формату, размещается в хранилище новостного контента, и с помощью рекомендательной системы формируется персонифицированная новостная лента.

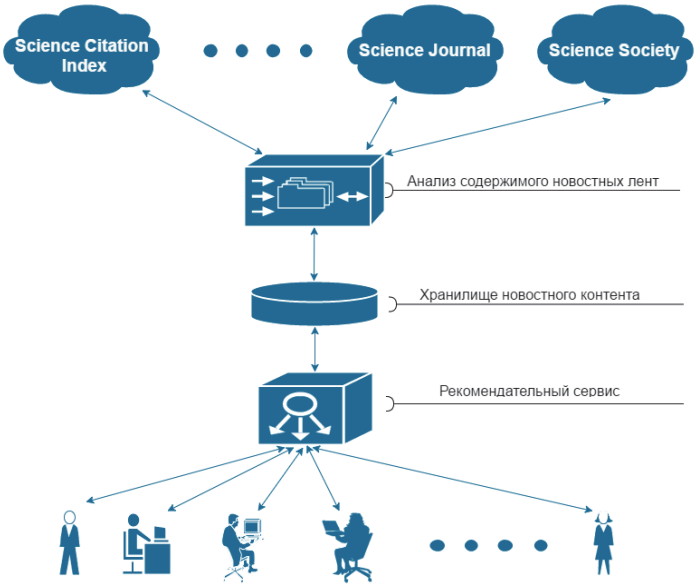


Рис. Схема функционирования рекомендательного сервиса на платформе Open Journal Systems

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проекты №№ 15-07-08522, 15-47-02472).

Литература

1. John K M. DotNetNuke 5.4 Cookbook. Gardners Books, 2010. 432 p.

2. Ахметов Д.Ю., Грачев А.О., Герасимов А.Н., Елизаров А.М., Липачёв Е.К. Облачная платформа поддержки электронных научных изданий // Учёные записки Института социальных и гуманитарных знаний. 2014. № 1 (12), ч. 1. С. 13–19.

3. Stranack K. Getting found, staying found, increasing impact. Enhancing readership and preserving content for OJS journals // Public Knowledge Project. 2006. 40 p.

4. Суэринг С., Конверс Т., Парк Д. PHP и MySQL. Библия программиста. М.: Изд-во «Диалектика», 2010. 912 с.

5. Buyya R., Broberg J., Goscinski A. Cloud computing: principles and paradigms. John Wiley & Sons Inc., 2011. 674 p.

6. Елизаров А.М., Жижченко А.Б., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К. Онтологии математического знания и рекомендательная система для коллекций физико-математических документов// Докл. РАН. 2016. Т. 467, №4. С. 392–395.

УДК 004.91

СЕМАНТИЧЕСКИЙ АНАЛИЗ БОЛЬШИХ КОЛЛЕКЦИЙ НАУЧНЫХ ДОКУМЕНТОВ

А.М. Елизаров[1], Е.К. Липачёв[2], Ш.М. Хайдаров[3]
Казанский (Приволжский) федеральный университет
1 – amelizarov@gmail.com, 2 – elipachev@gmail.com,
3 – 15jkeee@gmail.com

Предложен метод автоматической обработки больших коллекций физико-математических документов, хранящихся в формате OpenXML, включающий валидацию документов и их преобразование в соответствии с правилами формирования коллекций, семантический анализ документов, извлечение метаданных и др. Описан алгоритм метода, приведен пример успешной его реализации при организации XI Всероссийского съезда по фундаментальным проблемам теоретической и прикладной механики (Казань, 20 – 24 августа 2015 г.).

Ключевые слова: *Big Data, семантический анализ документов, структурный анализ текстов, метаданные, сервисы автоматической обработки больших коллекций*

Как известно (см., например, [1, 2]), большинство современных электронных коллекций научных документов (научные журналы,

сборники научных трудов, диссертации, научные отчеты, архивы и др.) представляет собой наборы неструктурированных документов, на базе которых трудно организовать семантический поиск, извлечение метаинформации и различные информационные сервисы. Кроме того, в настоящее время наблюдается значительное увеличение объема данных, включаемых в коллекции, что в свою очередь создает дополнительные трудности при обработке информации. Поэтому в условиях непрерывного роста объёмов, а также многообразия информации сейчас активно развиваются новые подходы, инструменты и методы обработки огромных объёмов данных, обозначаемых термином «большие данные» (Big Data). При управлении электронными научными коллекциями больших данных в полной мере остаются актуальными, а также появляются новые задачи, в их числе: семантическая разметка, организация поиска, выделение метаданных, формирование тематических кластеров документов, сбор наукометрической информации, подготовка сборников материалов и др. Насущными становятся проблемы анализа и управления данными в различных областях с интенсивным использованием данных (см., например, материалы конференции [3]).

К большим массивам научных документов сегодня можно отнести и материалы, поступающие на конференции. Их ручная обработка чаще всего не эффективна или даже невозможна. Именно такая ситуация возникла при подготовке проведения XI Всероссийского съезда по фундаментальным проблемам теоретической и прикладной механики (Казань, 20–24 августа 2015 г.). В частности, при подготовке к печати материалов Съезда потребовалось решить задачу автоматизированной подготовки метаданных этих публикаций (в соответствии с правилами баз научного цитирования) общим объемом более 1500 статей в формате .docx. Естественно, что традиционными методами оперативно выполнить эту работу было невозможно. Основной задачей, решенной при формировании коллекции материалов Съезда, было приведение поступивших материалов к единому стилевому оформлению:

- единообразное представление названий статей и списка авторов докладов, структура аффилиации авторов, формат аннотации;
- приведение списков литературы к выбранному формату библиографического описания;
- единообразное шрифтовое оформление разделов текста статей;
- выбор форматов рисунков, схем, диаграмм;
- набор математических формул и системы ссылок на них;
- оформление ссылок на поддержку исследований грантами, благодарности.

Основным технологическим инструментом решения названных задач был структурный анализ рассматриваемых документов, проведенный с использованием техники регулярных выражений, а также различных эвристических методов: информация, извлекаемая из документа, содержит название статьи, список авторов с выделением для каждого аффилиации и адреса электронной почты, аннотацию и благодарности, список ключевых слов, основные разделы статьи, библиографический список. Результаты структурного анализа позволили сформировать семантическое представление формируемой коллекции. Опишем подробнее конкретные шаги проведенного структурного анализа.

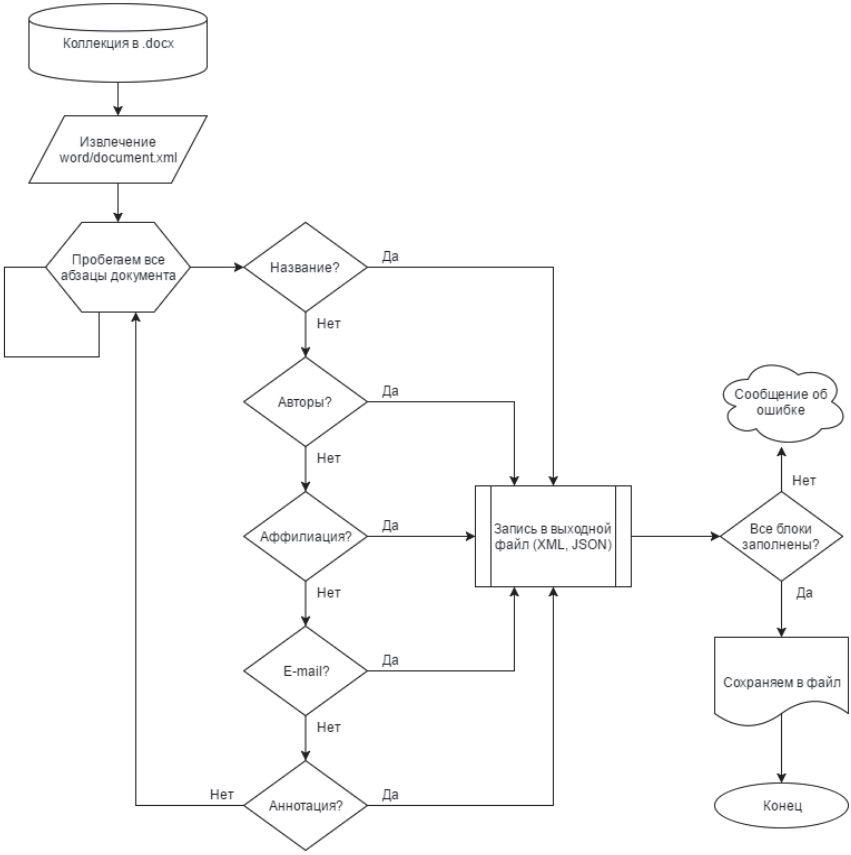


Рис. Алгоритм структурного анализа коллекции

Все материалы, поступившие на Съезд, были преобразованы в формат .docx, который основан на языке разметки XML и допускает семантическую обработку документов (см. [4]). Затем методом, предложенным в [5, 6], из каждого документа извлекался соответствующий ему файл «word/document.xml», который содержит информацию о шрифтовом оформлении и расположении основных блоков документа (название работы, перечень авторов, их аффилиация и др.). Для выделения метаданных использовалась техника регулярных выражений. Например, для выделения списка авторов статей использовались регулярное выражение

```
$fio="/([А-ЯA-Z]\.(?:[А-ЯA-Z]\.)?\s[А-ЯA-Z][a-za-я]+)(,\s)?(?1)?(,\s)?(?1)?/u";
```

и соответствующий скрипт

```
while(($w_ps->item($k+1)->nodeValue)!=""){
if(preg_match($fio,$w_ps->item($k+1)->nodeValue))break;
$articlename.=$w_ps->item($k+1)->nodeValue."%%";
$k++; }
```

Подчеркнем, что существенным условием применения описанного метода является единообразное стилевое оформление документов, что позволяет проводить структурный анализ документов коллекции в автоматическом режиме. Общий алгоритм метода представлен на рисунке выше.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проекты №№ 15-07-08522, 15-47-02472).

Литература

1. Афонин С.А., Бахтин А.В., Бухонов В.Ю., Васенин В.А., Ганкин Г.М., Гаспарянц А.Э., Голомазов Д.Д., Иткес А.А., Козицын А.С., Тумайкин И.Н., Шапченко К.А. Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА). Под ред. акад. В.А. Садовничего. М.: Изд-во Московского университета, 2014. 262 с.

2. Елизаров А.М., Липачев Е.К., Хохлов Ю.Е. Семантические методы структурирования математического контента, обеспечивающие расширенную поисковую функциональность // Информационное общество. 2013. № 1–2. С. 83–92.

3. Аналитика и управление данными в областях с интенсивным использованием данных: XVII Международная конференция DAMDID/RCDL’2015 (Обнинск, 13–16 октября

2015 года, Россия): Труды конференции/ под ред. Л.А. Калиниченко, С.О. Старкова. – Обнинск: ИАЕЭ НИЯУ МИФИ, 2015. 525 с.

4. Standard ECMA-376: Office Open XML File Formats. URL: http://www.ecmainternational.org/publications/ standards/Ecma-376.htm

5. Хайдаров Ш.М. Методы управления математическим контентом в информационных издательских системах // Тр. Матем. центра им. Н.И. Лобачевского. Материалы 14-й Всерос. Молодежной школы-конференции «Лобачевские чтения–2015 (Казань, 22–27 октября 2015 года). Казань. 2015. С. 162–165.

6. Хайдаров Ш.М. Семантический анализ документов в системе управления цифровыми научными коллекциями // Электронные библиотеки. 2015. Т. 18. № 1–2. С. 61–85.

УДК 004.91

ЭКОСИСТЕМА ONTOMATH И ПРОЕКТ ВСЕМИРНОЙ ЦИФРОВОЙ МАТЕМАТИЧЕСКОЙ БИБЛИОТЕКИ

А.М. Елизаров[1], Н.Г. Жильцов[2], А.В. Кириллович[3], Е.К. Липачёв[4], О.А. Невзорова[5],

Казанский (Приволжский) федеральный университет

1 – amelizarov@gmail.com,

2 – nikita.zhiltsov@gmail.com, 3 – alik.kirillovich@gmail.com,

4 – elipachev@gmail.com, 5 – onevzoro@gmail.com

Описаны возможности использования при проведении новых исследований всего корпуса накопленных научных знаний. Такое использование предполагает повсеместное внедрение информационно-коммуникационных технологий (ИКТ), обеспечивающих оптимальное управление имеющимися знаниями, организацию эффективного доступа к ним, а также совместное и многократное использование новых видов структур знаний. Наибольший эффект от внедрения современных ИКТ для дальнейшей организации научных знаний и повышения их понятности можно ожидать в области математики. Эти ожидания в полной мере подтверждены проектом создания Всемирной цифровой математической библиотеки (World Digital Mathematical Library – WDML). Представлены основные направления реализации проекта WDML и результаты по созданию экосистемы OntoMath как его составной части.

Ключевые слова: *WDML, Всемирная цифровая математическая библиотека, экосистема OntoMath, онтологии, семантический поиск*

В настоящее время благодаря повсеместному внедрению информационно-коммуникационных технологий (ИКТ) в научно-исследовательскую деятельность стало возможным при проведении новых

исследований использовать весь корпус накопленных научных знаний. Такое использование предполагает создание комплекса технологий, обеспечивающих оптимальное управление имеющимися знаниями, организацию эффективного доступа к ним, а также совместное и многократное использование новых видов структур знаний. По-видимому, наибольший эффект от внедрения современных ИКТ для дальнейшей организации научных знаний и повышения их понятности можно ожидать в области математики, которая характеризуется наибольшей по сравнению с другими науками формализацией результатов. Эти ожидания в полной мере подтверждены проектом создания Всемирной цифровой математической библиотеки (World Digital Mathematical Library – WDML). Термин WDML введен в 2006 году на Генеральной ассамблее международного математического союза (см. [1]).

Назначение WDML – объединить в распределенной системе взаимосвязанных хранилищ оцифрованные версии всего корпуса математической научной литературы, включая как современные источники, так и ставшие уже историческими (см. [2, 3]). Основные задачи построения WDML и технологии, необходимые для их решения, описаны в [4]. В частности, проект предполагает, что следующим шагом в продвижении математики будут выход за пределы традиционных математических публикаций и построение сети информации, основанной на знаниях, содержащихся в этих публикациях. Благодаря сочетанию методов машинного обучения и усилий редакций и редколлегий математических научных журналов, значительная часть информации и знаний (как связанных открытых данных) в глобальном математическом корпусе станет доступной для исследователей через WDML. Частью проекта WDML, связанной с семантическим представлением математического знания, стал симпозиум, прошедший в 2016 году в Филдсовском институте, г. Торонто [5]. В работе симпозиума приняло участие 37 приглашенных экспертов. В нашем докладе [6] на симпозиуме представлены технологии управления математическими знаниями на основе онтологий (см. также [7–12]). Результатом исследований, описанных в этих работах, стало создание основ цифровой экосистемы OntoMath, основные положения которой полностью коррелируют с идеологией проекта WDML.

OntoMath – это экосистема онтологий, инструментов текстовой аналитики и приложений для управления математическими знаниями. Кратко опишем ее основные элементы.

Платформа семантической публикации – это центральный элемент экосистемы (см. [7]). Эта платформа принимает на вход коллекцию математических публикаций в формате LaTeX и строит их

семантическое представление, интегрированное в облако Linked Open Data. Семантическое представление публикаций включает:

- метаданные: названия, даты, авторы, аффилиации, наименования журналов и т.д.; метаданные описаны с использованием онтологии AKT Portal;
- логическую структуру публикаций: раздел, теорема, доказательство, формула и т. д., которая описана с помощью онтологии Mocassin;
- терминологию, выраженную посредством онтологии OntoMathPro;
- формулы: переменные внутри них привязаны к понятиям, которые эти переменные обозначают.

Mocassin – это онтология логической структуры математических документов, предназначенная для автоматического анализа математических публикаций в формате LaTeX. Онтология описывает семантику структурных элементов математических документов (например, теоремы, леммы, доказательства, определения и т. д.) и связей между ними (см. [8]).

OntoMathPro – это онтология математического знания, которая организована в виде двух иерархий:

- иерархии областей математики: математическая логика, теория множеств, алгебра, геометрия, топология и т. д.;
- иерархии математических объектов: множество, функция, интеграл, элементарное событие, многочлен Лагранжа и т. д.).

OntoMathPro содержит пять типов отношений: Класс → Подкласс, Определяется с помощью, Ассоциативная связь, Задача → Метод решения и Область математики → Математический объект. Концепты онтологии содержат название концепта на русском и английском языках, определение, ссылки на внешние ресурсы из облака Linked Open Data и связи с другими концептами (см. [9–11]).

Семантический поисковик по математическим формулам: отыскивает формулы, содержащие переменную, обозначающую заданное математическое понятие (например, формулы, содержащие переменную, обозначающую угол, или формулы, связывающие давление и массу).

Рекомендательная система математических публикаций: для заданной публикации строит список «похожих» статей (см. [12]).

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проекты №№ 15-07-08522, 15-47-02472).

Литература

1. Digital Mathematics Library: a vision for the future. International Mathematical Union, 2006, URL: http://www.mathunion.org/fileadmin/IMU/Report/dml_vision.pdf.

2. Olver P.J. What's happening with the World Digital Mathematics Library? URL: http://www.math. umn.edu/~olver/t_/wdmlb.pdf.

3. Olver P.J. The World Digital Mathematics Library: report of a panel discussion // Proceedings of the International Congress of Mathematicians, August 13–21, 2014, Seoul, Korea. Kyung Moon SA, 2014. V. 1. P. 773–785.

4. Developing a 21st century global library for mathematics research. Washington, D.C.: The National Academies Press, 2014. 131 p.

5. Semantic representation of mathematical knowledge workshop, 5 February 2016. URL: https://www.fields.utoronto.ca/programs/scientific/15-16/semantic/

6. Elizarov A.M., Zhiltsov N.G., Kirillovich A.V., Lipachev E.K., Nevzorova O.A., Solovyev V.D. The OntoMath ecosystem: ontologies and applications for math knowledge management // Semantic Representation of Mathematical Knowledge Workshop 5 February 2016. URL: http://www.fields.utoronto.ca/video-archive/2016/02/2053-14698.

7. Nevzorova O., Zhiltsov N., Zaikin D., Zhibrik O., Kirillovich A., Nevzorov V., Birialtsev E. Bringing Math to LOD: a semantic publishing platform prototype for scientific collections in mathematics // 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21–25, 2013, Proceedings, Part I. Lecture Notes in Computer Science, Vol. 8218. Springer Berlin Heidelberg, 2013. P. 379–394.

8. Solovyev V., Zhiltsov N. Logical structure analysis of scientic publications in mathematics // In: Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS'11). ACM. 2011. P. 21:1–21:9.

9. Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E. OntoMathPro ontology: a linked data hub for mathe-matics // Communications in Computer and Information Science. 2014. V. 468. P. 105–119.

10. Elizarov A., Kirillovich A., Lipachev E., Nevzorova O., Solovyev V., Zhiltsov N. Mathematical knowledge representation: semantic models and formalisms // Lobachevskii Journal of Mathematics, 2014. V. 35, No 4. P. 347–353.

11. Elizarov A.M., Lipachev E.K., Nevzorova O.A., Solov'ev V.D. Methods and means for semantic structuring of electronic mathematical documents // Dokl. Math. 2014. V. 90, No 1. P. 521–524.

12. Елизаров А.М., Жижченко А.Б., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К. Онтологии математического знания и рекомендательная система для коллекций физико-математических документов// Докл. РАН. 2016. Т. 467, №4. С. 392–395.

УДК 004.622:004.822

ПОСТРОЕНИЕ МОДЕЛИ ПРЕДМЕТНОЙ ОБЛАСТИ И СЕТИ СОАВТОРСТВА В ОБЛАСТИ ЮРИСПРУДЕНЦИИ НА ОСНОВЕ ЗОНДИРОВАНИЯ СЕРВИСА GOOGLE SCHOLAR CITATIONS

Д.В. Ландэ, В.Б. Андрущенко
Институт проблем регистрации информации НАН Украины, г. Киев, Украина
dwlande@gmail.com, valentyna.andrushchenko@gmail.com

Предлагается методика построения сетей – моделей предметных областей и сетей соавторства на основе зондирования контентных сетей. В работе рассматриваются сети понятий, соответствующих тегам и авторам сервиса Google Scholar Citations. Модели построены для правовой науки, однако предложенный подход можно применять и для других областей.

Ключевые слова: *предметная область, сеть соавторства, правовая наука, зондирование сети, информационная сеть*

С развитием информационных ресурсов сети Интернет появились новые возможности описания предметных областей и изучения закономерностей научного взаимодействия. В то время, как для автоматизированного построения онтологий, моделей предметных областей, все чаще используются документальные корпусы [1], в том числе сетевые [2], основным инструментом изучения закономерностей научного сотрудничества являются сети соавторов, формируемые наукометрическими службами [3]. С помощью сетей соавторов можно получать не только наукометрические оценки, но и определять экспертов для решения сложных задач [3]. Одним из крупных сервисов научной информации является Google Scholar Citations, который позволяет ученым создавать их профили, среди прочего, содержащие библиографическую информацию, а также осуществлять поиск публикаций.

В этой работе представляется подход к созданию модели предметной области (юриспруденция) на основе зондирования большой информационной сети и построения сети понятий, которые отражаются в тегах наукометрического сервиса Google Scholar Citations (http://scholar.google.com/citations) [4], [5]. Интерфейс этого сервиса позволяет выводить списки авторов и приписываемых им тегов (понятий, концептов), соответствующих заданному первоначальному тегу (в нашем случае, law). В данной работе также предлагается алгоритм построения сетей

соавторства – моделей сотрудничества ученых на основе зондирования этой же наукометрической сети.

В поисковом интерфейсе, соответствующем заданному тегу (label: law) постранично в ранжированном виде отображаются имена ученых, которые отметили свою деятельность этим тегом, а также другие теги, приписанные ими. Множество тегов образуют сеть, производную от биграфа «ученый-теги». Эту сеть можно рассматривать как некоторую онтологическую модель предметной области. Узлы в этой сети соответствуют понятиям, маркированным тегами, а связи – семантическую связь между ними. С другой стороны, для каждого ученого, зарегистрированного в Google Scholar Citations, в интерфейсе сервиса приводится список его соавторов.

Для построения модели предметной области был предложен алгоритм [2] к реальной сети тегов сервиса Google Scholar Citations следующим образом:

1. Экспертным путем был определен небольшой перечень базовых тегов (law, justice, criminology), для каждого из которых выполняется следующая последовательность действий.
2. Выбирается тег из данного перечня.
3. Выполняется поиск и открывается страница веб-сервиса, соответствующие этому тегу.
4. К создаваемой сети добавляются все теги, содержащиеся на этой странице (соседние теги).
5. Далее из соседних тегов выбирается тот, на страницы которого планируется перейти для дальнейшего анализа. Этот тег с наибольшей степенью среди соседних тегов, который также удовлетворяет тематике выбранной предметной области (в нашем случае, содержит такие фрагменты слов, как _law, right, crimin, crime, just), к поисковой странице которого еще не был осуществлен переход.
6. Если такой тег выбран, то происходит переход к пункту 3.
7. Если такого тега не существует, но перечень базовых тегов не завершен, то осуществляется переход к следующему базовому тегу из начального перечня, т.е. переход к пункту 2. Иначе считается, что сеть зондирования построена.

На рис. 1 приведен пример сети понятий предметной области, построенной в соответствии с приведенным алгоритмом по указанным базовым тегам.

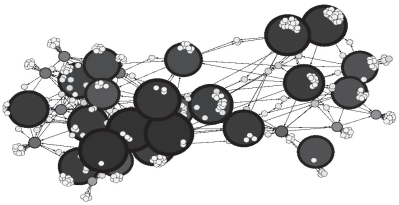


Рис. 1. Структура сети понятий по заданному тегу

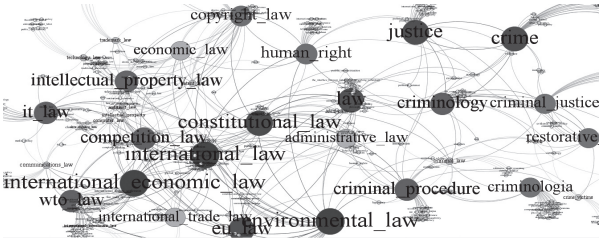


Рис. 2. Фрагмент сети понятий

Приведенный выше алгоритм был адаптирован для построения соавторов следующим образом [4]:

1. Экспертным путем определяется небольшой перечень базовых тегов, для каждого из которых выполняется следующее:
2. Открывается страница веб-сервиса, соответствующая выбранному тегу.
3. Выбирается самый цитируемый автор, представленный на данной странице.
4. К создаваемой сети добавляются все соавторы, содержащиеся на странице выбранного автора. Формируются ребра-связи к этим узлам (соавторам) из исходного узла (автора).
5. Из списка узлов формируемой сети выбирается тот, на страницу которого планируется перейти для дальнейшего анализа. Это самый весомый узел, удовлетворяющий тематике выбранной предметной области (его теги содержат фрагменты слов, выбранные экспертами) и не входит в состав тех узлов, к страницам которых уже был осуществлен переход.
6. Если такой узел-автор выбран, то происходит переход к пункту 4.
7. Если такого автора не существует, то считается, что сеть соавторства построена.

В соответствии с приведенным алгоритмом была построена сеть соавторов при заданном заранее ограничении на количество сканируемых узлов. С помощью программного средства Gephi была получена визуализация данной сети соавторов (рис. 3 и 4).

Применение методов кластерного анализа позволяют выявлять наиболее тесно связанные между собой группы ученых-соавторов, научных школ, экспертных групп.

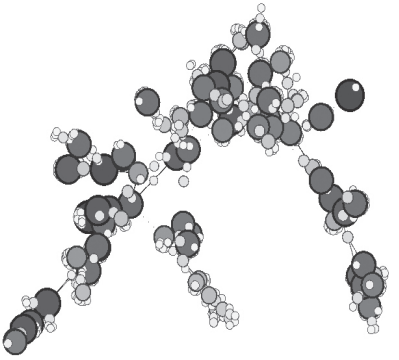


Рис. 3. Структура сети соавторства

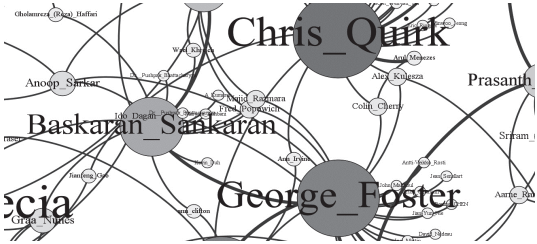


Рис. 4. Фрагмент сети соавторства

Таким образом, предложен и реализован подход к формированию модели предметной области и сетей соавторства в рамках этой предметной области, ограничительными элементами которого составляют некоторые маркеры знаний (теги), заранее заданные учеными – участниками проекта Google Scholar Citations. Следует отметить принципиальное отличие предложенной модели автоматического формирования модели предметной области от существующих, базирующихся на анализе текстовых корпусов или непосредственном участии экспертов при выборе конкретных узлов и связей. В данном

случае эксперт-пользователь вкладывает лишь крупицы знаний в виде набора базовых тегов и небольших по объему словарей (до десятка слов). В дальнейшем программа использует знания, заложенные самими авторами публикаций, теги отмеченные ими как главные.

Модели применены для отрасли правовой науки, однако предложенный подход можно использовать и для других областей научных знаний.

Литература

1. Добров Б.В., Соловьев В.Д., Лукашевич Н.В., Иванов В.В. Онтологии и тезаурусы. Модели, инструменты, приложения. Бином, 2009. – 173 с.

2. Lande D. A Domain Model Created on the Basis of Google Scholar Citations // CEUR Workshop Proceedings (ceur-ws.org). Vol-1536 urn:nbn:de:0074-1536-8. Selected Papers of the XVII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015) Obninsk, Russia, October 13-16, 2015. – pp. 57-61.

3. Liu J., Li Y., Ruan Z., Fu G., Chen X., Sadiq, Deng Y. A new method to construct co-author networks // Phisica A. – 2015. – 419. – pp. 29-39.

4. Ландэ Д.В., Балагура И.В., Андрущенко В.Б. Построение сетей соавторства по данным сервиса Google Scholar Citations: материалы VI междунар. науч.-техн. конф. [“Открытые семантические технологии проектирования интеллектуальных систем” (OSTIS-2016)], (Минск, 18-20 февраля 2016 года). – Минск: БГУИР, 2016. – С. 233-237.

5. Brezina V. Use of Google Scholar in corpus-driven EAP research // Journal of English for Academic Purposes.– 2012. – 11. – P. 319-331.

УДК 004.9:510

ИСПОЛЬЗОВАНИЕ СИСТЕМ СВЯЗЫВАНИЯ ДАННЫХ ДЛЯ УСТАНОВЛЕНИЯ СООТВЕТСТВИЙ МЕЖДУ ХРАНИЛИЩАМИ БИБЛИОГРАФИЧЕСКИХ ДАННЫХ

К.С. Николаев, О.А. Невзорова
Казанский федеральный университет, Казань
Институт прикладной семиотики АН РТ, Казань
konnikolaeff@yandex.ru, onevzoro@gmail.com

В статье описываются технология представления реляционных данных в формате RDF, программные решения для связывания данных, используемые в проекте Linked Open Data и эксперименты по связыванию библиографических данных.

Ключевые слова: *связывание данных, RDF, Linked Open Data.*

Введение

Проект Linked Open Data – одна из самых заметных по результатам реализаций принципов Linked Data [1, 2]. В последние годы семантические технологии, используемые в Linked Data (связанные данные), используются поисковыми системами, web-сайтами и т.д. Методы, используемые в Linked Data, позволяют удобным способом выражать, распространять и связывать данные на основе стандартов URIs, RDF, XML. Модель RDF – это принятый W3C формат данных, используемый для представления отдельных сущностей, представляющих собой самостоятельную единицу знаний. Сущности связаны друг с другом с помощью предикатов по шаблону «субъект – предикат - объект». Данные для открытых хранилищ в основном извлекаются из реляционных баз данных или в частичном виде из веб-страниц и текстовых документов. Ссылки между сущностями образуют глобальный граф данных, используемый поисковыми роботами и браузерами для перемещения по источникам данных.

В статье рассматриваются технологии связывания данных из реляционной базы данных, содержащей в себе библиографические описания публикаций из разных источников, и хранилища библиографических данных DBLP Computer Science Bibliography [3], для поиска соответствий между этими источниками данных.

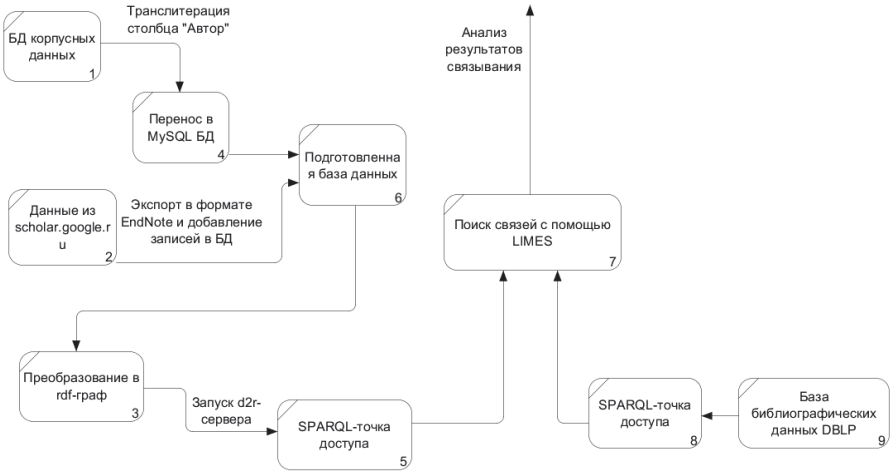


Рис. 1. Схема подготовки и связывания данных

1. Схема преобразования и связывания источников данных

Для выполнения поставленной задачи необходимо организовать связку источника локальных данных (база данных MySQL), преобразовать ее в RDF-хранилище и произвести связывание данных с помощью программной системы LIMES. На рис. 1 приведена схема выполнения задачи.

На первом этапе выполняется подготовка локальной базы данных, в нашем случае, базы библиографических описаний корпусных данных corpus.antat.ru, дополненной мета-описаниями открытых публикаций по компьютерным наукам из Google Scholar.

Подготовка заключается в корректировке структуры базы данных путем разбиения строки с библиографическим описанием на отдельные составляющие (Город, Издательство, Год издания и Количество страниц). В Таблице 1 приведены самые значимые поля преобразованной базы

данных и примеры их значений (в исходной базе данных отсутствовали поля, начинающиеся с «Разбиение»).

Таблица 1

Структура экспериментальной БД

№	Имя_ файла	Название	Автор	Объем_ в_словах	Разбиение_ Количество_ страниц	`Разбиение_ Основное_ описание`	`Разбиение_ Год`	`Разбиение_ Город`
139	maturlik_ tat.txt	"Матурлык"	'Amirkhan Eniki'	630	247	ИЯЛИ АН РТ	2010	Казан

Количество записей в дополненной БД составляет 7646.

Далее производится конвертация данных с помощью программы ESF Database Migration Toolkit [4]. Данный программный продукт переносит все записи из исходной БД в целевую с сохранением связей между таблицами. Программа поддерживает двустороннюю конвертацию между базами данных следующих типов: MySQL, PostgreSQL, Oracle, SQL Server, IBM DB2, Informix, Microsoft Access, Microsoft Excel, Paradox и другие. В проведенных экспериментах осуществлялся перенос всех записей из исходного файла (korpus.accdb) в MySQL базу данных. На рис. 2 указаны параметры, которые позволяют выполнить указанный перенос записей.

Следующим этапом является создание d2r-сервера. Данный d2r-сервер нужен для того, чтобы автоматизировать доступ к переведенным в RDF-формат данным через локальную SPARQL-точку доступа.

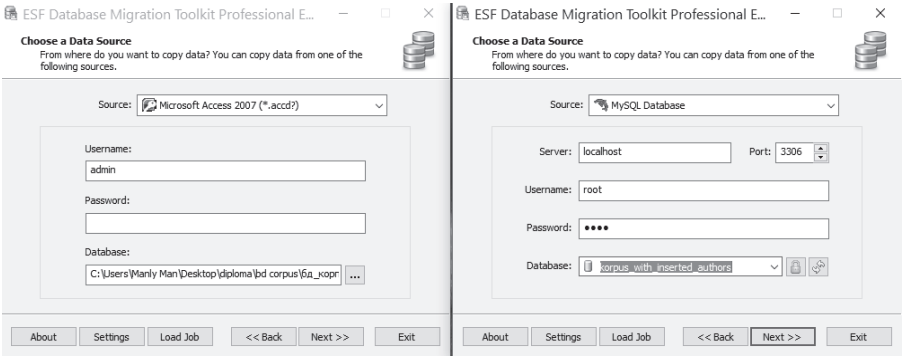


Рис. 2. Настройка ESF Database Migration Toolkit

1.1. Система преобразования реляционных БД в RDF-формат D2RQ Platform

Система преобразования реляционных БД в RDF-формат D2RQ Platform используется для настройки и реализации доступа к реляционным базам данных, который осуществляется путем представления реляционной базы данных в виде RDF-графа с возможностью создания дампа RDF-данных.

Запуск D2RQ Platform происходит через командную строку из папки с исходными файлами командой следующего вида: «generate-mapping -o <Название файла мэппинга>.ttl -d org.gjt.mm.mysql.Driver -u root -p password jdbc:mysql://localhost:3306/<Название локальной MySQL БД>». На рис. 3 приведен шаблон запуска скрипта generate-mapping.bat. В параметрах запуска программе необходимо указать название файла, в который будет записан способ представления необходимой базы данных (мэппинг) (-o <Название файла мэппинг >.ttl), указать драйвер (-d org.gjt.mm.mysql.Driver). С помощью параметров –u и –p задаётся пара логин-пароль для доступа к базе данных, указанной далее (jdbc:mysql://localhost:3306/<Название локальной MySQL БД>).

Данная команда создает файл мэппинга, анализируя схему указанной базы данных. Каждая таблица представляется в виде нового RDFS класса, одноименного с этой таблицей. Каждый столбец переводится в одноименное свойство [5].

С помощью файла мэппинга мы можем выгрузить данные из исходной базы данных в rdf-файл.

Этот шаг выполняется с помощью следующей команды: dump-rdf -f RDF/XML -b http://localhost:2020/ <Название файла мэппинга>.ttl > <Название выходного RDF-файла>.nt

После разметки базы данных нужно запустить сервер с настройками, указанными в файле с мэппингом. Это происходит при запуске команды «d2r-server mapkorpus.ttl» через командную строку.

Через точку доступа https://localhost:2020/sparql можно получить доступ к данным в полученном RDF-хранилище. Именно через эту точку доступа мы будем получать доступ к данным в системе LIMES.

1.2. Система связывания источников данных LIMES

Система LIMES – система поиска связей в Web of Data. Она применяет эффективные по времени алгоритмы для поиска связей в крупных хранилищах данных с использованием аксиомы треугольника в метрических пространствах.

Запуск системы LIMES происходит с помощью jar-программы из архива, скачанного с репозитория программы [6]. Перед запуском исполняемого файла необходимо произвести настройку конфигурационного файла (config.xml), и передать его в параметры при запуске. В config.xml хранятся параметры связывания для LIMES. На рис. 4 показана схема работы LIMES, из которой видно, что LIMES требует на вход информацию об источнике данных, целевых данных, параметр, по которому будет производиться связывание и метод сравнения.

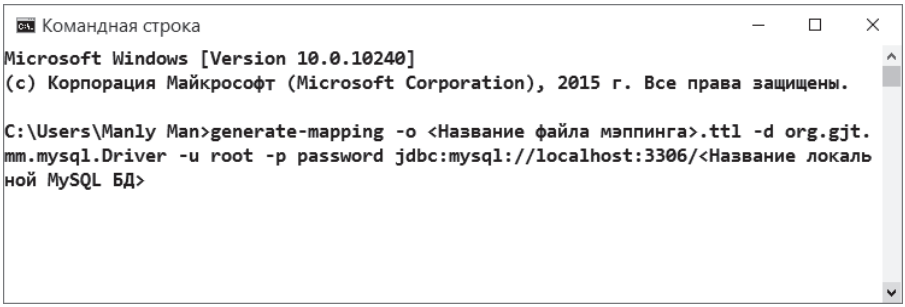


Рис. 3. Шаблон запуска скрипта generate-mapping.bat

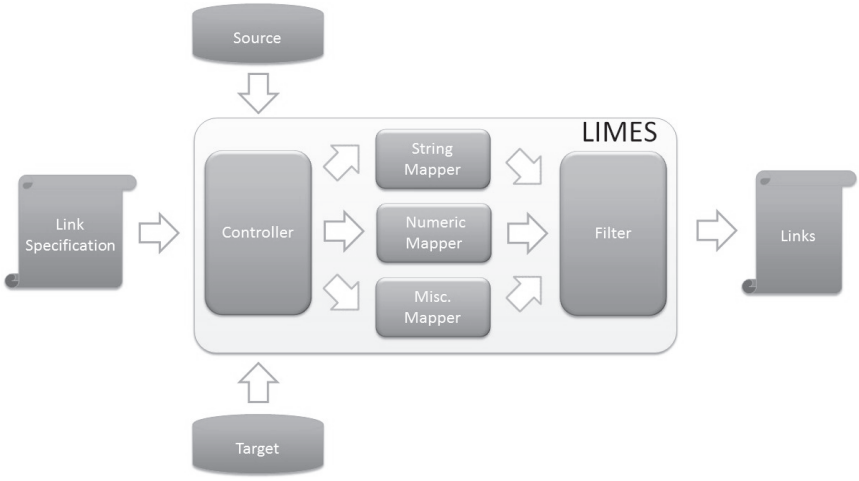


Рис. 4. Схема работы LIMES (из официального руководства)

2. Эксперименты

Для выполнения поиска связей между двумя источниками данных будем использовать набор библиографических описаний корпусных данных. В эксперименте для связывания будет использоваться хранилище библиографических данных DBLP Computer Science Bibliography, записи в которых имеют структуру, отображенную в Таблице 2.

Таблица 2

Структура онтологии DBLP

dc:identifier	DBLP journals/acta/AczelE97 (xsd:string)
dc:identifier	DOI 10.1007%2Fs002360050100 (xsd:string)
dcterms:issued	1997 (xsd:gYear)
swrc:journal	<http://dblp.l3s.de/d2r/resource/journals/acta>
rdfs:label	A New Formula for Speedup and its Characterization. (xsd:string)
foaf:maker	<http://dblp.l3s.de/d2r/resource/authors/J%C3%A1nos_Acz%C3%A9l>
foaf:maker	<http://dblp.l3s.de/d2r/resource/authors/Wolfgang_Ertel>
swrc:number	8 (xsd:string)
swrc:pages	637-652 (xsd:string)
dc:title	A New Formula for Speedup and its Characterization. (xsd:string)
dc:type	<http://purl.org/dc/dcmitype/Text>
rdf:type	swrc:Article
rdf:type	foaf:Document

Опишем схему связывания источников данных по шагам.

1. Мэппинг локальной базы данных производится следующей командой:

«generate-mapping -o mapkorpus.ttl -d org.gjt.mm.mysql.Driver -u root -p 6480 jdbc:mysql://localhost:3306/korpus_with_inserted_authors»

2. Запуск d2r-сервера производится с помощью команды «d2r-server mapkorpus.ttl». Сервер предоставляет доступ к RDF-хранилищу, краткая схема и соотношение с исходной БД которого показано в Таблице 3.

Таблица 3

Сравнение структур реляционной БД и RDF-хранилища

joinedandauthors	map:joinedandauthors a d2rq:ClassMap;
№	<#joinedandauthors_№> a d2rq:PropertyBridge;
Имя_файла	<#joinedandauthors_Имя_файла> a d2rq:PropertyBridge;
Название	<#joinedandauthors_Название> a d2rq:PropertyBridge;
Автор	<#joinedandauthors_Автор> a d2rq:PropertyBridge;
Объем_в_словах	<#joinedandauthors_Объем_в_словах> a d2rq:PropertyBridge;
Разбиение_ Количество_ страниц	<#joinedandauthors_ Разбиение_Количество_страниц> a d2rq:PropertyBridge;
Разбиение_ Основное_описание	<#joinedandauthors_ Разбиение_Основное_описание> a d2rq:PropertyBridge;
Разбиение_Год	<#joinedandauthors_ Разбиение_ Год> a d2rq:PropertyBridge;
Разбиение_Город	<#joinedandauthors_ Разбиение_Город> a d2rq:PropertyBridge;

В таблице 4 приводятся параметры системы LIMES, отвечающие за настройку исходного (Source) и целевого (Target) набора данных.

Таблица 4

Входные параметры LIMES

Source		Target	
Строка	Описание	Строка	Описание
<ID>local</ID>	Идентификатор источника	<ID>dblp</ID>	Идентификатор цели
<ENDPOINT>http://localhost:2020/sparql</ENDPOINT>	Точка доступа к источнику	<ENDPOINT> http://dblp.l3s.de/d2r/sparql </ENDPOINT>	Точка доступа к цели
<VAR>?x</VAR>	Объявление переменной для SPARQL-запроса	<VAR>?y</VAR>	Объявление переменной для SPARQL-запроса
<PAGESIZE>2000</PAGESIZE>	Ограничение на количество объектов в одном запросе	<PAGESIZE>2000</PAGESIZE>	Ограничение на количество объектов в одном запросе
<PROPERTY>vocab:books_author</PROPERTY>	Ограничение, по которому выбираются объекты	<PROPERTY>foaf:maker/rdfs:label AS cleaniri RENAME name</PROPERTY>	Ограничение, по которому выбираются объекты
<PROPERTY>vocab:joinedandauthors_Автор AS nolang->lowercase</PROPERTY>	Свойство, по которому сравниваются объекты из двух источников	<RESTRICTION>?y rdf:type foaf:Document</RESTRICTION>	Свойство, по которому сравниваются объекты из двух источников.
trigrams(x.vocab:joinedandauthors_Автор,y.name) – алгоритм поиска связей.			

Как можно понять из фрагмента файла, приведенного выше, нахождение соответствий будет происходить по следующей схеме:

1) В локальной БД - по полю vocab:joinedandauthors_Автор, которое указывает имя автора в сущностях типа vocab:joinedandauthors;

2) В DBLP хранилище – по полю rdfs:label, хранящимся в сущности типа foaf:maker из из foaf:Document.

При связывании мы используем онтологию FOAF (**F**riend **O**f **A** **F**riend), основанную Либби Миллером и Дэном Брикли для представления данных о личности в удобном формате [7].

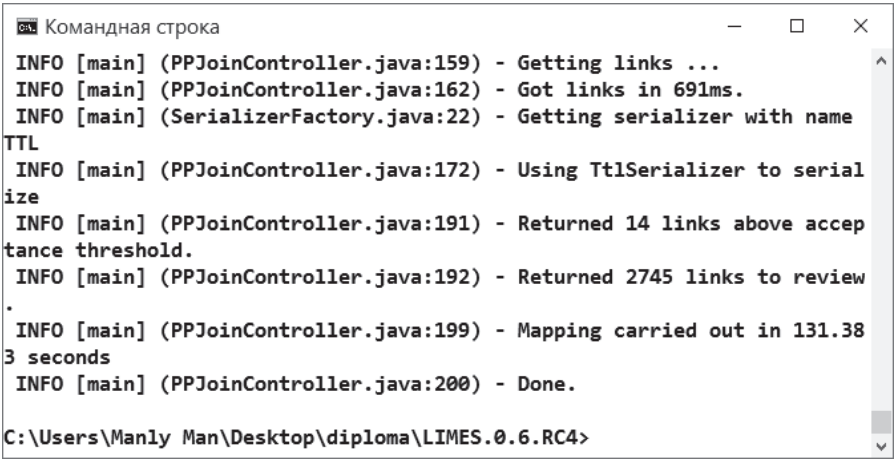


Рис. 5. Результат работы системы LIMES

В результате связывания вышеупомянутых источников данных было выявлено 14 точных совпадений (Threshold = 1) и 2745 предположительных (Threshold = 0.5). На Рисунке 5 показан результат работы программы.

В Таблице 5 показан вывод консольной версии системы LIMES.

Таблица 5

Результаты связывания

Запись из локальной БД	Тип соответствия	Запись из DBLP Computer Science Bibliography
<http://localhost:2020/resource/joinedandauthors/7894>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/abi/DivoliNH12>
<http://localhost:2020/resource/joinedandauthors/7909>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/aamas/GlassG03>
<http://localhost:2020/resource/joinedandauthors/7909>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/aamas/SarneG13>

<http://localhost:2020 /resource/joinedandauthors /7909>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/aamas/ElmalechSG15>
<http://localhost:2020 /resource/joinedandauthors /7909>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/aamas/OrtizG02>
<http://localhost:2020 /resource/joinedandauthors /7855>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/aai/ToroAR11>
<http://localhost:2020 /resource/joinedandauthors /7865>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/aai/BoothECW15>
<http://localhost:2020 /resource/joinedandauthors /7931>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/aai/DaganJLLR95>
<http://localhost:2020 /resource/joinedandauthors /7864>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/aai/Mitkov01>
<http://localhost:2020 /resource/joinedandauthors /7930>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/aai/DaganJLLR95>
<http://localhost:2020 /resource/joinedandauthors /7889>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/aamas/GlassG03>
<http://localhost:2020 /resource/joinedandauthors /7889>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/aamas/SarneG13>
<http://localhost:2020 /resource/joinedandauthors /7889>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/aamas/ElmalechSG15>
<http://localhost:2020 /resource/joinedandauthors /7889>	owl:sameAs	<http://dblp.l3s.de/d2r/resource/publications/journals/aamas/OrtizG02>

Как показывает анализ Таблицы 5, точные совпадения получены на документах из Google Scholar (англоязычные публикации), а для корпусных данных найдены только предположительные совпадения.

Заключение

В результате проделанной работы было произведено улучшение структуры базы корпусных данных, добавление в них высоко цитируемых публикаций из Google Scholar, преобразование этой базы в rdf-хранилище и связывание его с открытыми данными схожей структуры и тематики из DBLP Computer Science Bibliography.

Дальнейшее развитие экспериментов в описанной области будет связано с разработкой программной среды для повышения скорости и эффективности выполнения процесса преобразования и связывания источников данных, которая будет обладать удобным интерфейсом.

Литература

1. Невзорова О.А. Технологии связывания данных в пространстве открытых данных на примере математической коллекции / О.А. Невзорова, А.В. Кириллович // Материалы конференции OSTIS -2012. - С. 1-7.

2. Linked Data - Connect Distributed Data across the Web [Электронный ресурс], - http://linkeddata.org (Дата обращения: 01.04.2016 г.)

3. DBLP Computer Science bibliography Homepage [Электронный ресурс], - http://dblp.uni-trier.de/db/ (Дата обращения: 04.04.2016 г.)

4. Домашняя страница ESF Database Migration Toolkit [Электронный ресурс], - http://easyfrom.net (Дата обращения: 16.01.2016 г.)

5. Описание модуля generate-mapping [Электронный ресурс], - http://d2rq.org/generate-mapping (Дата обращения: 04.04.2016 г.)

6. GitHub.com: AKSW/LIMES [Электронный ресурс], - https://github.com/AKSW/LIMES (Дата обращения: 04.04.2016 г.)

7. Википедия: FOAF [Электронный ресурс], - https://ru.wikipedia.org/wiki/FOAF (Дата обращения: 04.04.2016 г.)

РАЗРАБОТКА ОНТОЛОГИЧЕСКОЙ МОДЕЛИ ОПЕРАЦИИ КОММЕРЧЕСКИХ БАНКОВ

С.А. Поздеева
Московский Педагогический Государственный Университет, Москва
englishinoz@gmail.com

Банковская сфера пронизывает все отрасли экономики, обеспечивая их денежно-кредитным потоком. Можно сказать, что банки страны - это показатель состояния экономики, поэтому так необходимо структуризировать знания о них. Это можно осуществить при помощи онтологии.

Ключевые слова: *онтология, банковские операции*

Банк является реальной производительной силой, его деятельность напрямую связана с экономикой, обеспечением непрерывности и ускорением производства, приумножением богатства общества. Банки способны сделать многое для увеличения материального производства и обмена продуктами труда. По состоянию экономики судят об активности банков, также как по состоянию банков можно судить в целом об экономическом состоянии общества.

На данный период, практически отсутствуют исследования, в которых бы обсуждались онтологические модели в области микроэкономических моделей экономики и онтологических моделей предприятий, ориентированных на экономику, а не на производство. К одной из данных отраслей можно отнести банковскую отрасль.

Под онтологией понимается подробная спецификация структуры определенной проблемной области. Иными словами, онтология представляет собой формализованное описание общепринятого понимания некоторой предметной области, обеспечивающее использование широким кругом пользователей.

Потребность в разработке онтологий возникает для решения следующих причин:

- возможность использования людьми и компьютерами с общим пониманием структуры информации;
- для повторного использования знаний предметной области;
- для того чтобы сделать допущения предметной области явными;
- для отделения знаний предметной области от оперативных знаний;
- для анализа знаний предметной области.[1]

Рассмотрим онтологии, которые предназначены для использования в различных направлениях финансовой сферы:

• Suggested Upper Merged Ontology (SUMO). Область применения: обеспечение совместимости данных, поиск информации, автоматический логический вывод, нейролингвистическое программирование. К преимуществам SUMO можно отнести возможность трансляции на любой из языков представления знаний.[2]

Рис. 1. Suggested Upper Merged Ontology (SUMO)

• Financial Instruments and Trading Strategies (FITS ontology). Данная онтология разработана с применением концепция ROD(Rapid ontology development), главная цель которой является непрерывная оценка онтологии в течение всего процесса разработки. Онтология реализована с использованием языка OWL. Область её применения: анализ финансовых инструментов и торговля финансовыми инструментами на фондовом рынке. Она позволяет использовать уже существующие торговые стратегии, или создавать новые, что в совокупности позволяет изучение отдельных экземпляров финансовых инструментов.[2]

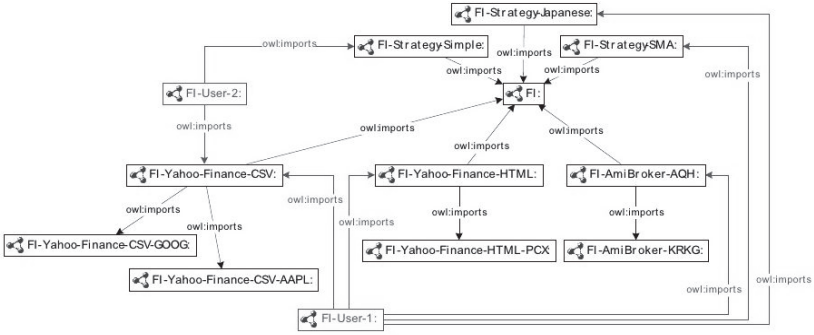


Рис. 2. Financial Instruments and Trading Strategies (FITS ontology)

- Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE). Эта онтология применяется для коммуникации участников рынка (финансового в общем и фондового в частности). Используется для согласования между интеллектуальными агентами, использующими разную терминологию.[3]

Рис. 3. Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)

- Финансовая онтология WP10. Предметной областью данной онтологии являются финансовые отношения участников экономического мира. В ней вводятся термины и понятия международного экономического мира.[4]

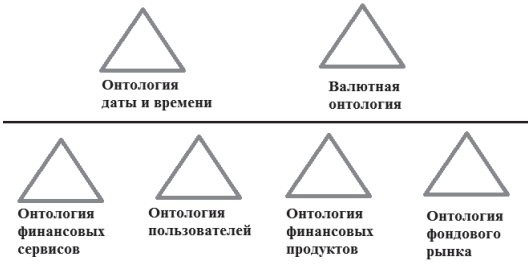


Рис. 4. Финансовая онтология WP10

Разработка онтологии проходит по следующим этапам:

1 этап. Определение классов в онтологии;

2 этап. Организация классов в некоторую иерархию, т. е. базовый класс –> подкласс;

3 этап. Определение слотов и их допустимых значений;
4 этап. Заполнение значений слотов для экземпляров классов.

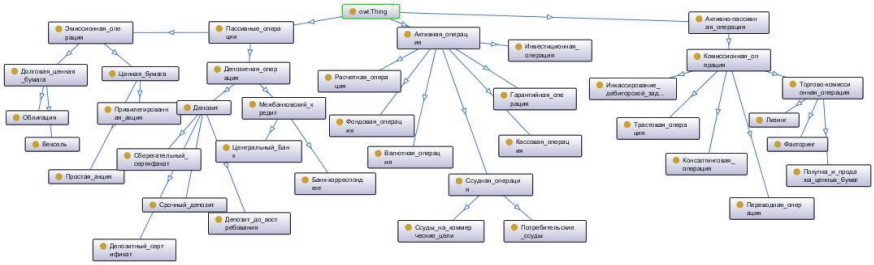


Рис. 5. Граф онтологии операции коммерческих банков

Таким образом, была спроектирована онтология операций коммерческих банков. Видно насколько в ней структурированы понятия, термины, связи, поэтому данная база знаний будет понятна и пользователю, и разработчику, не имеющему специальных знаний и навыков в финансовой сфере.

Литература

1. *Муромцев Д.И.* Онтологический инжиниринг знаний в системе Protégé [Электронный ресурс]: СПб.: СПб ГУ ИТМО, 2007. URL: http://books.ifmo.ru/file/pdf/243.pdf (дата обращения: 02.04.2016).

2. *Dejan Lavbič, Marko Bajec* Employing Semantic Web technologies in financial instruments trading // International Journal on New Computer Architectures and Their Applications (IJNCAA). 2012. Vol. 2(1). P. 167-182

3. *Steffen Lamparter, Björn Schnizler.* Trading services in ontology-driven markets. In Proceedings of the 2006 ACM symposium on Applied computing, SAC '06, New York, NY, USA, 2006. P.1679–1683

4. *Alonso L. S., Bas L. J., Bellido S., Contreras J., Benjamins R. and Gomez M. J.* WP10: Case Study eBanking D10.7 Financial Ontology [Электронный ресурс] // Data, Information and Process Integration with Semantic Web Services, FP6-507483, 2005 URL: http://dip.semanticweb.org/documents/D10-7-Stock-Market-Ontology.pdf (дата обращения: 02.04.2016).

5. *Лапшин В.А.* Онтологии в компьютерных системах // М.: Научный мир, 2010. с. 222.

УДК 004.822:514

ПРЕДИКАТИВНО-ОБРАЗНЫЙ КОНТРОЛЬ ПОСТАНОВОК ЗАДАЧ

П.И. Соснин, М.В. Галочкин, А.А. Лунецкас
Ульяновский государственный технический университет, Ульяновск
sosnin@ulstu.ru, m.galochkin@ulstu.ru, lunacorp@inbox.ru

В статье представляются средства псевдо кодовой программируемой графики поддержки проектирования автоматизированных систем на концептуальном этапе. Специфику подхода определяет преобразование текстовых единиц в прологоподобные конструкции и семантические граф-схемы в условиях взаимодействия с онтологией проекта. Возможность обратного преобразования схемы после её коррекции позволяет итеративно доводить её и исследуемую текстовую единицу до взаимно согласованных состояний, констатирующих и регистрирующих проверяемую версию понимания проектировщиком освоенного текста.

Ключевые слова: онтология, понимание, постановка задачи, предикат, семантическая графика

Введение

В проектировании автоматизированных систем широко используются средства текстового и табличного сопровождения, а также средства чертёжной и псевдо графики, обслуживающие создание диаграмм и виде block-and-line схем. Текстовое и табличное сопровождение в САПР заимствует практически всё достигнутое в разработках текстовых и табличных процессоров, но чаще всего только в привязках к решению задач формирования и регистрации проектной документации. Однако, как в поддержке работ с графикой, так и в работе с текстами практически отсутствует выход на автоматизированное решение задач семантики, занимающих очень важное место в оперативной работе проектировщика и особенно процессах принятия проектных решений.

Для проектировщика выход на ту семантику, для которой полезно её графическое представление, начинается с первых шагов его взаимодействия с техническим заданием и особенно принципиален на этапах концептуального проектирования, когда только еще нащупываются потенциальные прототипы решений, когда выход на реальную геометрию и нормативную графическую (образную) регистрацию преждевременен

[1]. На этом этапе основная информация поступает к проектировщику через тексты, во взаимодействии с которыми он практически лишён автоматизированной графической поддержки. Такое положение дел послужило основанием для разработки средств автоматизированного преобразования текстовых описаний ситуаций в семантические граф-схемы и схемы других типов. Предлагаемые средства реализованы как расширение инструментально-моделирующей среды OwnWIQA [2].

Основная идея

При увеличении сложности программных систем на первый план встает проблема понимаемости. Человеческие ресурсы ограничены и разработчику трудно представить всю систему в целом. Для решения такого рода проблем обычно вводят дополнительные уровни абстракции, модели наподобие черного ящика, стараются увеличить эргономические показатели среды разработки. Одной из хорошо зарекомендовавших методик является предоставление разработчику возможности экспериментировать с разрабатываемой системой. При исследовании работы кода большой системы удобно пользоваться разного рода анализаторами кода (автоматическое построение различных UML диаграмм, просмотр зависимостей между классами и т.д.) и отладчиками. Последнее позволяет остановиться в различных местах кода, исследовать стек вызова, посмотреть текущие значения переменных. При исследовании постановки задачи обычно проектировщику не предоставляется возможность экспериментирования с предметной областью. Отсутствие средств поддержки накладывает ряд трудностей, так как постановка задачи может содержать противоречия, неточности, не полностью описывать то, что необходимо сделать. Для предоставления таких возможностей был разработан комплекс средств, позволяющих из прологоподобного описания постановки задачи сгенерировать и исполнить код на языке пролог, тем самым предоставив проектировщику богатый набор средств для экспериментирования. Особое внимание в статье уделено переводу исследуемого текстового описания на прологоподобный язык и исполнение построенной декларативной программы на пролог-интерпретаторе.

Особенности реализации

В качестве основы для среды исполнение предикатов была взята свободная реализация языка пролог – SWI-Prolog. Эта среда обладает богатым набором возможностей (богатые библиотеки, интерфейс к языку java, ODBC и т.д.), хорошей стабильностью работы (среда развивается

с 1987г.) а также является достаточно популярной в университетской среде. Для взаимодействия с интерпретатором была написана прослойка, занимающаяся проксированием запросов из внешнего интерфейса в интерпретатор и обратно. Такая схема взаимодействия со средой отличается большой гибкостью и может работать практически с любым интерпретатором пролог. Общая схема взаимодействия среды с SWI-Prolog[3, 4, 5] представлена на рисунке 1.

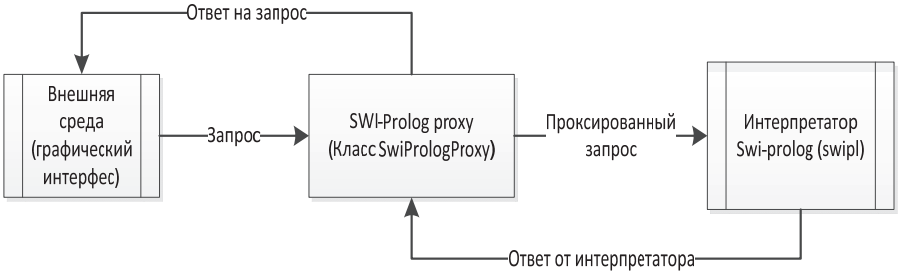


Рисунок 1 Общая схема взаимодействия с пролог интерпретатором

Компонент SwiPrologProxy является частью среды WIQA и призван решать следующие задачи:

• Проксирование запросов из среды в интерпретатор. Синтаксисы предикатов используемого в среде и в интерпретаторе отличаются, поэтому приходится производить преобразование из одного синтаксиса в другой на лету. Это становится возможным, так как синтаксис, используемый в WIQA и синтаксис SWI-Prolog [3] имеют определенное соответствие и не зависят от состояния интерпретатора.

• Обработка дополнительных команд. Некоторые команды, посланные в прокси не должны проксироваться в интерпретатор. К таким командам преимущественно относятся служебные команды (очистка экрана консоли, статистика и т.д.) [6].

• Обработка ответов интерпретатора. Для выполнения команды интерпретатору необходимо некоторое время. После выполнения запроса в интерпретатор класс прокси ожидает прихода определенной последовательности символов, которые служат сигналом о том, что запрос был выполнен. К таким символам относятся символы перевода строки (“\r\n”). После получения такой последовательности прокси считывает результат и передает его дальше на обработку.

• Обработка успешного результата. В случае если результат выполнения успешен, прокси передает ответ в неизменном виде

в графическую среду. В случае если ответ представляет собой ошибку, происходит обработка ответа с целью приведения к оригинальному синтаксису (как было описано выше синтаксисы предикатов отличаются). Часто эта операция сводится к простому поиску и замене подстроки в строке [7].

В связи с вышеизложенными требованиями реализация класса SwiPrologProxy не является достаточно простой. Общий размер кода класса порядка 400 строк.

Пример работы. Сравнение синтаксисов.

Рассмотрим пример работы с интерпретатором пролога, который представлен на рисунке 2.

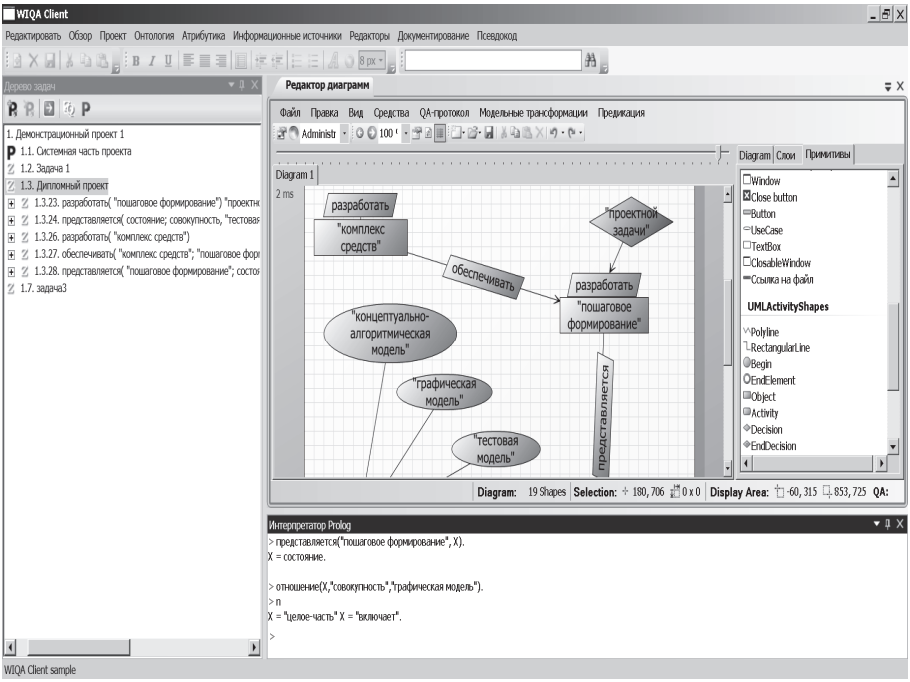


Рисунок 2 Общий вид окна клиента WIQA
с открытой консолью интерпретатора

Начнем с рассмотрения того каким образом представляются прологоподобные формы в вопросно-ответном дереве (рисунок 2, блок 2). Для примера возьмем следующую постановку задачи «Разработать комплекс средств обеспечивающих пошаговое формирование постановки

проектной задачи, состояние которой представляется взаимосвязанной совокупностью текстовых, графических и концептуально-алгоритмических моделей»[8]. После представления данного предложение в вопросно-ответном протоколе мы получим следующее прологоподобное описание:

разработать("пошаговое формирование") "проектной задачи"
представляется(состояние; совокупность, "тестовая модель", "графическая модель", "концептуально-алгоритмическая модель")
разработать("комплекс средств")
обеспечивать("комплекс средств"; "пошаговое формирование")
представляется("пошаговое формирование"; состояние)

Как видно представление не является предикатом в строгой форме, так как содержит часть, названную дополнением. При разборе предложения существуют элементы, которые не выражают в явном виде отношения, но несут некоторую смысловую нагрузку. Очень часто такие элементы являются дополнением.

Так как интерпретатор пролог не работает с дополнениями, то при переводе они будут исключаться. После трансформации запросов в интерпретатор пролог будет передан следующий код:

assert(разработать("пошаговое формирование")).
assert(представляется (состояние; совокупность, "тестовая модель", "графическая модель", "концептуально-алгоритмическая модель")).
assert(разработать ("комплекс средств")).
…

Как видно отличия достаточно не существенные. Во-первых, необходимо убрать дополнения и превратить в конструкцию вида «assert(P(X,Y,…)).». Это конструкция служит для описания утверждений.

Каждая конструкция при передаче ее в интерпретатор возвращает true в случае успеха и false с кодом ошибки в случае наличия проблем в синтаксисе. На рисунке 3 представлена, сессия работы с интерпретатором.

Кроме перевода и исполнения прологоподобного кода из вопросно-ответного дерева, в интерпретатор может загружаться онтология, связанная с этим проектом. В словаре онтологий хранятся связи между различными терминами предметной области. Загрузка словаря представляет собой выполнение команд вида:

assert(отношение (“целое-часть”, ”совокупность моделей”, ”графическая модель”)).
assert(отношение (“включает”, ”совокупность моделей”, ”графическая модель”)).
…

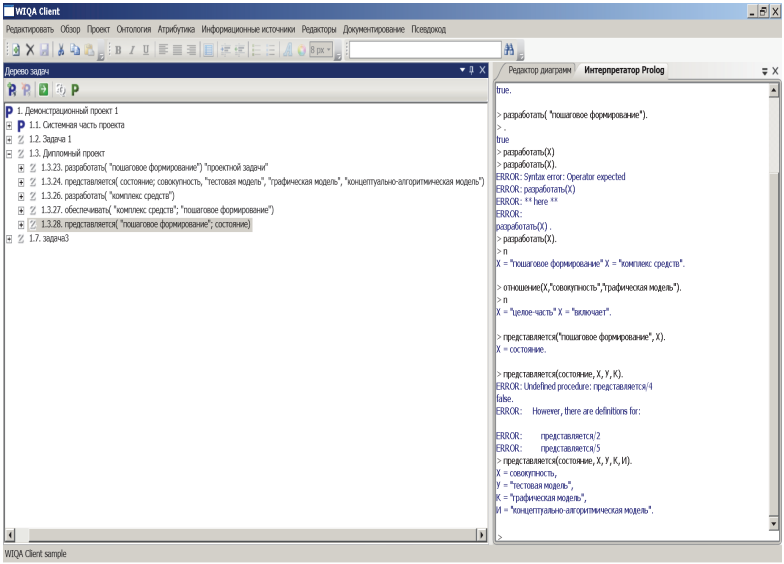


Рисунок 3. Пример работы с интерпретатором

Кроме перевода и исполнения прологоподобного кода из вопросно-ответного дерева, в интерпретатор может загружаться онтология, связанная с этим проектом. В словаре онтологий хранятся связи между различными терминами предметной области. Загрузка словаря представляет собой выполнение команд вида:

assert(отношение (“целое-часть”, ”совокупность моделей”, ”графическая модель”)).

assert(отношение (“включает”, ”совокупность моделей”, ”графическая модель”)).

…

Это необходимо для того что на этапе экспериментирования с постановкой задачи может возникнуть ситуация когда проектировщику не известны некоторые термины предметной области или отношения между ними. В данном случае можно вызвать следующую команду и получить ответ (пример из рисунка 3):

отношение(X, ”совокупность моделей”, ”графическая модель”).

X=“целое-часть”

X=“включает”

Когда значение переменной X может принимать больше чем 1, значение интерпретатор после каждого найденного значения останавливается в ожидании команды на продолжение или остановку поиска ответа. Для продолжения необходимо ввести символ “n” (next). Более детальную информацию по среде SWI-Prolog можно найти на официальном сайте http://www.swi-prolog.org/pldoc/doc_for?object=manual.

Согласование текстовой единицы и соответствующей ей семантической граф-схемы представлено на рисунке 4.

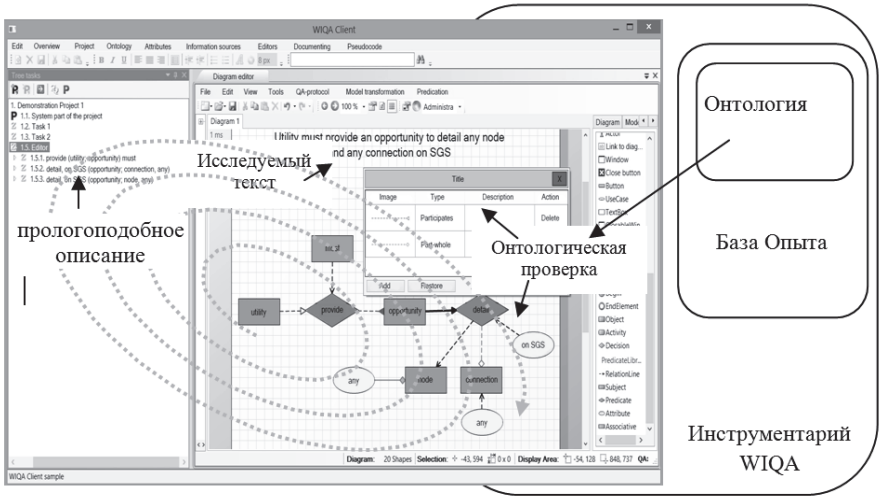


Рисунок 4. Итеративное согласование текста и графики

На рисунке приведены поясняющие метки и показано взаимодействие диаграммы с онтологией. Такое взаимодействие обеспечивает обогащение семантического содержания графики, что переносится и на прологоподобное описание. Взаимодействие, дополнительно, позволяет раскрыть системные связи между блоками диаграммы, которые не представлены явно в исследуемой текстовой единице.

Схем на рисунке 4 также отражает (спиралью) процесс итеративного согласования текста и его графического представления. Основой такого согласования служит обратная связь между текстом и графикой, способствующая переносу любых «позитивных изменений» от текста на графику и наоборот.

Заключение

Представленный в статье подход позволяет выполнять предикативный контроль предметной области с помощью исполнения постановки задачи в интерпретаторе пролог. Такое отображение позволяет задавать вопросы интерпретатору и детально исследовать предметную область. Наличие возможности подгружать словарь онтологий связанный с конкретной предметной областью делает процесс экспериментирования более дружелюбным. Благодаря этому у проектировщика появляется возможность в интерпретаторе узнать определение неизвестного ему термина, или выяснить возможные связи между терминами. Для достижения такой функциональности был разработан класс SwiPrologProxy осуществляющий трансляцию команд из среды WIQA в интерпретатор, получение и обработку ответов. В случае возникновения ошибок производится корректировка кода ошибки с целью правильного указания места в исходной форме. Реализация достаточно оригинальна и позволяет легко расширять количество поддерживаемых интерпретаторов и сред.

Литература

1. Галочкин, М.В., Соснин, П.И. средства псевдокодовой программируемой графики в проектировании автоматизированных систем// Автоматизация процессов управления. — 2015. — № 1 (39). — С. 82-88.

2. Sosnin, P. Question-Answer Approach to Human-Computer Interaction in Collaborative Designing// Cognitively Informed Intelligent Interfaces: Systems Design and Development. IGI Global. —2012. — pp. 157-176.

3. SWI-PROLOG Reference Manual [Электронный ресурс]. May 2014. URL: http://www.swi-prolog.org/download/stable/doc/SWI-Prolog-6.6.6.pdf (Дата обращения: 18.01.2016).

4. Программирование на языке ПРОЛОГ[Электронный ресурс] // МГУ имени Ломоносова. М., 2013. URL: http://recyclebin.ru/BMK/prolog/PrProlog.pdf (Дата обращения: 18.01.2016).

5. Хабаров, С.П. Prolog – язык разработки интеллектуальных и экспертных систем [Электронный ресурс]. С-П, 2013. URL: http://www.habarov.spb.ru/book_prolog_ 2013/SerpBook_Prolog.pdf (Дата обращения: 18.01.2016).

6. Schatz, B. Verification of Model Transformations [Электронный ресурс]. 2009. URL: http://www4.in.tum.de/~schaetz/papers/GT-VMT-Verification.pdf (Дата обращения: 18.01.2016).

7. Паронджанов, В. Язык дракон. Краткое описание[Электронный ресурс]. 2001. URL: http://drakon.su/_media/biblioteka/drakondescription.pdf(Дата обращения: 18.01.2016).

8. Sosnin, P. Scientifically Experimental Way-of-Working in Conceptual Designing of Software Intensive Systems, In Proceedings of the IEEE 12th International Conference on Intelligent Software Methodologies, Tools and Techniques, pp. 43-51, 2013

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проект № 15-07- 04809а).

УДК 81’33

МОДЕЛИРОВАНИЕ СЕМАНТИКИ ЕСТЕСТВЕННОГО ЯЗЫКА НА ОСНОВЕ МУЛЬТИАГЕНТНОЙ РЕКУРСИВНОЙ КОГНИТИВНОЙ АРХИТЕКТУРЫ[1]

Б.П. Тажев, Д.Г. Макоева, И.А. Пшенокова
ФГБУН «Институт информатики
и проблем регионального управления КБНЦ РАН»,
360000, КБР, г. Нальчик, ул. И.Арманд 37-а
E-mail: boristazhevart@mail.ru, d.makoeva@iipru.ru,
pshenokova_inna@mail.ru

В работе представлена модель формализации семантики естественного языка, реализованная на базе мультиагентной рекурсивной когнитивной архитектуры.

***Ключевые слова**: мультиагентные системы, когнитивная архитектура, формальная семантика, естественно-языковой интерфейс.*

1. Модели и методы формализации семантики естественного языка

Первые попытки по созданию интеллектуальных систем понимания речи и компьютерных переводчиков предпринимались еще в середине прошлого столетия, однако ни одна из них не увенчалась особым успехом.

В 70-х годах пошлого века Роджер Шенк предложил теорию концептуальных зависимостей. В его теории предложение строится не из лексических единиц, а посредством концептуальных примитивов [Шенк, 1980].

Еще одним способом формальной репрезентации семантики являются фреймы. Основоположником данной теории был Марвин Минский. Фреймы – это структуры, которые позволяют выстраивать знания в виде сложных сущностей, отображающих взаимосвязь объектов определенной предметной области [Минский, 1979].

Среди работ отечественных лингвистов особое место занимает теория И.А. Мельчука, в создании которой принимали активное участие А.К. Жолковский и Ю.Д. Апресян. Семантическая репрезентация в теории «Смысл ↔ Текст» основывается на определенном наборе сем, т.е. атомов смысла. Семантическое значение высказывания представляется в виде «структурных формул», так называемых, семантических графов.

В вершинах графов находятся семы, а дуги между ними – это связывающие их взаимоотношения [Мельчук, 1999].

Область науки под названием прикладная семиотика появилась около двадцати лет назад. Семиотике, как части лингвистики, традиционно отводились только гуманитарные вопросы, касающиеся в основном главной знаковой системы, т.е. естественного языка. Яркими представителями данного направления в России являются: Г.С. Осипов, Д.А. Поспелов, Р.Г. Бухараев, Д.Ш. Сулейманов [Поспелов, Осипов 1999].

Естественные системы, успешно работающие со знаками, вдохновили ученых на создание биоинспирированных алгоритмов [Dorigo, Gambardella, 1997], [Holland, 1975], [Курейчик, Лебедев, 2006]. Среди них особо выделяются искусственные нейронные сети [Мак-Каллок, Питтс, 1956].

Нейроны можно представить в качестве интеллектуальных агентов, а их совокупность – в качестве вычислительной сети, в которой происходят процессы самоорганизации [Нагоев а, б, 2013].

Основой для интеллектуальных агентов является когнитивная архитектура. В рамках когнитивных наук, появлению которых способствовала «когнитивная революция» 1959 [Хомский, 1959], было разработано много подходов для объяснения сущности психических процессов. На стыке информатики и когнитивистики возникла гипотеза о необходимости моделирования интеллектуальных систем в неразрывной связи с моделями нервной системы живых организмов [Anderson, 2005].

2. Формализация семантики на основе мультиагентной рекурсивной когнитивной архитектуры

Мультиагентная рекурсивная когнитивная архитектура (МуРКА) – это абстрактная модель самоорганизации мозга, предполагающая автоматическую интерпретацию естественно-языковых высказываний, при которой агенты сообщают о своих решениях в рамках когнитивной архитектуры посредством контрактных отношений, т.е. продажи и покупки знаний.

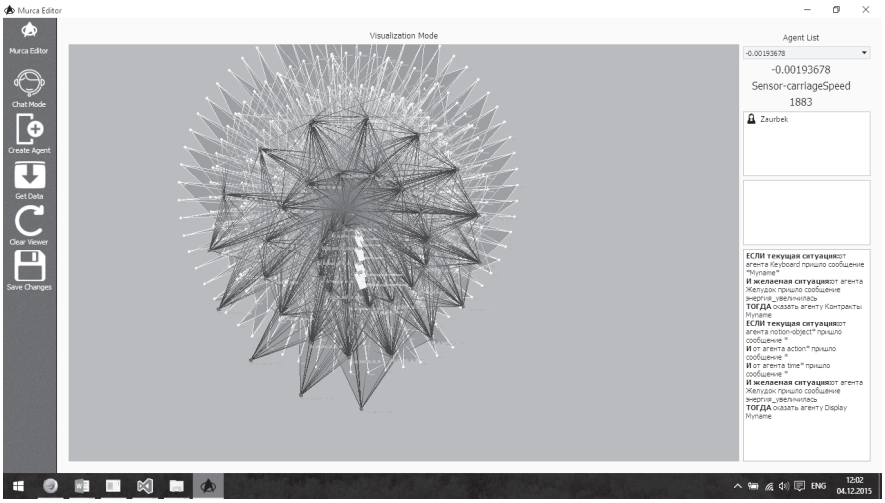


Рис.1. Мультиагентная рекурсивная когнитивная архитектура

Предпосылки по созданию МуРКА были заложены в следующих работах [Nagoev, 2012].

Как известно из семиотики, слово, как и любой другой знак обладает двумя «планами»: план содержания (смысл) и план выражения (форма). Соотношение двух этих планов, с одной стороны, может показаться тривиальным: каждому слову соответствует некоторый предмет внешнего мира. Однако, с другой стороны, корреляция плана содержания слова с его планом выражения не наблюдается во внешнем мире, где между словом и предметом отсутствует какая-либо осязаемая связь. Данное соотношение происходит в психической, мыслительной сфере деятельности человека [Шенк, 1980].

Следовательно, для того, чтобы соединение двух планов произошло, человеку необходимо иметь в голове отражение (представление) данного предмета. Таким образом, соотношение слова с предметов происходит по цепочке через отражение этого предмета в мышлении. Эта связь планов языкового знака часто изображается, так называемым, семантическим треугольником Фреге.

В МуРКА соотношение плана выражения и содержания, предположительно, будет осуществляться следующим образом: входной аудиосигнал или буквенный ввод с клавиатуры будем называть знаком. На любое введенное в систему слово (знак) в реальном мире есть свой денотат, т.е. предмет, явление, свойство. Объектным сигнификатом мы

называем свойства денотата (круглый, оранжевый и т.д.), т.е. понятие. Под языковым сигнификатом мы подразумеваем агента, отвечающего за отдельное слово, т.е. его вербальное представление в системе, другими слова, понятие-слово. Рисунок 2 демонстрирует данное семантическое соотношение, представленное в МуРКА.

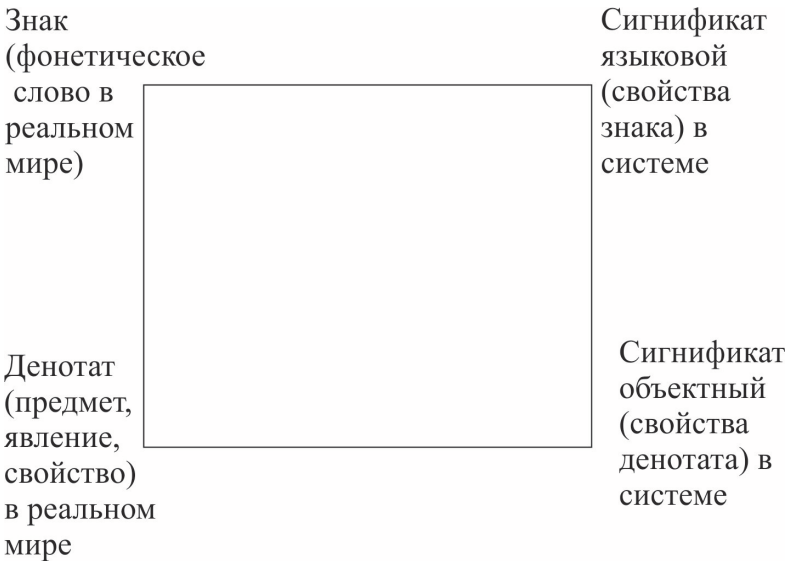


Рис. 2. Семантическая связь в МуРКА

Заключение

Предложенная модель формальной семантики естественного языка, реализованная на базе мультиагентной рекурсивной когнитивной архитектуры, может быть использована для решения задач организации высокоуровневого естественно-языкового интерфейса с интеллектуальными системами или роботами. Основополагающим отличием предлагаемого нами способа формализации семантики, предположительно, является то, что он включает в себя все вышеперечисленные способы в качестве подклассов.

Литература

1. Бабкин Э.А., Козырев О.Р., Куркина И.В. Принципы и алгоритмы искусственного интеллекта. Нижний Новгород, 2006, 133 с.

2. Елекова О.А., Нагоев З.В., Анчеков М.И. Интерактивная среда обучения виртуальных человекоподобных агентов интеллектуальному поведению // Известия КБНЦ РАН, 2006.

3. Ильин Г.М., Игнатова В.Н. Система "Рисунок-текст". // Программные продукты и системы. -1992. - № 2. С. 48-53

4. Курейчик В. М., Лебедев Б. К., Лебедев О. К. Поисковая адаптация: теория и практика. — М: Физматлит, 2006. — С. 272.

5. Мак-Каллок У.С., Питтс В. Логическое исчисление идей, относящихся к нервной активности // Автоматы / Под ред. К. Э. Шеннона и Дж. Маккарти. — М.: Изд-во иностр. лит., 1956. — С. 363—384. (Перевод английской статьи 1943 г.)

6. Мельчук И.А. Опыт теории лингвистических моделей «Смысл↔Текст» М.: «Языки русской культуры», 1999. 345 с.

7. Минский М. Фреймы для представления знаний М.: Энергия, 1979. 151 с.

8. Нагоев З.В. Интеллектика, или мышление в живых и искусственных системах // Нальчик: Издательство КБНЦ РАН, 2013 – 211 с.

9. Нагоев З.В. Мультиагентные экзистенциальные отображения и функции // Известия КБНЦ РАН, Нальчик: Издательство КБНЦ РАН, 2013, № 4 (54), с. 64-71.

10. Поспелов Д.А., Осипов Г.С. Прикладная семиотика //Новости искусственного интеллекта №1, 1999

11. Филлмор Ч. Дело о падеже // Новое в зарубежной лингвистике М.: ПРОГРЕСС, 1981 С. 369-496

12. Шенк Р., Бирнбаум Л., Мей Дж. К интеграции семантики и прагматики // Новое в зарубежной лингвистике М.: ПРОГРЕСС, 1989. С. 31- 47

13. Шенк Р. Обработка концептуальной информации / Пер. с англ. М.: Энергия, 1980. – 360 с.

14. Anderson, J. R. (2005) Human symbol manipulation within an integrated cognitive architecture. CognitiveScience, 29(3), 313-341.

15. Chomsky N. A. (1959), A Review of Skinner's Verbal Behavior

16. Dorigo M., L.M. Gambardella (1997). Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem. IEEE Transactions on Evolutionary Computation, 1, 1, 53-66.

17. Holland J.H. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan Press, 1975 - 183

18. Nagoev Z.V. Multiagent recursive cognitive architecture // Biologically Inspired Cognitive Architectures 2012, Proceedings of the third annual meeting of the BICA Society, in Advances in Intelligent Systems and Computing series, Springer, 2012, pp. 247-248.

РАЗДЕЛ II | КОРПУСНАЯ ГРАММАТИКА И СИСТЕМЫ АННОТИРОВАНИЯ (UNITURK)

УДК 81'33 Прикладная лингвистика

СТАТИСТИЧЕСКИЕ МЕТОДЫ ИССЛЕДОВАНИЯ КОНКУРИРУЮЩИХ ГЛАГОЛЬНЫХ ФОРМ В РУССКОМ ЯЗЫКЕ

Т.И. Галеев
Казанский федеральный университет, Казань
TIGaleev@kpfu.ru

В работе описывается конкуренция грамматических синонимов – глаголов, образованных при помощи суффиксов –а/я–, и – ива/ыва (*оздоровлять*/*оздоравливать*). На основе данных корпуса Google Books были описаны базовые модели изменения динамики частотности конкурирующих форм.

Ключевые слова: *вариативность, Google Books, языковая динамика, глагольная парадигма, когнитивная лингвистика.*

Один из самых острых вопросов филологии, который помимо отечественных специалистов [3], [5] интересует широкие слои общества – это вопрос о норме. Особое внимание в современных исследованиях стало уделяться ситуации, при которой новый вариант по частоте сравнивается со старым. Например, *изготавливать* и *изготовлять*.

Из зарубежных лингвистов интересных результатов добились группы Э. Либермана для английского языка [1] и Л. Янды [2] для русского языка. В исследовании группы Э. Либермана впервые для изучения механизма смены неправильной формы глагола на правильную получен важный результат: чем неправильный глагол более частотен, тем больше времени требуется для его перехода в разряд правильных. Когнитивисты Л. Янда и Т. Нессет на примере глаголов с вариативностью типа *хнычет – хныкает* показали, что словоизменительная парадигма

имеет радиальную структуру, т.е. в ней можно выделить центр (3 Sg > 3 Pl > 1 и 2 Sg) и периферию(императив > причастия > деепричастия), причем элементы парадигмы упорядочены. Вывод сделан на основе детального изучения частоты встречаемости вариантов для всех словоизменительных форм по данным НКРЯ. Оказалось, что хотя в целом в русском языке в этих глаголах наблюдается сдвиг от формы на *–а* к форме на *–aj*, глаголы в 3 Sg дольше сохраняют исходную форму, а более периферийные легче переходят на новую. Так *хнычет* употребляется все еще чаще, чем «*хныкает*», но *хнычущий* уже реже, чем *хныкающий*.

В русском языке продуктивным способом образования глаголов НСВ является вторичная имперфективация с помощью суффикса *–ива/ыва–*, например: *изготовлять – изготавливать*. В речи появляется вариативность.

Целью исследования является выявление закономерностей эволюции вариативных форм центра глагольной парадигмы. Конкретные решаемые задачи: выделение случаев смены нормы за 2 века; получение численных характеристик изменений наиболее «консервативных» членов парадигмы. На основе данных корпуса Google Books будет осуществлён анализ частотности словоупотребления консервативных форм т.н. «избыточных» глаголов.

Для изучения эволюции вариативных форм предлагается применить квантитативный метод. На основе данных корпуса Google Books, осуществляющего поиск по книгам, изданным в основном с 1800 по 2000 гг., будут построены графики изменения частотности глаголов, имеющих избыточную парадигму. Характер корпуса определяет стилистический аспект исследования глаголов: сфера употребления глаголов – художественная, научная и научно–популярная литература.

В «Стилистическом словаре вариантов» Л.К. Граудиной [4] было найдено 38 пар приставочных глаголов несовершенного вида с чередованиями *–а/я–//–ива/ыва–*, Для получения более полных и объективных данных помимо инфинитива было решено проанализировать ещё 11 пар форм: 1, 2, 3 Sg/Pl, PastMSg, Part. (Act., Pres.), Part. (Act., Past) Gerund, Imperative. При составлении поисковых запросов по 456 парам глагольных форм было получено и проанализировано 344 графика, а в 112 случаях запрашиваемые формы в корпусе встречены не были. В меньшинстве случаев (38,4%) встречается только одна из форм, другую найти не удалось. Почти всегда единственной функционирующей словоформой оказывался именно вариант с суффиксом *–а/я–*, в форме 1 и 2 Sg/Pl или деепричастия. Остальные графики по причине отсутствия их единообразия было решено разделить на 6 групп по типу взаимодействия анализируемых форм: 1. В 25,9%

графиков в начале XIX в. преобладает форма с *–а/я–*, в конце XX в. – её «конкурент» с *-ива/ыва–* (*приспособлять → приспосабливать*). Такого вида графики способны проиллюстрировать в том числе и смену нормы; 2. В 15% случаев в конце XX в. утверждается только один из вариантов, хотя в течение XIX в. обе формы употреблялись одинаково часто; 3. В 5,2% случаев одна из форм (обычно с суффиксом *–а/я–*) стабильно употребляется чаще, чем другая (*обособлять/обосабливать*); 4. В 4,7% случаев на протяжении 200 лет обе формы употребляются одинаково часто (*простужаться/простужи-ваться*); 5. В 4% случаев обнаруживается начало процесса смены нормы, то есть в XIX в. одна из форм преобладала над другой, а в XX в. они стали использоваться одинаково часто (*оздоровлять/оздоравливать*); 6. В 3,7% произошла архаизация глагола, то есть одна из форм в XIX в. употреблялась часто, другая – редко, а в XX в. обе формы редки и говорить об изменении нормы говорить уже не приходится (*обрезаете/обрезываете*).

Закономерность, связанная с неравномерностью унификации глагольной парадигмы, выявленная когнитивистами, лишь частично находит применение к другой группе глаголов с избыточной парадигмой – одинаковые графики были получены для большинства форм одних и тех же глаголов. Например, во всех формах глагола глагол *простужаться* на протяжении 2 веков была конкуренция форм, но смены нормы не произошло. Таким образом, гипотеза Л. Янды о консервативности «центральных» форм глагольной парадигмы не подтверждена для данной группы глаголов.

Также было обнаружено большое количество фактов, противоречащих многочисленным работам, посвящённых исследованию нормы [3], [4], [5]. Так, в большом количестве случаев отмечается возврат к старой форме на –а/я–.

Вторым этапом исследования стало составление списка частотности данных глаголов на материале Google Books и НКРЯ. Если группа Э. Либермана для работы со списком глаголов в более 200 глаголов применяла метод ранжирования, то в нашем исследовании 38 пар глаголов более целесообразно сравнить 10 наименее частотных глаголов (у каждого до 10 тыс. словоупотреблений за 200 л.) с тем же количеством наиболее частотных глаголов (от 90 до 280 тыс. словоупотреблений за 200 л.). Примечательно, что списки, сделанные на основе НКРЯ и GBN совпадали примерно на 70%.

В динамике изменения частотности 8 из 10 редчайших глаголов (*надломлять–надламывать, надломляться–надламываться, опорожнять–опоражнивать, опорожняться–опоражниваться, простужаться–простуживаться, обмерять–обмеривать, вымерять–вымеривать,*

оздоровлять–оздоравливать, засоряться–засариваться, накалять–накаливать) превалирует тенденция к выбору формы с суффиксом *–а/я–*. При этом, как показано на Графике № 1, в XIX в. оба варианта употреблялись одинаково часто.

График № 1. Изменение частотности глаголов *простужаться* и *простуживаться*

Аналогичная динамика наблюдается в 3 из 10 наиболее распространённых глаголов (*вырезать–вырезывать, срезать–срезывать, ускорять–ускоривать, ускоряться–ускориваться, накоплять–накапливать, накопляться–накапливаться, приготовлять–приготавливать, приспособляться–приспосабливаться, сужаться–суживаться, изготовлять–изготавливать*) с тем отличием, что в XIX в. преобладала форма с суффиксом –ива/ыва–. График № 2 иллюстрирует возвращение исходной формы.

График № 2. Изменение частотности деепричастий *вырезая* и *вырезывая*

В четырёх случаях (*приготовлять–приготавливать, ускорять–ускоривать, ускоряться–ускориваться, сужать–суживать)* форма с суффиксом –а/я– является единственной найденной или, по крайней мере, доминирующей в плане частотности. Таким образом, 7 из 10 самых ча-

стотных глаголов своей динамикой могут проиллюстрировать теорию группы Э. Либермана и других лингвистов об устойчивости самых частотных форм к морфологическим или фонетическим изменениям. И, наоборот, 8 из 10 самых малоупотребительных глаголов служат опровержением данной теории: по Э. Либерману и др., в этой группе должен был произойти переход к –ива/ыва–, но в русском языке эта норма не утвердилась и произошёл возврат к формам с –а/я–.

В трёх парах из первого списка (*приспособляться–приспосабливаться, накоплять–накапливать, изготовлять–изготавливать*) и одной паре из второго (*оздоровлять–оздоравливать*) прослеживается обратная тенденция: формы с –а/я–, отличавшиеся частотностью десятки лет назад, постепенновытесняются формами на –ива/ыва–. Графики № 3 и 4 иллюстрируют разные этапы смены нормы.

График № 3. Изменение частотности глаголов *приспособляться* и *приспосабливаться*

График № 4. Изменение частотности глаголов *изготовлять* и *изготавливать*

Причины появления случаев, которые противоречат общей тенденции возврата к форме с суффиксом –а/я–, судя по (), связаны с фоносемантикой. Нежелательная ассоциация, вызванная сочетанием эксплозивных

губных согласных звуков [б], [в] и [п] с вставочным *л* (*l-epeneticum*) и суффиксом –я(ть), делает вариант с -ива/ыва– более предпочтительным в силу его благозвучности. Обнаружение данного явления стало возможно благодаря только квантитативным методам при помощи корпуса Google Books. Ранее в соответствующей литературе упоминалась только неблагозвучность конфиксальных глаголов с –ива–, образованных от существительного *срок* (**отсрачивать*) [6].

Говоря о других результатах исследования, можно сказать, что на основе сверхобъёмного корпуса текстов книжного стиля впервые была построена классификация глагольных пар в зависимости от динамики частотности их употребления. Методы описания динамики вариативных форм глаголов могут быть применены в дальнейшем и в других случаях. При этом они позволят не только описывать, и не только объяснять лингвистические явления, но и делать обоснованные количественные предсказания развития языковых форм.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проект №16–06–00165 А «Когнитивная модель словоизме-нительной глагольной парадигмы в русском языке: квантитативный анализ динамики частотности словоформ»).

Литература

1. Lieberman E., Michel J.–B., Jackson J., Tang T., Nowak M.A. Quantifying the evolutionary dynamics of language // Nature, 2007. Vol. 449. pp. 723–716. http://www.nature.com/nature/journal/v449/n7163/abs/nature06137.html [Электронный ресурс. Дата обращения: 31.03.2016]
2. Nesset T., Janda L. Paradigm structure: Evidence from Russian suffix shift. // Cognitive Linguistics, 2010. Vol. 21(4), pp. 699–725.
3. Горбачевич К.С. Вариантность слова и языковая норма: На материале современного русского языка. – М: URSS: ЛИБРОКОМ, 2009. – 240 с.
4. Граудина Л.К. Грамматическая правильность русской речи. Стилистический словарь вариантов. – М.: АСТ: Астрель, 2004. 557 с.
5. Земская Е. А. Русская разговорная речь: лингвистический анализ и проблемы обучения. – М.: Наука; Флинта, 2004. – 240 с.
6. Розенталь Д. Э. А как лучше сказать? М.: Просвещение, 1988. – 176 с.

УДК 811.512.145; 81'366.5

ФУНКЦИОНИРОВАНИЕ АФФИКСА СИМИЛЯТИВА -ДАЙ В ТАТАРСКОМ ЯЗЫКЕ (НА КОРПУСНЫХ ДАННЫХ)

А.М. Галиева
НИИ «Прикладная семиотика» АН РТ
amgalieva@gmail.com

В работе рассматриваются особенности функционирования аффикса симилятива -ДАй в татарском языке. Статья состоит из двух частей: в первой рассматриваются типы аффиксальных цепочек, типичных для образований с аффиксом симилятива -ДАй, во второй исследуется вопрос о том, какого типа конструкции вводятся при помощи данного аффикса.

Ключевые слова: *аффикс, симилятив, татарский язык, значение уподобления.*

Особенности аффиксальной цепочки тюркской словоформы проанализированы в ряде специальных работ, в последние годы работы такого рода посвящены разработке корпусов тюркских языков (см. [3, 5, 6] и др.). Тем не менее, функционирование ряда аффиксов татарского языка остаются неизученной.

Аффикс -ДАй является одним из специфических аффиксов татарского языка, вводящих конструкции со значением сравнения и уподобления. Агглютинативная морфология освобождает тюркские языки от необходимости использования союзов (большая часть которых является либо заимствованными, либо представляют собой исконные формы (например, местоимения), выполняющие функции союзов под влиянием синтаксического строя индоевропейских языков [2]. Вместо союзов часто используются аффиксы, которые образуют не только словоформы, но и более сложные конструкции, например, являются средствами осложнения простого предложения или построения сложного.

Семантико-функциональный аспект татарских сравнительных конструкций, вводимых главным образом союзами, представлен в работе Р.М. Болгаровой и С.С. Сафоновой [1]. В данной статье рассмотрим на корпусных данных [4] особенности функционирования аффикса симилятива -ДАй. Статья состоит из двух частей: в первой рассматриваются типы аффиксальных цепочек, типичных для образований с аффиксом симилятива -ДАй, во второй - исследуется вопрос о том, какого типа конструкции вводятся при помощи данного аффикса.

Аффикс симилятива имеет четыре алломорфа (-дай/-дәй, -тай/-тәй) и позволяет создавать конструкции со значением сравнения и уподобления, при этом основа, к которой присоединяется аффикс симилятива, называет стандарт сравнения:

саламдай *сары чәч* 'волосы, желтые, как солома';

коштай *тиз оча* 'летит быстро, как птица'.

Рассмотрим структуру аффиксальных цепочек образований, содержащих аффикс -ДАй.

1. Именные основы.

Модель N+ SIM:

боздай салкын 'холодный, как лед';

шардай түгәрәк 'круглый, как шар';

коштай оча 'летит, как птица';

уттай яна 'горит, как огонь';

мамыктай йомшак 'мягкий, как вата'.

Модель N+PL+SIM:

дуслардай аералыштык 'расстались, подобно друзьям';

егетләрдәй көчле 'сильные, как парни';

кызлардай сылу 'стройная, как девушка'.

Основой может служить как имя нарицательное, так и собственное:

Себердәй киң 'широкая, как Сибирь';

Агыйделдәй ага 'течет, как река Белая';

Җәлилдәй батыр 'храбрый, как Джалиль'.

Стандартом сравнения может выступать как прецедентное имя, входящее в общий фонд знаний говорящего (см. примеры выше), так и вполне обычное, прагматически обусловленное ситуацией и контекстом:

Зөләйхадай уңган 'умелая, как Зулейха';

Кларадай чибәр 'красивая, как Клара';

Нәфисәдәй тыгыз тәнле 'с плотным телом, как Нафиса'.

Словоформа с аффксом -ДАй может содержать аффикс посессивности:

туганымдай 'как мой родственник';

туганнарымдай 'как мои родственники'.

Модель N+ATTR_GEN+ SIM:

аккошныкыдай 'как у лебедя';

сандугачныкыдай 'как у соловья';

дуңгызныкыдай 'как у свиньи';

су анасыныкыдай 'как у водяной';

атлетныкыдай 'как у атлета'.

Данный тип также может содержать аффикс множественного числа или посессивности:

хатын-кызларыныкыдай 'как у женщин';
кош баласыныкыдай 'как у птенцов'.

Основой может служить как имя нарицательное, так и собственное:
Никитинныкыдай 'как у Никитина';
Отеллоныкыдай 'как у Отелло';
Сократныкыдай 'как у Сократа'.

В конструкциях с генетивным атрибутивом стандарт сравнения почти всегда называет живое существо — человека, животное или мифологический персонаж.

Модель PN +SIM.

Наиболее часто встречаются образования с основой — личным местоимением:
миндәй 'подобно мне';
синдәй 'подобно тебе';
бездәй 'подобно нам';
сездәй 'подобно вам';
алардай 'подобно им'.

В качестве основы могут выступать и местоимения других семантических классов, в том числе и осложненные дополнительными аффиксами:
үзедәй 'подобно [ему] самому';
башкалардай 'подобно другим';
аныкыдай 'как его';
шулардай 'подобно этим/тем'.

Вопросительно-относительные местоимения также могут присоединять аффикс симилятива, но обычно они входят в состав более сложных номинаций:
кирәксез нәрсәдәй 'подобно чему-то ненужному';
әллә кемдәй 'подобно кому-то'.

Модель ADV+релятивизатор ГЫ+SIM:
баягыдай 'как недавно';
бүгенгедәй 'как сегодня';
кичәгедәй 'как вчера';
андагыдай 'как там'.

Глагольные образования включают два стандартных случая.

1. Модель V + аффикс перфекта -ГАн + SIM
карагандай 'словно смотрел';
кочаклагарндай 'словно обнимал';
ябыштыргандай 'словно приклеил'.

Глагол может стоять в отрицательной форме:
булмагандай 'словно не был';

алмагандай 'словно не брал'.

Основное значение данных образований— уподобление действию, которое могло бы состояться в прошлом:

Олы бүлмә кешеләр белән шыплап тулгандай булды. 'Большая комната словно наполнилась людьми'.

2. Модель V + аффикс потенциального будущего времени -Ыр/-мАс+ SIM:

ташлардай 'словно намереваясь' бросить;

чыгардай 'словно намереваясь выйти';

атардай 'словно намереваясь бросить'.

В том числе и с отрицательной основой:

сыймастай 'словно не вместится';

алмастай 'словно не возьмет';

кабатланмастай 'словно не повторится'.

Основное значение данных форм— уподобление потенциальному действию в будущем:

(1) *Кемгә казның зурысы, кемгә йомырка* ***салырдай*** *тавыклар кирәк*. 'Кому нужны гуси покрупнее, кому — куры, которые могут нести яйца'.

Данная форма часто выражает ирреальную форму желания или намерения:

(2) *Башына* ***орырдай*** *булам шул бөдрә чәчнең*. 'Иногда хочет дать по голове этой кудрявой девушке'.

Можно выделить три основных синтаксических типа двухчастных уподобительных конструкций – **именной**, ат**рибутивный и сентенциальный**. Они различаются тем, какая составляющая занимает позицию перед аффиксом симилятива (то есть выражает стандарт сравнения).

- **Именной тип** – стандарт сравнения выражен именной группой (*боздай салкын 'холодный, подобно льду'; дуслардай якын 'близкие, подобно друзьям';*):
- **Атрибутивный тип** – стандарт сравнения выражен образованием на -ГЫ или образованием с генетивным атрибутивом (*бүгенгедәй хәтерлим 'помню, словно это было сегодня', аккошныкыдай муены 'шея, как у лебедя'*).
- **Сентенциальный тип** – стандарт сравнения выражен клаузой.

Субъект подчинённой клаузы может совпадать с субъектом главной (3) или отличаться от него (4):

(3) *Зур соры күзләре дә аның, үз дәрәҗәсен яхшы* ***белгәндәй****, бик тыныч, салкын, горур карыйлар*. 'Ее большие глаза, словно зная ее цену, смотрели очень спокойно, холодно, гордо'.

(4) *Яныма килеп* ***яшәрдәй*** *берәр хатын-кыз булса, тормыш корыр идем, – ди.* 'Говорит: «Если будет женщина, которая согласится жить со мной, завел бы семью»'.

Таким образом, аффикс -ДАй может вводить симилятивные конструкции разной природы: именные, местоименные, наречные, глагольные. Признак, по которому производится уподобление, может иметь разную природу: оно может быть построено на знании объективных свойств предметов и явлений или иметь субъективную природу, часто осложненную модальными значениями. Образования с аффиксом -ДАй часто используются для образных сравнений реально далеких объектов, принадлежащих к совершенно разным классам.

Литература

1. Болгарова Р.М., Сафонова С.С. Сравнения в русском и татарском языках: семантико-функциональный и сопоставительный аспекты — Казань: Отечество, 2015. - 136 с.

2. Гузев В.Г., Бурыкин А.А. Общие строевые особенности агглютинативных языков // Acta linguistica Petropolitana. Труды ИЛИ РАН. - Т. 3. - Ч. 1. - СПб., 2007. - С. 109-117. // http://www.philology.ru/linguistics1/guzev-burykin-07.htm

3. Дыбо А.В., Шеймавич А.В. Автоматический морфологический анализ для корпусов тюркских языков // Филология и культура. - 2014. – № 3 (36). - С. 20-26.

4. Татарский национальный корпус “Туган тел” // http://corpus.antat.ru.

5. Хакимов Б.Э., Гильмуллин Р.А. Система морфологической разметки для корпуса татарского языка // Компьютерная лингвистика: научное направление и учебная дисциплина: сборник научных статей. Вып. 1 / В.И. Коваль (ответств. ред.); М-во образования РБ, ГГУ им. Ф. Скорины. – Гомель: ГГУ им. Ф. Скорины, 2010. - С.34-38.

6. Galieva A.M. The Locative Attributive in the Tatar Language: the Structure and Semantics // Proceedings of the International Conference“Turkic Languages Processing” TurkLang-2015 — Kazan, 2015. - Pp. 386-395.

УДК 811.512.145; 81'366.5

ГРАММАТИЧЕСКИЙ ПОРТРЕТ ТАТАРСКОГО ТЕКСТА И ЕГО СТИЛЕВАЯ ПРИНАДЛЕЖНОСТЬ

А.М. Галиева, Р.Р. Гатауллин
НИИ «Прикладная семиотика» АН РТ
amgalieva@gmail.com, ramil.gata@gmail.com

В работе исследуется вопрос о корреляции между частеречными характеристиками текста на татарском языке (соотношение слов различных частей речи в тексте) и стилевой принадлежностью текста. Получены

предварительные результаты по разработке методики автоматической классификации татарских текстов из корпусной коллекции по набору морфологических признаков.

Ключевые слова: *стиль, татарский язык, стилевая характеристика текста.*

В настоящее время одной из важных задач прикладной лингвистики является задача по разработке методов автоматической классификации массива текстов по стилевым и жанровым характеристикам, и эта задача решается для ряда языков, в частности, русского (см.[1-3]).

Для материала русского языка лексические и морфолого-синтаксические параметры функциональных стилей представлены в большом количестве исследований по стилистике, и эти данные становятся отправной точкой для прикладных разработок. Исследования по стилистике татарского языка ориентированы главным образом на анализ языковых средств выразительности и особенностей языка того или иного писателя. Поэтому остро стоит задача определения корреляции между частеречными и иными формализуемыми характеристиками текста и его стилевой принадлежностью.

Нами вручную подготовлена коллекция текстов разных функциональных стилей — официально-делового, публицистического (на примере новостных текстов), научного и художественного, и получены количественные данные по четырем типам критериев для каждого стиля по разработанным критериям. Таблица 1 представляет общие данные о текстах.

Таблица 1

Данные о текстовой выборке

№ п/п	Стиль текста	Количество документов	Общее количество лексических единиц в документе
1	Официально-деловой	21	12.310 (8.919)
2	Публицистический	15	5.181 (4.158)
3	Научный	15	46.348 (27.775)
4	Художественный	15	26.988 (20.954)

Задачей исследования являлось выделение различных видов корреляции различных характеристик данных в задачах классификации

данных. Так, например, в задаче классификации текстов по стилям и жанрам исследовались корреляции между внешними характеристиками текста (распределение слов по частям речи, средняя длина предложения, наличие/отсутствие средств осложнения предложения, количество заимствованных слов и др.) и жанровой и стилевой принадлежностью текста. Получены предварительные результаты по разработке методики автоматической классификации татарских текстов из корпусной коллекции по набору морфологических признаков.

Анализ основывался на измеряемых параметрах текстов (преимущественно морфологических) без подключения словарей. Для исследования были отобраны тексты четырех типов: официальные документы, новостные, научные и художественные тексты. Сформули-рованы ряд критериев для классификации текстов, а именно:

- морфологический критерий;
- критерий «Средняя длина слов и предложений в тексте»;
- критерий «Наличие числовых данных»;
- критерий «Типы знаков препинания».

В частности, нами предложено 20 частных морфологических критериев, например:

- количество существительных по отношению к общему количеству слов в тексте.
- общее количество глагольных форм по отношению к общему количеству слов в тексте.
- количество спрягаемых форм глагола (формы глагола, которые имеют аффиксы 1, 2, 3 лица ед. и мн. числа):
- количество форм 1 лица ед. числа (по отношению к общему количеству глаголов).
- количество форм 2 лица ед. числа (по отношению к общему количеству глаголов).
- количество форм 3 лица ед. числа (по отношению к общему количеству глаголов).
- количество форм 1 лица мн. Числа (по отношению к общему количеству глаголов).
- количество форм 2 лица мн. числа (по отношению к общему количеству глаголов) и т. п.

Таблица 2

Распределение грамматических параметров текстов

Количество по отношению к общему количеству слов	**Официально-деловой**	**Публицистический**	**Научный**	**Художественный**
Существительных	51,70%	41,97%	46,96%	37,23%
Глагольных форм	31,08%	25,77%	26,02%	35,20%
Форма 1 лица ед. числа	0,29%	0,05%	0,10%	1,63%
Форма 2 лица ед. числа	0,00%	0,02%	0,04%	0,66%
Форма 3 лица ед. числа	4,50%	9,50%	8,87%	12,51%
Форма 1 лица мн. числа	0,00%	0,27%	0,24%	0,71%
Форма 2 лица мн. числа	0,03%	0,03%	0,09%	0,43%
Средняя длина слово	4,76	4,80	4,20	5,33
Средняя длина предложений	12,23	10,13	9,20	12,83

Таблица 2 дает представление о распределении ряда параметров в зависимости от стилевой принадлежности текста. Ряд критериев, которые считаются надежными с точки зрения автоматического выделения стилей текстов русского языка [3], показали свою неэффективность для татарского языка. Например, общее соотношение имен существительных и глаголов в разных типах текстов на татарском языке отличается незначительно. Это связано с тем, что в официально-деловых, новостных и научных текстах активно используется такая специфическая форма глагола, как имя действия, что значительно повышает общую глагольность текстов. Поэтому критерий глагольности нами конкретизируется с учетом специфики спрягаемых и временных форм.

Средняя длина слов и предложений в текстах разных стилей также отличается несущественно и не может служить маркером типа текста.

Наибольшей спецификой обладают тексты официальных документов. Для автоматического выделения текстов официальных документов значимы следующие морфологические признаки:

1) очень низкая частотность ряда личных форм, а именно:

-полное отсутствие глаголов 2 лица единственного и множественного числа;

- полное отсутствие глаголов 1 лица множественного числа.

2) очень низкая частотность форм категорического прошедшего времени;

3) очень низкий процент форм определенного будущего и полное отсутствие форм неопределенного будущего времени;

4) очень низкая частотность частиц;

5) высокая частотность имен действий (в среднем в текстах официально-делового стиля их в два раза больше, чем в научных и новостных текстах и в шесть раз больше, чем в художественных текстах);

6) полное отсутствие форм личных местоимений 1 и 2 лица.

Установлено, что критерий средней длины слов и предложений в тексте не является релевантным, а также слабо релевантным является критерий наличия числовых данных. По критерию «Типы знаков препинания» выявлены зависимости признака «Количество вопросительных и восклицательных знаков по отношению к общему количеству знаков препинания в тексте» и некоторых стилей (официально-делового, научного и новостного). Полное отсутствие вопросительных и восклицательных знаков отличает тексты официально-делового стиля, относительно низкая частотность восклицательных знаков – научные и новостные тексты.

В ходе экспериментов пока не получила подтверждения гипотеза о фонемном различии типов текстов (первоначально предполагалось, что книжные стили содержат значительное количество заимствований разного рода — арабизмов, фарсизмов, русизмов, интернационализмов, обладающих фонетической спецификой, и предполагалось, что по данному критерию можно дифференцировать художественные и нехудожественные тексты. Отрицательный результат может быть связан с рядом факторов: 1) арабо-персидские заимствования давно освоены татарским языком и их частотность не зависит от типа текста; 2) эксперименты проводились только с текстами книжной речи, включая художественные; возможно, критерий позволяет дифференцировать тексты книжные и разговорные); 3) сама методика требует дальнейшего развития.

В целом, художественные тексты отличаются от нехудожественных гораздо большим морфологическим богатством. Важным критерием для разграничения официальных текстов от научных и новостных является использование форм глагольных времен (в новостных текстах, в целом, превалирует определенное прошедшее время, в научных — настоящее). Полученные результаты являются важным шагом для разработки

методики автоматического определения типа текстов из корпусной коллекции. На сегодняшний день существуют различные методы для автоматической классификации текстов по признаку стиля. Основные отличия предложенного нами подхода для автоматического определения стилевой принадлежности текстов на татарском языке от аналогичных методик анализа текстов на русском языке:

1) введение большого количество морфологических признаков (а не просто учет общей глагольности и адъективности текстов);

2) учет пунктуационных особенностей текста;

3) учет частотности числовых данных (факультативный признак, который в ряде случаев позволяет четко выделить тексты информационного содержания);

4) отказ от таких факторов, как средняя длина слов и средняя длина предложений в тексте.

Таким образом, разработана группа формальных грамматических, синтаксических семантических моделей, описывающих языковые единицы и их совокупности на материале корпусных данных; получены предварительные результаты по разработке методики автоматической классификации татарских текстов из корпусной коллекции по набору морфологических признаков.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проект № 15-07-09214).

Литература

1. Браславский П.И. Автоматическая классификация документов Internet по стилям: реализация макета: Доклад V рабочего совещания по электронным публикациям-EL-PUB-2000 - Новосибирск, Академгородок, ИВТ СО РАН. – 2000. – С. 21-23.

2. Емашова О.А., Мальковский М.Г. Функциональные стили русского языка и их влияние на задачу автоматического реферирования текстов //Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог. – 2007. // http://www.dialog-21.ru/digests/dialog2007/materials/html/25.htm

3. Поспелова А., Ягунова Е. Опыт применения стилевых и жанровых характеристик для описания стилевых особенностей коллекций текстов //Новые информационные технологии в автоматизированных системах. – 2014. – №. 17. С. 347-356.

УДК 81'37; 811.512.145

КОГНИТИВНАЯ СТРУКТУРА ОТРИЦАНИЯ В ТАТАРСКОМ ЯЗЫКЕ

А.М. Галиева, Д.Ш. Сулейманов
НИИ "Прикладная семиотика" АН РТ, Казань
amgalieva@gmail.com, dvdt.slt@gmail.com

В работе анализируется основные типы и способы выражения отрицания в татарском языке: раскрывается когнитивная, структурная и семантическая сложность отрицания как лингвистической категории. Авторы выделяют 4 базовых способа выражения отрицания, которые по-разному классифицируют и концептуализируют ситуации отрицания, и показывают, что ни одно из этих средств не является семантически элементарным.

Ключевые слова: *отрицание, татарский язык, семантика, лингвистические категории.*

Отрицание - одна из универсальных лингвокогнитивных категорий, дающих возможность адекватного описания реальности. Концептуальная сложность отрицания как категории определяется отсутствием непосредственной референции у отрицательных словоформ и предложений: отрицание становится возможным лишь на фоне "положительного" знания о ситуации: отрицание констатирует отсутствие предполагаемого объекта или его признаков или отсутствие связи между концептами.

Если русская отрицательная частица *не* универсальна по своей природе и может находиться практически перед любой словоформой (относительная лингвистическая простота выражения отрицания в русском языке), то виды отрицания в татарском языке являются специализированными и требуют того или иного грамматического (морфолого-синтаксического) представления (сортировка ситуаций, подпадающих под отрицание, обусловливает грамматическую избирательность). В работе анализируется способы выражения отрицания в татарском языке, раскрывается когнитивная, структурная и семантическая сложность отрицания как лингвистической категории. Работа выполнена на данных из Татарского Национального корпуса "Туган тел" (corpus.antat.ru).

Можно выделить следующие основные типы отрицания в татарском языке:

1. Именное отрицание.
2. Глагольное отрицание.

Глагольное отрицание формируется при помощи отрицательного форманта (аффикса) -мА, который присоединяется к глагольной основе. Глагольное отрицание используется только при отрицании процессуального признака, обычно (но не обязательно) локализованного на временной оси:

Сара мәктәпкә бармый. ‘Сара не идет в школу’.

В татарском языке мы выделяем 3 типа именного отрицания (при описании именного отрицания использована терминология из работы А.М. Певнова, 2007):

1) привативное - ситуация ‘референт не имеет *x’: койрык-**сыз*** ‘бесхвостый’;

2) абсентивное - ситуация ‘референт отсутствует’: *Суыткычта сөт **юк*** ‘В холодильнике нет молока’;

3) дезидентификационное - ситуация ‘референт не является *x’: Марат укытучы **түгел.*** ‘Марат не учитель’; *Кар ак **түгел*** ‘Снег не белый’.

Именное отрицание всех трех выделенных типов является семантически сложным, включая в свой состав, кроме собственно отрицания, семантические компоненты иной природы.

Привативное отрицание выражается при помощи аффикса абессива -сЫз и передает идею отсутствия объекта обладания или отсутствия дополнительного агенса в ситуации:

*машина**сыз** кеше* ‘не имеющий машины человек’,

*канат**сыз** кош* - ‘бескрылая птица’,

*яклаучы**сыз** калды* - ‘остался без защитника’.

В составе привативного отрицания могут быть выделены следующие компоненты значения: 'отрицание' + 'посессивность' + 'признаковость' (приписывание непроцессуального признака), толкование: ‘не имеющий *x*’.

Аффикс абессива присоединяется к существительным с конкретным и абстрактным (реже) значением. Антонимичной категории абессива является категория мунитатива (аффикс -лЫ), которая передает идею обладания:

*машина**лы** кеше* ‘человек, имеющий машину’.

Абсентивное отрицание выражается при помощи предикатива *юк* ‘нет’ и передает идею отсутствия объекта:

*Өстәлдә китап **юк.*** ‘На столе нет книги’.

Обычно абсентив выражает идею наблюдаемого или осознаваемого отсутствия, проверяемого эмпирически. Для выражения идеи несуществ-

вания обычно также используется слово *юк*, но предполагается не несуществование объекта как такового, а его отсутствие в мире (пресуппозиция *дөньяда* ***юк*** ‘нет в мире’).

Антонимом к слову *юк* является предикатив *бар* ‘есть, имеется, наличествует’:

Өстәлдә китап ***бар.*** ‘На столе есть книга’.

В составе абсентивного отрицания мы выделяем следующие семантические компоненты: 'отрицание' + 'наличие', где наличие представляет собой сложный концепт: 'существование' + 'локативность' (‘существование в месте *z’*).

Слово *юк* сочетается как с конкретными, так и абстрактными существительными.

Дезидентификативное отрицание выражается при помощи отрицательного предикатива (в татарских грамматиках - частицы) *түгел* и передает идею отсутствия связи между концептами, когда первый концепт обозначает референта в широком смысле (идея предметности), а второй - предмет или непроцессуальный признак (в прототипическом случае выраженный существительным или прилагательным).

Марат укытучы ***түгел*** ‘Марат не учитель’.

Дезидентификатив обычно выражает идею отсутствия связи между концептами - отрицание неглагольного (непроцессуального) предиката, актанта или сирконстанта высказывания:

Бу Марат ***түгел.*** ‘Это не Марат’.

Китап өстәлдә ***түгел.*** ‘Книга не на столе’.

Кар ак ***түгел***. ‘Снег не белый’.

В составе дезидентификатива могут быть выделены следующие компоненты значения: 'отрицание' + 'признаковость' (приписывание непроцессуального признака), толкование: ‘не есть *x*’.

Отрицательный предикатив *түгел* может использоваться также при отрицании глагольных словоформ (причастий, деепричастий, имен действий и даже финитных глаголов), но обычно это происходит при сентенциальных актантах (в случаях, когда актантное место заполняется инфинитивной конструкцией или целым предложением):

Эшкә соңга калу яхшы ***түгел***. ‘Нехорошо опаздывать на работу’.

В этом случае происходит транспонирование в позицию субъекта предиката имени, признак которого отрицается.

Когнитивная сложность отрицания в татарском языке выражается не только в том, что для понимания отрицания необходимо знание о стереотипной “положительной” ситуации, но и в том, что отсутствует универсальное средство выражения отрицания; имеется 4 базовых способа выражения отрицания, которые по-разному классифицируют и концептуа-

лизируют ситуации отрицания, и ни одно из средств не является семантически элементарным. Два типа отрицания имеют специальные маркеры для выражения положительных коррелятов, а два - не имеют таких маркеров. Широкое использование форм привативного и абсетивного отрицания и их положительных коррелятов позволяет татарскому языку обходиться без базового глагола, выражающего идею обладания (в татарском языке отсутствует эквивалент русского глагола *иметь*).

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проект №15-07-09214 А).

Литература

1. Певнов А.М. Об именном отрицании // Типология языка и теория грамматики: Материалы Международной конференции, посвященной 100-летию со дня рождения С.Д. Кацнельсона.- СПб.: Нестор-История, 2007. - С. 155-161.

2. Логический анализ языка. Ассерция и негация / Отв. ред. Н.Д. Арутюнова. - М.: Индрик, 2009. - 560 с.

УДК 81’33

О МОРФОЛОГИЧЕСКОЙ РАЗМЕТКЕ ТАТАРСКИХ ПРИЧАСТИЙ

А.М. Галиева
НИИ «Прикладная семиотика» АН РТ
amgalieva@gmail.com,
А.Р. Гатиатуллин
НИИ «Прикладная семиотика» АН РТ
agat1972@mail.ru

В работе рассматриваются предложения по морфологической разметке причастий татарского языка в электронном корпусе татарского языка. Поскольку ситуация с причастными формами является общей для всех тюркских языков, то это предложение применимо для морфологической разметки всех тюркских языков.

Ключевые слова: *морфологическая разметка, татарские причастия.*

Причастия, как одна из сложных функциональных форм глагола, традиционно выделяются в тюркских языках [1-9]. Так, все основные академические грамматики и базовые учебники по татарскому языку выделяют категорию «причастие» [10-12]. Тем не менее, ряд теорети-

ческих вопросов, связанных с причастиями, остается дискуссионным или же носит предварительный характер, так как для исследований не привлекались корпусные данные, представляющие реальное распределение единиц подобного типа.

В тюркских языках стандартным способом выражения атрибутивных отношений является следующее: *уточнение предшествует уточняемому*, т. е. определение предшествует определяемому. Например: *озын таяк 'длинная палка', Маратның китабы 'книга Марата'*, при этом определительные отношения могут выражаться только порядком слов, а не морфологически *таш йорт 'каменный дом'*.

Описательные грамматики татарского языка выделяют категорию причастия, которое, тем не менее, не имеет специфических средств выражения, а образуется от соответствующих:

а) временных форм (на –ГАН, -ЫР, -АчАК):

кешеләр кайткан 'люди вернулись' — кайткан кешеләр 'вернувшиеся люди';

кешеләр кайтыр 'люди вернуться' — кайтыр кешеләр 'люди, которые вернуться';

кешеләр кайтачак 'люди вернуться' — кайтачак кешеләр 'люди, которые вернуться'.

б) имени действия:

кайтучы кеше 'возвращающийся человек'.

Таким образом, по способу образования класс традиционных причастий не является однородным и делится на 2 типа:

а) словоформы, образованные на базе форм прошедшего и будущего времени;

б) словоформы, образованные на базе имени действия при помощи стандартного показателя деятеля (лица по действию) -чЫ:

Например: *эшче - эшләүче.*

По сути, причастия на -УчЫ — это обычные субстантивы, обозначающие деятеля, которые могут выступать в атрибутивной функции, подобно другим существительным.

В татарском языке форма имени действия является производящей основой не только для традиционно выделяемых форм настоящего времени, но и для атрибутивных форм узитатива (квалификатива): *карау – караучан 'любящий смотреть', эшләү- эшләүчән 'работящий'*, которые также сохраняют глагольные признаки отрицание, залог и раритив:

караучан — карамаучан;

караучан — каратучан;

караучан — караштыручан.

Причастие сохраняет основные признаки, характерные для финитного глагола. Он сохраняет общекатегориальное значение процессуальности и обладает рядом морфологических и синтаксических признаков, которые объединяют его с обычным глаголом [11]:

- стандартный аффикс отрицания: *караган – карамаган*;
- стандартные показатели залогов: *караган — каралган — каранган — караткан — карашкан кеше;*
- стандартные показатели раритива: *караштырган*.

Кроме того, причастие управляет падежами *баланы караган кеше 'человек, который присматривал за ребенком'*.

Признаки атрибутивных слов (прилагательного) у данных форм проявляются, когда они стоят в определенной синтаксической позиции — в препозиции к существительному;

При этом активный и пассивный залог различаются лишь по контексту:

укыган кеше (актив) 'человек, который читал' — укыган китап (пассив) 'книга которую читали', укыган еллар 'годы, в которые читали' (обстоятельственное значение).

Данные формы присоединяют аффиксы именной группы, формируя сентенциальные актанты и сирконстанты.

Например:

Кунаклар Туфан аганың актерлыкка ***укыганын*** *искә төшерделәр.*

Чит авыл мәктәбендә ***укыганда****, малайларның монда дус-ишләре булмады.*

Таким образом, тюркские причастия имеют ряд признаков, которое существенным образом отличают его от русского причастия. Тюркские причастия не имеют собственных грамматических маркеров, они не являются чисто атрибутивными формами и исполняют большой набор грамматических функций, в зависимости от структуры аффиксальной цепочки словоформы и синтаксического окружения, они имеют ряд признаков, общих с личными формами глагола, а также с именами действия. Кроме того, под термином «причастие» объединены формы с разным происхождением и разными грамматическими функциями.

Мы считаем, что введение этой категории по аналогии со славянскими языками создает 'искусственную' многозначность.

В отличии от славянских языков в тюркских языках категория причастия - это не морфологическая категория, а синтаксическая роль – в которой может выступать словоформа с рядом аффиксов: в татарском это -ГАн, -Ыр, -АчАк. Кроме того, имеется еще аналогичный аффикс -

АсЫ/ -ЫЙсЫ: барасы кеше, которое не вошло в категорию причастия в татарской грамматике.

Вопрос о природе причастия в тюркских языках давно привлекает внимание тюркологов. Так, Н.А. Баскаков отмечает, что в казахском языке «в отличие от русского, причастие является спрягаемой формой глагола...» [10]. Ряд исследователей считает, что категория «причастие» в тюркологии введено под непосредственным влиянием индоевропеистики [13].

Вместо категории «причастие» мы предлагаем говорить об атрибутивном употреблении форм глаголов или их субстантивации. Также при разметке электронных корпусов тюркских языков не выделять аффиксы –ГАН, -ЫР, -АчАк и –УчЫ как аффиксы категории причастия.

Литература

1 Мирзоев Г.И. Причастия в современном азербайджанском литературном языке (морфологические особенности). Автореф. канд. дисс. - Баку, 1965.

2 Мухтаров Дж. История развития причастных форм в узбекском языке. Автореф. канд. дисс. - Ташкент, 1971.

3. Ергалиев Т. Причастия в казахском языке. Автореф. канд. дисс. - Семипалатинск, 195.

4. Мусаев С.Ж. Структурно-функциональные типы и семантика причастных конструкций в киргизском языке. Автореф. докт. дисс. - М., 1982.

5. Хангишиев Дж.М. Причастия е кумыкском языке. Автореф. канд. дисс. - Баку, 1966.

6. Хисамова Ф.М. Причастия в современном татарском литературном языке. Автореф. канд. дисс. - Казань, 1970.

7. Сат Ш.Ч. Причастия в тувинском языке. Автореф. канд. дисс. - М., 1961.

8. Насыров Д.С. Причастия и его синтаксические функции в каракалпакском языке. Автореф. канд. дисс. - М., 1954.

9. Филиппов Г.Г. Причастия якутского языка: комплексное типологическое функционально-семантическое исследование. Автореф. докт. дисс. - Якутск, 1999.

10. Баскаков Н.А. Сопоставительная грамматика русского и казахского языковю Морфология — 1966.

11. Татар грамматикасы. – Т.2. – М.: ИНСАН, Казан: ФИКЕР, 2002. — 448 б.

12. Татарская грамматика: В 3т. Т.2. Морфология / Рос.АН, АН Татарстана, Институт языка, литературы и истории им. Г. Ибрагимова; Казан. науч. центр; Редкол.: М.З. Закиев и др. – Казань: Татарское книжное изд-во, 1993. – 397с.

13. Дубровина М.Э. О термине причастие в тюркских языках (на примере якутского языка) // «Модернизация и традиции»: XXVI Международная конференция по источниковедению и историографии стран Азии и Африки, 20-22 апреля 2011 г.: Тезисы докладов / Отв. ред. Н.Н. Дьяков и А.С. Матвеев. СПб.: Издательство РХГА 2011. С 357-358.

УДК 004.82+004.912+81.322.2

РОЛЬ ИМЕН ПРИЛАГАТЕЛЬНЫХ В ОПРЕДЕЛЕНИИ ТОНАЛЬНОСТИ ТЕКСТА

Б.Ж. Ергеш, А.А. Шарипбай, Г.Т. Бекманова
Евразийский национальный университет
имени Л.Н. Гумилева, Астана
b.yergesh@gmail.com, sharalt@mail.ru, gulmira-r@yandex.ru

В данной работе описывается лингвистический подход сентимент анализа текстов на казахском языке основанный на правилах.

Ключевые слова: *сентимент анализ, казахский язык, тональность, классификация по тональности.*

Сентимент анализ текстов на естественном языке один из быстро развивающих технологий обработки естественного языка. Сентимент анализ это процесс извлечения эмоции, мнений, настроений или отношения людей к продуктам, сервису, организациям и т.д. через анализ текстов, изображений или других источников. Сентимент анализ очень интересен компаниям, предприятиям для определения новых возможностей рынка [1]. Для людей эмоции и мнения играют важную роль в повседневной жизни и принятии решений. Системы семантического анализа широко используются в коммерческих и социальных сферах. С каждым днем растет количество блогов, рецензии, форумов, веб-страниц социальных сетей во всемирной сети. А ручная обработка этих информации становится невозможным. Для обработки такого рода информации используются разные методы лингвистические и машинного обучения.

Для английского языка созданы много лексических ресурсов и инструментов для определения тональности [1,2]. В последнее время ведутся исследовательские работы для других языков. В России с 2011 года в рамках этого семинара РОМИП организовываются тестирования систем сентимент анализа русскоязычных текстов [3,4,5]. Ведутся работы извлечения оценочных слов для русского и других языков [6,7]. Реализованы методы для извлечения признаков для улучшения результата сентимент анализа рецензии на турецком языке [8, 9, 10]. Сентимент анализ текстов на казахском языке все еще не изучен. Данная работа является введением и попыткой применить лингвистический подход вопределений тональности текста.

Основные методы сентимент анализа

Согласно работе [1] автоматический анализ тональности текстов на естественном языке осуществляется с помощью следующих основных методов: методы машинного обучения, лингвистические методы.

Системы анализа тональности на основе методов машинного обучения «обучаются» на коллекции заранее размеченных текстов. К таким методам относятся метод опорных векторов (SVM), логистическая регрессия, наивный байесовский классификатор, максимум энтропии, k ближайших соседей (k-NN) и др.

Лингвистические методы используют морфологический анализ, специально создаваемые словари оценочных слов и выражений и лингвистические правила [11].

Определение тональности текстов на казахском языке

Здесь рассматривается подход для классификации на два класса: позитивный (POS) и негативный (NEG). Для казахского языка был вручную создан словарь оценочных слов.

С помощью определения частей речи можно определить тональность текста. В казахском языке тональность тексту придают такие части речи, как прилагательное, глагол, наречие.

После проведения морфологического анализа из текста извлекаются слова и/или фразы содержащие оценочные слова. Как показывает исследования, прилагательные в основном определяют семантическую ориентацию (полярность) текста, а существительное является аспектом (объектом) обсуждения. Из извлеченных фраз можно определить полярность текста. Тональность оценочного слова может зависеть от контекста и предметной области. Также тональность может изменятсяили усиливаться в зависимости от наречия, глагола и союзов.

Согласно исследованиям с помощью этих словосочетаний можно определить тональность:

[ADJ]+[NOUN]
[ADJ]+[VERB]
[ADJ]+[VERV]+[отрицание]
[ADV]+[ADJ]

Правила для определения тональности

Таблица 1

Тональность словосочетия: [ADJ]+[NOUN]

ADJ	NOUN	Пример	Тональность
NEG	+	Нашар қалам (плохой карандаш)	NEG
POS	+	Жақсы кітап (хорошая книга)	POS

Таблица 2

Тональность словосочетия: [ADJ]+[VERB]

ADJ	VERB	Отрицание	Пример	Тональность
NEG	VERB	-	Жаман дайындалған (плохо приготовлен)	NEG
POS	VERB	-	Жақсы істейді (хорошо работает)	POS
NEG	VERB	емес/жоқ (не)	Жаман айтқан емес/жоқ (плохо не говорил)	POS
POS	VERB	емес/жоқ (не)	Жақсы болған емес/жоқ (не было хорошо) Дәмді болған жоқ (вкусно не было)	NEG
NEG	Отрицательная форма глагола	-	жаман болмайды (не будет плохо)	POS
POS	Отрицательная форма глагола	-	жақсы істемейді (не работает хорошо)	NEG

Таблица 3

Тональность словосочетия: [ADV]+[ADJ]

ADV	ADJ	Пример	Тональность
ADV	POS	ең әдемі (самый красивый), өте дәмді (очень вкусный),	POS
ADV	NEG	тым ащы (слишком горький).	NEG

Усилительные и сравнительные наречия придают дополнительный окрас тексту.К ним относятся: айрықша, аса, әбден, ең, ерекше, нағыз, өте, тым, азғана, әзер, әжептеуір и др. Например, ең әдемі(самый красивый), өте дәмді(очень вкусный),тым ащы(слишком горький). Это может использоваться, если тексты будут классифицироваться на групп, более двух(позитивные, негативный).

На рисунке 1 приведен пример работы программы для определения тональности текста. Положительные цифры (1, 3) определяют позитивную, а отрицательные негативную (-1) тональность. Тональность всего текста в целом можно определить как арифметическую середину лексических тональностей составляющих его единиц (предложений) и правил их сочетания.

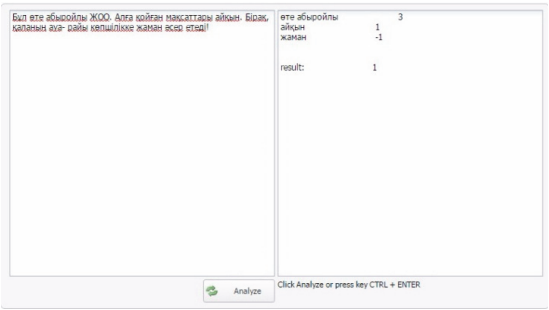


Рисунок 1. Фрагмент работы программы

Также в определений тональности может повлиять союзы между словами или предложениями. К примеру если встречаются соединительные союзы, то тональность не изменеятся. А если между словами или предложениями стоят разделительные или противительные то семантическая ориентация меняется в противоположеную. Например, бұл кәмпит тәтті, бірақ қатты екен (эти конфеты вкусные, но твердые).

В данной работе мы представили применение словаря и лингвистических правил для определения тональности текстов на казахском языке. Данная работа является введением в данную область и будет в будущем изучен. Также планиурется применение методов машинного обучения.

Литература

1. BingLiu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

2. Pang B., Lee L. Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval. Now Publishers. 2008.

3. Лукашевич Н.В. Четвёркин И.И. Открытое тестирование систем анализа тональности на материале русского языка.

4. Chetviorkin I., Braslavskiy P., Loukachevich N. Sentiment Analysis Track at ROMIP 2011. In Proceedings of International Conference Dialog-2012, volume 2, 2012. pp. 1-14.

5. Chetvirokin I., Loukachevitch N. Sentiment Analysis Track at ROMIP 2012. In Proceedings of International Conference Dialog-2013, volume 2, 2013. pp. 40-50.

6. Chetviorkin I., Loukachevitch N. Extraction of Russian Sentiment Lexicon for Product Meta-Domain In Proceedings of COLING 2012, 2012. pp. 593-610.

7. Steinberger J., Lenkova P., Ebrahim M., Ehrmann M., Hurriyetogly A., Kabadjov M., Steinberger R., Tanev H., Zavarella V., Vazquez S. Creating Sentiment Dictionaries via Triangulation. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT, 2011, pp. 28-36.

8. Akba F., Uçan A., Sezer EA., Sever H. Assessment of feature selection metrics for sentiment analyses: Turkish movie reviews. 8th European Conference on Data Mining 2014, 180-184]

9. Ezgi Yıldırım, Fatih Samet Çetin, Gülşen Eryiğit, Tanel Temel. (2014). The Impact of NLP on Turkish Sentiment Analysis. In Proceedings of the TURKLANG'14 International Conference on Turkic Language Processing, Istanbul, 06-07 November 2014.

10. Gülşen Eryiğit, Fatih Samet Çetin, Meltem Yanık, Tanel Temel, İlyas Çiçekli. TURKSENT: A Sentiment Annotation Tool for Social Media. In Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, ACL 2013, Sofia, Bulgaria, 4-9 August 2013.

11. Taboada M., Brooke J., Tofiloski M., Voll K., Stede M. Lexicon-based methods for Sentiment Analysis. Computational linguistics, 37(2) 2011, pp. 267-307

УДК 81'37

ФЕНОМЕН ЗНАЧЕНИЯ В КОГНИТИВИСТИКЕ

Г.В. Колпакова
Казанский федеральный университет
Galina.Kolpakova@kpfu.ru

Статья посвящена анализу подходов к исследованию значения в структурной и когнитивной лингвистике. В статье рассматриваются проблемы структурирования словаря, ментального лексикона на примере концепций семантики современных немецких исследователей.

Ключевые слова: значение, семантика, ментальный лексикон, структурная лингвистика, когнитивная лингвистика

Феномен значения является одним из самых сложных вопросов лексической семантики. В отечественной лингвистике лексическую семантику и предмет ее изучения – значение языковой единицы традиционно относят к области исследования лексикологии. В западноевропейской лингвистике феномен значения является объектом изучения семантики, сформировавшейся в самостоятельное направление исследования в отличие от лексикологии, интенсивно развивающейся лишь в последнее десятилетие. Свидетельством тому является выход в свет коллективной монографии ведущих немецких лингвистов «Lexikologie. Lexikology. Ein internationals Handbuch zur Natur und Struktur von Woertern und Wortschaetzen“ [1].

Лексикология, рассматриваемая как теория и практика структурирования словаря, представляет собой ввиду комплексности объекта изучения диффузное явление. Определяя в качестве объекта исследования «структурные группировки в словаре», П.Р.Лутцайер подчеркивает возможность структурирования словаря на локальном уровне, т.е. на уровне лексических единиц, и на глобальном уровне, т.е. на уровне релевантных разделов словаря, и предлагает три аспекта рассмотрения словаря: *лексис* – словарь как наиболее полное собрание слов естественного языка, *ментальный лексикон* как ментальный процесс накопления лексических единиц, лексической информации у индивида и *лексикон* – словарь как динамический накопитель информации, находящийся под воздействием грамматических процессов. Указанные три аспекта рассмотрения словаря тесно взаимосвязаны. Естественной логической связью является соотношение между *лексисом* как совокупностью слов, обладающей внутренней структурой, и *лексиконом,* в котором грамматические процессы оказывают регулирующее воздействие на единицы при образовании высказываний. Рассмотрение словаря как лексиса имеет системную ориентацию, трактовка словаря как лексикона характеризуется процессуальной направленностью. *Ментальный лексикон* как связующее звено между лексисом и лексиконом воплощает обе характеристики: системную и процессуальную направленность. Ментальный лексикон индивида востребован при речепроизводстве. Словари, рассматриваемые как *лексис* и *ментальный лексикон,* имеют краткосрочную и долгосрочную подвижность. К краткосрочной подвижности относится активизация различного вида локальных структур, к долгосрочной подвижности относятся глобальные изменения в словаре в течение определенного временного периода, обусловленные действием внешних факторов. Эта изменчивость, гибкость словаря является выражением внутренней, системно заданной динамики в словаре. Лексикология занимается не только словарем, но и его элементами, лексическими единицами. Вопрос о критериях идентификации, классификации словарей и лексических единиц исследователь рассматривает в качестве наиболее сложного вопроса лексикологии. Слова являются наиболее пригодными лексическими единицами для категоризации индивидом получаемой извне информации в силу своей интуитивной доступности и связанности с представлениями о жизни, природе и человеке [2].

В данной программной статье П.Р.Лутцайера, открывающей издание «Lexikologie. Lexikology”, отсутствует разработка понятия «ментального лексикона», являющегося одним из основных понятий в когнитивной лингвистике. П.Р.Лутцайер ограничивается лишь отдельными краткими

характеристиками ментального лексикона, указывая, например, что лексикология соприкасается с психолингвистикой, клинической лингвистикой и когнитивной лингвистикой при трактовке ментального лексикона. Упоминая о включениях отдельных значений слова в пространстве лексикона (Lexikoneintraege), П.Р.Лутцайер считает полезным применение понятия «фрейма», не давая его подробного описания. Неясным остается различие между лексиконом и ментальным лексиконом в концепции П.Р.Лутцайера. По мнению исследователя, лексикон – это словарь, рассматриваемый как динамический накопитель информации, на который оказывают регулирующее воздействие грамматические процессы. Но и ментальный лексикон предлагается рассматривать как ментальный процесс накопления лексических единиц, лексической информации у индивида, при этом дефиниции ментального процесса и ментального лексикона в работе отсутствуют. Возможно, что понятие «лексикона» введено с целью обозначения сферы действия грамматических процессов, имеющих межиндивидуальный характер, в отличие от «ментального лексикона», трактуемого как принадлежность индивида. Основное внимание П.Р.Лутцайер уделяет третьему аспекту рассмотрения словаря – *лексикону*, указывая, что подобная модель словаря позволяет охарактеризовать словари естественных языков, описать их системно-структурные свойства. На деле это означает не что иное, как рассмотрение словаря со структуралистских позиций. Но в этом случае подвергается сомнению динамическая природа лексикона в трактовке П.Р.Лутцайера, рассматривающего лексикон (словарь) как динамический накопитель информации. Это впечатление о трактовке словаря с позиций структурализма в концепции П.Р. Лутцайера еще более усиливается при анализе предложенной им интерпретации лексических единиц. Исследователь отмечает комплексный характер лексических единиц, состоящих из различных частей, либо лексическая единица представлена как выделенный элемент в объемном лексическом пространстве, где наряду с данной единицей находятся другие равнозначные лексические единицы. Но признание комплексного характера лексической единицы, семантика которой раскладывается на составные части, и возможность противопоставления лексической единицы другим лексемам в словаре является основным положением структурной лингвистики [3]. Проанализированная концепция П.Р.Лутцайера представляет собой пример того, как синтезируются структурализм, ставший со времен Соссюра классическим направлением в лингвистике, и когнитивная лингвистика. Декларируя когнитивную направленность своего исследования, лингвисты на деле применяют хорошо известные, испытанные методы анализа.

Повышенный интерес исследователей к словарю в последние десятилетия объясняется становлением когнитивной лингвистики и интенсификацией исследований лексической информации в компьютерной лингвистике. Классические вспомогательные средства лексикологических теорий – словари и корпуса текстов в значительной степени облегчают практическую работу лексикологов, но, как справедливо отмечает П.Р.Лутцайер, требуют ввиду масштабности проводимых исследований значительных временных затрат [4].

Иной подход к решению проблем семантики с психологических позиций обусловил развитие в лингвистике направления исследования, объектом изучения которого являются процессы понимания актов актуального говорения, получившие название «концептуальных начал». В основе данной семантической теории лежит представление о том, что между уровнем знаков и значениями отдельных слов имеется универсальный, независимый от конкретных языков уровень концептов. Приверженцы этого направления пытаются экспериментальным путем овладеть этими концептами как основополагающими или типичными [5]. Концептуальные подходы к изучению семантики слова относятся к парадигме когнитивной семантики. Направление «концептуальные начала» в немецкой лингвистике представлено семантической теорией М.Шварц [6], характерной особенностью которой является синтез психологического и когнитивно-семантического подходов к анализу слов, концептуальных структур, ментального лексикона. В работе М.Шварц анализируется краеугольная проблема семантики – соотношение знания и значения. Когнитивная семантика в интерпретации М.Шварц рассматривает значения как ментальные единицы, зафиксированные в лексиконе и актуализирующиеся в процессе переработки языковой информации. Таким образом, когнитивная реальность семантических феноменов находится, по мнению исследователя, на переднем крае семантических исследований. Следовательно, модели и теории когнитивной семантики должны быть совместимы с результатами исследований когнитивного процесса, процесса познания в целом. Основные положения когнитивной семантики затрагивают организацию и репрезентацию семантического знания в памяти индивида и его актуализацию в актуальных процессах переработки информации. Целью исследования при концептуальном подходе является описание ментального уровня значений слов как субсистемы познания, а также выявление отношения между значениями и внеязыковыми референтами при семантической интерпретации. Для когнитивной теории семантики анализ языковых значений, утверждает М.Шварц, в отрыве от аспектов структуры и процессуальности концептуальной системы невозможен. Это утверждение не означает

тождественности семантического и концептуального уровней. Различие между концептуальным и семантическим уровнями подтверждается исследованиями отношения между значениями в лексиконе и знанием о мире в памяти индивида, а также между значениями в лексиконе и контекстуальной информацией при интерпретации значений [7].

Концепты рассматриваются в когнитивной лингвистике как структурные строительные элементы человеческого познания. Концепты представляют собой единицы организации и хранения результатов нашего познания. Концептуальные единицы репрезентируют наше знание о мире в более абстрактном формате и делают возможной эффективную переработку внешних раздражителей путем категоризации. Наше познание строится на различных ментальных субсистемах, организованных в соответствии с разными принципами (моторное, тактильное, пространственно-воспринимающее знание). В процессах восприятия и познания мы соотносим информацию из этих различных систем знания, иными словами, мы интегрируем информацию в целостные единицы. Этот процесс возможен благодаря абстрактному уровню концептуальной системы, где информация хранится во внемодульном виде. Создается принципиальная возможность перевода одной специфической модульной репрезентации в другую: мы можем, например, репрезентировать моторные действия образно или вербально. Концептуальная система знаний накапливает, хранит в форме когнитивных категорий наше общее знание о мире. Концептуальная структура представляет собой сетку онтологических категорий для переработки и классификации раздражителей из внешнего мира. Она выполняет роль посредника между языком и миром: так как мы не имеем непосредственного доступа к репрезентациям окружающего мира, этот доступ осуществим лишь через наши внутренние репрезентации. Экстенсия, иными словами, референты языковых единиц должны постоянно идентифицироваться и описываться при помощи опосредующего уровня концептуальной структуры. При интерпретации языковых выражений концептуальная система играет огромную роль, так как значения перерабатываются в актуальном процессе на основе их вариабельности и контекстуальной зависимости путем отсылки к концептуальному знанию [8].

М.Шварц отмечает, что в когнитивной семантике образовались два теоретических подхода, которые различным образом освещают проблему хранения единиц в лексиконе и интерпретации этих единиц. Оба направления: одноуровневая теория и многоуровневая теория претендуют на достоверность отображения реальности в когнитивном процессе. Если одноуровневая теория не предусматривает разделения знания о мире и языкового знания при репрезентации значений в ментальном лексиконе,

то многоуровневая теория постулирует различие между семантическими и концептуальными репрезентациями [9]. Современное состояние исследований в когнитивистике характеризуется дискуссией по поводу холистической (целостной) и модульной концепциями семантики.

Литература

[1] *Lexikologie. Lexikology. Ein internationales Handbuch zur Natur und Struktur von Woertern und Wortschaetzen* [hrsg/edited by D.Allan Cruse, Fr.Hundsnurscher, M.Job, P.R.Lutzeier]. Berlin, New Jork: Walter de Gruyter, 2002. 1.Halbband/ Volume 1. 942 S.

[2] *Lutzeier P.R.* Der Status der Lexikologie als linguistische Disziplin // Lexikologie. Lexikology. Ein internationales Handbuch zur Natur und Struktur von Woertern und Wortschaetzen [hrsg/edited by D.Allan Cruse, Fr.Hundsnurscher, M.Job, P.R.Lutzeier]. Berlin, New Jork: Walter de Gruyter, 2002. 1.Halbband/ Volume 1. S. 1-8.

[3] *Ebend.* S. 6-8.

[4] *Ebenda. S. 6,8.*

[5] *Lange Kl.P.* Die Behandlung der Wortbedeutung in der Geschichte der Sprachwissenschaft // Lexikologie. Lexikology. Ein internationales Handbuch zur Natur und Struktur von Woertern und Wortschaetzen [hrsg/edited by D.Allan Cruse, Fr.Hundsnurscher, M.Job, P.R.Lutzeier]. Berlin, New Jork: Walter de Gruyter, 2002. 1.Halbband/ Volume 1. S. 242-243.

[6] *Schwarz M.* Konzeptuelle Ansaetze II: Einebenen – Ansatz vs. Mehrebenen – Ansatz // Lexikologie. Lexikology. Ein internationales Handbuch zur Natur und Struktur von Woertern und Wortschaetzen [hrsg/edited by D.Allan Cruse, Fr.Hundsnurscher, M.Job, P.R.Lutzeier]. Berlin, New Jork: Walter de Gruyter, 2002. 1.Halbband/ Volume 1. S. 277-284.

[7] *Ebenda.* S. 277.

[8] *Ebenda.* S. 277-278.

[9] *Ebenda.* S. 278.

УДК 81’33

О РЕАЛИЗАЦИИ СИСТЕМЫ МОРФОЛОГИЧЕСКОЙ РАЗМЕТКИ КРЫМСКОТАТАРСКОГО ЭЛЕКТРОННОГО КОРПУСА

Л.Ш. Кубединова
Крымский федеральный университет
kubedinova@gmail.com
А.Р. Гатиатуллин
НИИ «Прикладная семиотика» АН РТ
agat1972@mail.ru

В работе рассматриваются система морфологической разметки для электронного корпуса крымскотатарского языка и программа морфологического анализа крымскотатарских словоформ, использующая в своей работе

эту систему тэгов. Данная система разработана на основе тэгов, используемых при разметке электронного корпуса татарского языка “Туган тел”.

Ключевые слова: *морфологическая разметка, крымскотатарский электронный корпус.*

В настоящее время для многих языков тюркской языковой группы создаются текстовые электронные корпусы. Такие корпусы существуют для турецкого, татарского, казахского, башкирского, тувинского и других тюркских языков [1-5]. Все разработчики этих корпусов встречаются с одними и теми же проблемами и начинают проходить один и тот же путь создавая свои системы аннотирования корпусов. Хотя структурная близость тюркских языков позволяет создавать общую базу компьютерных моделей и программных модулей для обработки текстов на тюркских языках.

Аналогичные задачи встали и перед разработчиками корпуса крымскотатарского языка. Несмотря на то, что корпус электронных текстов крымскотатарского языка [6-7] был создан еще в 2006 году, у него до настоящего времени не было системы морфологической разметки. Авторами данной статьи проведена работа по созданию системы морфологической разметки крымскотатарского электронного корпуса.

Реализация этой задачи проходила в несколько этапов:

- сравнительный анализ системы разметок (тэгов) для других тюркских языков, в частности татарского и турецкого;
- предложение морфологических тэгов отсутствующих в сравниваемых языках;
- разработка системы морфотактических правил крымскотатарского языка;
- подготовка словаря крымскотатарских основ с морфонологической разметкой;
- заполнение информации в базу данных программного комплекса для многофункциональной модели тюркской морфемы.

Рассмотрим эти этапы. Нами было проведено сравнение таблицы аффиксальных морфем татарского и крымскотатарского языков, а также грамматических категорий выражаемых этими морфемами.

Проведенный сравнительный анализ показал, что в таблице крымскотатарских морфем есть целый ряд морфем, которые отсутствуют в таблице для татарского языка [8]. Список таких морфем приведен в таблице 1. Для всех этих морфем предложен свой набор тэгов.

Таблица 1

	Морфема	Алломорфы	Тэги
1	-нен		INST
2	-джА	-джа, -дже, -ча, -че	COMP
3	-мАктА	–макъта/ -мекте	PRES2
4	-Г[ъ]АйдЫ	-гъайды, -гейди, -къайды, -кейди	OPT
5	-мА+Й+Ып	-майып, -мейип	ADVV_NEG_1
6	-мАдАн	-мадан, -меден	ADVV_NEG_2
7	-мАлЫ	-малы, -мели	DEB
8	-АрАк[ъ]	-аракъ, - ерек, - яракъ, -йерек	ADVV_SIM
9	-ГЪАн+джА[къ]	-гъанджа(къ), -гендже(к), -къанджа(къ), -кендже(к)	ADVV_SUCC

Набор тэгов, добавленный для аффиксальных морфем крымскотатарского языка, был заполнен в базу данных программного комплекса для описания многофункциональной модели тюркской морфемы. Фрагмент интерфейса по заполнению модели для крымскотатарских морфем приведен на рис.1.

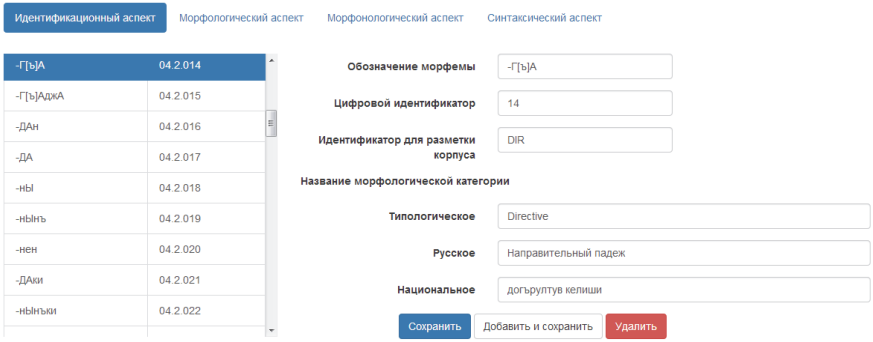


Рис.1. Форма для заполнения идентификационного аспекта для морфем крымскотатарского языка

Следующим этапом по созданию программы морфологической разметки является разработка системы морфотактических правил крымскотатарского языка. В нашей модели эти правила представлены правилами двух видов. Первый вид – это правила сочетания алломорфов крымскотатарских аффиксальных морфем. Второй вид – правила сочетания алломорфов аффиксальных морфем с корневыми морфемами.

После описания этих правил, они также были заполнены в базу данных многофункциональной модели тюркской морфемы для крымскотатарского языка. Всего получили 270 алломорфов для крымскотатарских аффиксальных морфем, а также 54 типа крымскотатарских основ в зависимости от присоединяемых аффиксальных алломорфов.

Следующий этап представлял собой подготовку словаря крымскотатарских основ и присвоение всем этим основам соответствую-щих морфонологических типов.

После завершения подготовки ресурсов все они были загружены в базу данных программного комплекса “Многофункциональная модель тюркской аффиксальной морфемы”. Пример работы модуля морфологического анализа для крымскотатарских представлен на рис.2. На этом рисунке показано, что на вход программы подается текст на крымскотатарском языке: *Япракълар тёкюльди. Отлар-оленлер сарарды, солдылар.* В нижней части окна выдаются результаты анализа.

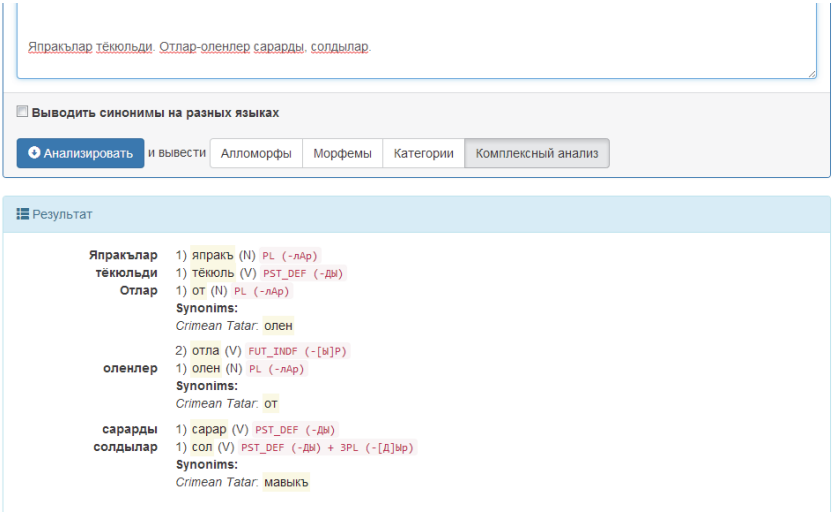


Рис. 2. Пример анализа крымскотатарского текста

Заключение

В статье дается описание процесса создания системы морфологической разметки для крымскотатарского электронного корпуса. Как показывает опыт работы над этой системой, конструктивным и продук-

тивным представляется использование для решения этой задачи многофункциональной многоязычной модели тюркских морфем в качестве одного из основных элементов технологии. Это позволяет в одной модели аккумулировать информацию по нескольким языкам и эффективно внедрять накопленный опыт в процессе разработки программных модулей для других языков тюркского семейства.

Литература

1. METU Turkish Corpus [Электронный ресурс]. URL: http://www.ii.metu.edu.tr/ content/metu-turkish-corpus-access-page

2. Tatar Corpus. [Electronic resource]. URL: http://web-corpora.net/TatarCorpus/ search/ ?interface_language=ru, August 2014.

3. Kazakh Corpus. [Electronic resource]. URL: http://kazcorpus.kz/klcweb/, August 2014.

4. Bashkir Corpus. [Electronic resource]. URL: http://mfbl.ru/bashkorp/korpus, August 2014.

5. Tuvinian Corpus. [Electronic resource]. URL: http://www.tuvancorpus.ru/, August 2014.

6. *Кубединова Л.Ш. Гарабик Р.*Лингвистический корпус крымскотатарского языка: перспективы развития // Труды Казанской школы по компьютерной и когнитивной лингвистике, TEL'2014. — Выпуск 16. — Казань, 2014. — С. 124-127.

7. *Kubedinova Lenara* Corpus Linguistics: Studies in Crimean Tatar Language / Kubedinova Lenara, Radovan Garabik // TURKLANG'14 International Conference on Turkic Language Processing", 6-7 November 2014 – http://turklang.itu.edu.tr/invited-speakers.htm.

8. Proceedings of the International Conference "Turkic Languages Processing: TurkLag-2015". – Kazan: Academy of Sciences of the Republic of Tatarstan Press, 2015. C.331-337.

9. *Кубединова Л.Ш., Гатиатуллин А.Р.* Морфологическая разметка крымскотатарского электронного корпуса // Proceedings of the International Conference "Turkic Languages Processing: TurkLag-2015". – Kazan: Academy of Sciences of the Republic of Tatarstan Press, 2015. C.331-337.

УДК 81'32

О НЕКОТОРЫХ СЛОЖНОСТЯХ МОРФОЛОГИЧЕСКОЙ РАЗМЕТКИ КАЗАХСКИХ ТЕКСТОВ

Макажанов Айбек Омиржанович,
Султангазина Айтолкын Нурлановна
National Laboratory Astana, Астана, Казахстан
aibek.makazhanov@nu.edu.kz, aitolkyn.sultangazina@nu.edu.kz

Морфологическая разметка со снятием омонимии заключается в сопоставлении каждого токена (единицы разбора) *единственному* разбору или парсу (тройке значений <лемма, часть речи, набор граммем>) в соответ-

ствии с заранее определенной схемой разметки. Неотъемлемой частью схемы разметки являются критерии токенизации, набор правил, определяющий что и когда считать единицей разбора. Как правило, один токен соответствует одному орфографическому слову и наоборот, однако в некоторых случаях это тождество может быть нарушено. В данной статье на примере некоторых тюркских языков описываются два подобных случая: 1) одно орфографическое слово состоит из нескольких токенов; 2) один токен состоит из нескольких орфографических слов. Также приводятся примеры реализации соответствующих критериев токенизации в одной из существующих схем морфосинтаксической разметки казахского языка.

Ключевые слова: *морфологическая разметка, токенизация, тюркские языки, казахский язык*

1. Введение

В настоящее время в вычислительной лингвистике языковые ресурсы для тюркских языков представлены корпусами татарского [1], турецкого [2], башкирского [3], казахского [4, 5] и других языков. В качестве минимума лингвистической аннотации данные корпуса имеют морфологическую разметку с частично или полностью снятой омонимией. Грамматическая информация словоформ кодируется типичным для агглютинативных языков способом – коды граммем перечисляются за леммой в порядке аффиксации, в некоторых случаях с явной кодировкой нуль-морфем (например, именительного падежа [1, 4, 5] или глагола связки [4]). В некоторых схемах разметки также сохраняются алломорфы [1, 5], которые могут быть использованы при снятии неоднозначности [6].

Насколько нам известно, в большинстве случаев (за исключением [1, 4]) негласно считается, что одно орфографическое слово всегда соответствует одному токену и наоборот, т.е. единица морфологического разбора не может содержать пробелов и что словоформа (без пробелов[1]) не может быть разбита на несколько отдельных парсов. Назовем такой критерий токенизации «1=1» (один к одному). Мы считаем, что для (некоторых) тюркских языков критерий «1=1» недостаточен для полной и адекватной разметки на морфологическом и синтаксическом уровнях. В таблице 1 перечислены некоторые примеры неприменимости данного критерия, а также приведены альтернативные критерии токенизации.

[1] Насколько нам известно, при наличии прочих разделителей, в частности дефиса, в некоторых случаях, словоформа разбивается на отдельные парсы. Например, в корпусах казахского [4, 5] частицы, примыкающие к слову посредством дефиса, размечаются отдельно.

Таблица 1

Примеры неприменимости критерия токенизации «1=1»

№	Пример	Альтернатива
1	орфография: личные окончания пишутся раздельно; сравните: *тув. сен* {*келир сен*} [7, с. 24] – ты придешь – *каз. сен* {*келерсін*}	«1+=1» (многие к одному)
2	запись слитного произношения: *каз.* {*барғасын*} = {|*барған*|+|*соң*|} – как только дойдешь / раз уж пошел	«1=1+» (один ко многим)
3	деривация: *тур. mavi* {{*arabada*}$_1$+{*kiler*}$_2$} *uyuyorlar* [8] – те, кто в синей машине, спят – синей {{в машине}$_1$+{которые}$_2$} спят	

Если при разметке руководствоваться только критерием «1=1», то на морфологическом и синтаксическом уровнях возникают сложности с кодированием морфем и синтаксических отношений. Например, какую часть речи присваивать орфографически разделенным личным окончаниям в тувинском (таблица 1, пример 1)? Даже если в данном случае в виде исключения не указывать часть речи, ограничившись кодом граммемы, или использовать «искусственную» часть речи, то каким синтаксическим отношением связывать эти токены? Далее, как в казахском анализировать словоформы вида *глагол*+FAcЫн (таблица 1, пример 2)? В частности, какую грамматическую категорию присвоить заключительному «аффиксу»? Наконец, как в турецком (и других тюркских языках) без разделения словоформы показать (на синтаксическом уровне), что в предложении *mavi arabadakiler uyuyorlar* (таблица 1, пример 3) синей является машина, а не спящие в ней люди?

Мы считаем, что перечисленные сложности морфосинтаксической разметки следует решать путем использования дополнительных критериев токенизации, нежели «подстраиванием» списков грамматических категорий и синтаксических отношений под ситуацию. В следующих двух разделах на примере казахского языка мы опишем применимость двух таких критериев, в частности: 1) «1+=1» – несколько орфографических слов составляют одну единицу разбора; 2) «1=1+» – одно орфографическое слово составляет несколько единиц разбора.

2. Токенизация «1+=1»

В нашем понимании, критерий «1+=1» следует применять к последовательности орфографических слов, составляющих одну

языковую единицу на смысловом и синтаксическом уровне, по одному из следующих признаков: 1) как минимум один член последовательности не представляется возможным описать тройкой значений <лемма, часть речи, набор граммем>; 2) как минимум одну пару членов последовательности не представляется возможным связать синтаксическим отношением[2].

По первому признаку мы применяем критерий к словосочетаниям казахского языка один или несколько членов которых потеряли свое лексическое значение и, не будучи служебными словами, используются только в составе данных словосочетаний. Например, в составных глаголах *міз бағу* (часто в отрицательной форме: *міз бақпау – не обращать внимания*) и *місе тұту* (довольствоваться) первые члены не встречаются в словарях вне соответствующих словосочетаний, и, насколько нам известно, не будучи служебными словами, в современном языке собственных лексических значений не имеют. В данном случае, мы не можем точно классифицировать эти слова по частям речи[3], и, следовательно, не можем произвести разметку словосочетания пословно. Однако, применив критерий «1+=1», мы можем рассмотреть в качестве токена словосочетание целиком. Например, форму *міз бақпайды* можно сопоставить тройке значений <лемма={*міз бақ}*, часть речи=*глагол не переходный*, набор граммем={*отрицание*; *аорист*; *третье лицо; единственное число}*>, что в схеме разметки открытого синтаксического корпуса казахского языка [4] и морфо-анализаторах тюркских языков проекта Apertium [11] соответствует записи {*міз бақ}*<v><iv><neg><aor><p3><sg> (мы лишь добавили фигурные скобки).

На момент написания статьи безусловная применимость второго признака (как в случае с тувинскими личными окончаниями) к конструкциям казахского языка нами не установлена. Сейчас по данному признаку диагностируются аналитические отрицания вида {*глагол*+ГАн *жоқ/емес*}, и по совокупности признаков – составные глаголы с русским инфинитивом и глаголом *ет* (например, *обжаловать етті – обжаловал*, дословно: *сделал обжаловать*).

Следует также отметить, что и в первом и во втором случае рассматриваемые конструкции могут включать в себя частицы *Ғана* и *ДА*, например: *місе де тұтты* (и довольствовался), *міз де бақпады* (даже не обратил внимания), *барған ғана/да жоқ* (только/даже/и не пошел), *обжаловать қана/та*

---

[2] Понятие «синтаксическое отношение» весьма условно и в некоторых схемах синтаксической разметки существует «общее» отношение, которое используется, в том числе, и для подобных случаев. Например, отношение «dep» в схеме Universal Dependencies (UD) [9] (применимо к казахскому [4, 10]).

[3] Так как сами глаголы в данных сочетаниях переходные (*тұту* в значении держать), было бы логично предположить, что первые члены, занимая место дополнения, являются существительными. Однако нам бы не хотелось включать в лексикон корпуса «слова пустышки» без собственного значения.

emmi (только/даже/и обжаловал). В таких случаях, мы также предлагаем анализировать всю единицу целиком: {*міз де бақпайды*} = {*міз де бақ*}-V-NEG-AOR-P3.SG.

3. Токенизация «1=1+»

В нашем понимании, критерий «1=1+» следует применять к одному орфографическому слову, состоящему из нескольких синтаксических единиц, по одному из следующих признаков: 1) в результате морфологической сегментации как минимум один сегмент невозможно сопоставить существующей грамматической категории; 2) как минимум две синтаксические единицы, входящие в состав слова, вступают в синтаксическую связь с остальными членами предложения/клаузы.

По первому признаку мы применяем критерий к вариантам записи слитного произношения некоторых конструкций языка, при морфо-сегментации которых появляются «псевдо морфемные» сегменты. К таковым мы пока относим только форму глагол+ҒАсЫн, которая является записью слитного произношения конструкций типа {V-GER соң-POST}. В отличие от других форм записи слитного произношения (например, *сосын*={содан соң} и неғыл={не қыл}), данная форма не является лексикализированной и используется практически с любыми глаголами. В таблице 2 показан вариант разметки формы «барғасын» в сокращенном варианте формата CONLL-U, UD [9].

Таблица 2

Пример морфосинтаксического анализа формы «барғасын»

№	Токен	Лемма	ЧР	Грам-мемы	Главный	Синтаксическое отношение
1-2	барғасын	-	-	-	-	-
1	-	бар	V	GER|NOM	0	root
2	-	соң	POST	-	1	case

По второму признаку мы применяем критерий к деривативным формам, получающимся при субстантивации суффикса –ҒЫ. Мы также диагностируем другие дериваты, рассмотренные в работе [8], откуда мы заимствуем формулировку данного признака. В таблице 3 показан вариант разметки предложения (1) «*Көк машинадағылар ұйықтады*» (*те, кто в синей машине, спят*).

Таблица 3

Пример морфосинтаксического анализа формы предложения (1)

№	Токен	Лемма	ЧР	Граммемы	Главный	Синтаксическое отношение
1	Көк	көк	ADJ	-	2	amod
2-3	Машина дағылыр	-	-	-	-	-
2	-	машина	N	SG|LOC	3	nmod
3	-	-ҒЫ	N	PL|NOM	4	nsubj
4	ұйықтады	ұйықта	V	PAST|3PL	0	root

4. Заключение

В настоящей работе были рассмотрены некоторые сложности морфологической разметки языковых единиц, в которых количество орфографических слов не соответствует количеству единиц разбора. Было предложено решать такого рода проблемы путем использования двух дополнительных критериев токенизации: 1) «1+=1» – несколько орфографических слов составляют одну единицу разбора; 2) «1=1+» – одно орфографическое слово составляет несколько единиц разбора. На примере казахского языка были описаны признаки применимости данных критериев и их реализация в одной из существующих схем разметки.

Благодарности. Работа выполнена при финансовой поддержке АОО «Назарбаев Университет» (проект “Building a Kazakh Dependency Treebank”). Авторы также выражают признательность Фрэнсису Таерзу и Джонатану Вашингтону за консультации по вопросам токенизации.

Литература

1 Suleymanov D. et al. National corpus of the Tatar language “Tugan Tel”: Grammatical Annotation and Implementation //Procedia-Social and Behavioral Sciences. — 2013. — T. 95. — C. 68–74.

2 Building a Turkish treebank / K. Oflazer, B. Say, D. Z. Hakkani-Tür, G. Tür // Treebanks. — Springer, 2003. — C. 261–277.

3 About Linguistic Corpuses of the Bashkir Language / Z. Sirazitdinov, L Buskunbayeva, A. Ishmukhametova // 3rd International Conference on Turkic Languages Processing (TurkLang 2015). — Kazan, Tatarstan, 2015. — C. 269–275.

4 Towards a Free/Open-source Universal-dependency Treebank for Kazakh / F.M. Tyers, J. Washington // 3rd International Conference on Turkic Languages Processing (TurkLang 2015). — Kazan, Tatarstan, 2015. — С. 276–289.

5 Assembling the Kazakh Language Corpus / O. Makhambetov, A. Makazhanov, Z. Yessenbayev, B. Matkarimov, I. Sabyrgaliyev, A. Sharafudinov // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. — Association for Computational Linguistics, 2013. — С. 1022–1031.

6 Data-Driven Morphological Analysis and Disambiguation for Kazakh / O. Makhambetov, A. Makazhanov, Z. Yessenbayev, I. Sabyrgaliyev // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — Springer International Publishing, 2015. — С. 151–163.

7 Бичелдей, К. А. Поговорим по-тувински (Тывалап) чугаалажыылынар / К.А. Бичелдей. — Кызыл : Тувинское книжное издательство, 2012. — 128 с.

8 A Grammar-book Treebank of Turkish / Ç. Çöltekin // Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14). — Warsaw, Poland, 2015. — С. 35–49.

9 Towards a Universal Grammar for Natural Language Processing / J. Nivre // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — Springer International Publishing, 2015. — С. 3–16.

10 Syntactic Annotation of Kazakh: Following the Universal Dependencies Guidelines. A report / A. Makazhanov, A. Sultangazina, O. Makhambetov, Z. Yessenbayev // 3rd International Conference on Turkic Languages Processing (TurkLang 2015). — Kazan, Tatarstan, 2015. — С. 338–350.

11 Finite-state Morphological Transducers for Three Kypchak Languages / J. Washington, I. Salimzyanov, F. Tyers // Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14). — European Language Resources Association, 2014. — С. 3378–3385.

УДК 81'33

ИСПОЛЬЗОВАНИЕ КОРПУСА ТЕКСТОВ ПРИ ОБУЧЕНИИ ИНОСТРАННОМУ ЯЗЫКУ

А.Ф. Мухамадьярова
Казанский федеральный университет, Казань
liliana_muhamad@mail.ru
Б.Э. Хакимов
Казанский федеральный университет, Казань
НИИ "Прикладная семиотика" Академии наук Республики Татарстан, Казань
khakeem@yandex.ru

В статье рассматриваются общие методические аспекты использования корпусов текстов при обучении иностранным языкам. Определяются

преимущества корпусно-ориентированного подхода в преподавании иностранного языка студентам профильных филологических специальностей.

Ключевые слова: *корпус текстов, иностранный язык, обучение языку.*

Сегодня корпусная лингвистика является быстро развивающейся областью языкознания, о чем свидетельствуют национальные корпуса различных языков, созданные для научных исследований, а также для обучения языку. Хотя корпус считается инструментом, предназначенным в первую очередь для обеспечения научных исследований лексики и грамматики языка, то роль корпуса становится еще значительнее, если принять во внимание обучение иностранному языку.

В целом, в настоящее время одной из наиболее важных и устойчивых тенденций развития мирового образовательного процесса является применение современных информационных технологий. Внедрение информационных технологий в учебный процесс является неотъемлемой частью успешной работы преподавателя. Поэтому обращение к корпусам текстов и корпусным технологиям как к средству поддержки обучения иностранному языку является одним из актуальных направлений современной методики.

В этом контексте создание специализированных методически организованных языковых корпусов становится все более актуальной задачей. Это продиктовано необходимостью разрешения сложившихся противоречий между потенциальной высокой результативностью корпусных исследований и недостаточной их реализацией в практике обучения иностранному языку в вузе. В своем исследовании мы стремимся выявить конкретные преимущества корпуса над традиционным учебником иностранного языка на примере обучения студентов профильных филологических специальностей.

Во-первых, корпус является источником естественных языковых примеров. В ходе многих исследований было указано, что существуют значительные расхождения между тем, что предписывается учебниками, и тем, как язык действительно используется носителями [5: 57]. Язык в учебниках во многом отличается от языка литературы, газет и устной речи. По крайней мере, данные корпусов надежнее искусственных примеров, придуманных преподавателем или автором учебника [11]. Для преподавателей, не являющихся носителями языка, возможность работы с корпусом означает непосредственную доступность достоверной информации о языке [8].

Отличительной чертой корпусных методик является обращение исследователей к реальному употреблению языковых единиц.

В.А. Плунгян утверждает, что акцент перемещен с языка на тексты, на реальность, на живое пространство языка, и отмечает, что для овладения языком человеку нужны словарь, грамматика и корпус текстов данного языка, потому что отдельно взятые словарь и грамматика бесполезны вне живого пространства, в котором функционирует язык [6].

Другой важный аспект корпусной web-дидактики - обучение через исследование. Студент-исследователь может включиться в процесс освоения грамматики и семантики родного или иностранного языка на любом этапе благодаря непосредственному доступу к языковому материалу корпуса, удобству поисковой системы информации в корпусе и гибкости в формировании критериев запроса, благодаря чему повышается его самостоятельность [2: 351]. Отбирая, систематизируя и анализируя языковые данные, студент проводит свое исследование. Преподаватель координирует работу студента и определяет траектории его самостоятельного исследования. В статье С.С. Дикаревой указывается на постепенный отказ от традиционного подхода к обучению, требующего объяснений, к *когнитивно-коммуникативным* практикам, формам и методикам, т.е. к «*обучению через исследование*», которое стало новой дидактической парадигмой [2: 351].

Итальянская исследовательница с области корпусной дидактики Сильвия Бернардини пишет, что в преподавании произошел сдвиг акцента с дедуктивного подходав преподавании к индуктивному, то есть опирающемуся непосредственно на данные и связывающему значение с формой. Важность новой дидактической парадигмы «обучение через исследование» настолько велика, что в дидактические рекомендации Совета Европы включается разработка задач и методов преподавания языка путем исследования и анализа. Корпусы дают студентам практический материал, с которым они столкнутся при использовании языка в реальных ситуациях межкультурной коммуникации. Корпусное преподавание – процесс открытый, творческий, где нет заранее данных оценок. С помощью корпуса студент получает возможность выбирать из заранее не заданного, неопределенного множества ответов на вопрос [2: 351].

Не имея наглядных примеров использования того или иного слова или выражения, студент делает основные ошибки при изучении языка. И, например, параллельные корпусы дают студенту возможность самостоятельно найти примеры применения изучаемых правил и предлагают типовые и нестандартные решения при переводе с одного языкана другой. Следует отметить, что полученные во время самостоятельного исследования результаты в большей степени откладываются в памяти. Таким образом, студент самостоятельно

подбирает в корпусе примеры языковых данных для дальнейшего, самостоятельного исследования языка.

Последние исследования Университета Ланкастера, посвященные программному обеспечению для обучения студентов младших курсов грамматике и основам грамматического анализа, показали, что программы, создаваемые достаточно легко на основе корпуса текстов, аннотированного или по частям речи или по грамматическим/ синтаксическим ролям, чрезвычайно эффективны и обеспечивают нужную степень как интерактивности, так и автономности [5: 58]. Получая задание грамматического разбора текста со скрытой аннотацией, студенты самостоятельно разбирают предложения, имея возможность запросить у программы помощь в виде списка обозначений информации о частотности употребления той или иной лексической единицы или частотностисовместного употребления примеров (коллокации).

О.В. Нагель приводит пример, каким образом можно использовать корпус текстов и поиск конкорданс в преподавании иностранного языка. Задание для студентов может выглядеть следующим образом:

Вопрос: Какова разница между словами remember и remind? Правильно ли следующее?

Do you remember me? I used to sit at the back of your class (correct).

Can you remember me? I used to sit at the back of your class. (Is this wrong?)

The flowers reminded him his garden. (Is this wrong?)

Можете ли вы дать мне набор предложений (около 20), где бы я мог заполнить пробелы, используя «remember» или «remind» [5: 58]

В результате студенты в режиме конкорданса получают наборпредложений с нужным словом и в процессе контекстного анализа выявляют семантическую разницу [5: 58].

В ходе экспериментов, проведенных МакЭнери, Бэйкером, Уилсоном с целью определения эффективности этой методики, была установлена степень усвоения знания частей речи среди студентов, обучавшихся по новой – корпусной методике, и в контрольной группе студентов, обучавшихся по традиционной – лекционной методике. В целом студенты, обучавшиеся с помощью корпусного программного обеспечения, последовательно демонстрировали более высокие показатели, нежели студенты контрольной группы. Рассматривая вопросы, с которыми ежедневно сталкиваются студенты на различных стадиях овладения иностранным языком, можно отметить, что методы корпусной лингвистики, например, составление конкорданса, оптимально приспособлены для того, чтобы обеспечить интересный и самостоятельный поиск ответов, в процессе которого студент имеет возможность как

получить искомые сведения по частным вопросам изучаемого языка, так и получить представление о реальном естественном состоянии языка, его историческом, географическом и социальном варьировании, регистрах речи,жанровом разнообразии [5: 58].

Преподавание языка включает в себя изучение не только грамматических правил и лексических единиц, но и изучение культуры, традиций и обычаев народа изучаемого языка. Корпус выступает в данном случае надежным источником, позволяющим получить достоверную информацию. Например, проанализировав 557 примеров с компонентом-колоронимом blau (синий, голубой) из Мангеймского корпуса немецкого языка (IDS-Corpora) [23], мы пришли к выводу, что blau может выступать в качестве метафоры для обозначения состояния человека: сентиментальности, уныния – Blauäugleinposie, Blauklang; нетрезвого состояния – Blaudenker, Blauzustand, stockblau.

Дальнейшая работа с идиомами заключается в составлении упражнений, самостоятельной и контрольной работ с использованием конкорданса. Далее можно провести корпусное исследование по дистрибуции выражения в разных жанрах корпуса. Также интересны для рассмотрения и обсуждения случаи деформации идиом, оживления их компонентов, намеренное обыгрывание или мета-метафора. Такие случаи часто используются в политическом дискурсе, политики нередко прибегают к этому стилистическому средству в целях убеждения. Студенты получают задание найти случаи деформации метафоры, фразеологизма в корпусных примерах в качестве домашнего задания. Например, студентам может быть предложено с помощью самостоятельного исследования определить, какое значение приобретает колороним blau в указанных выше примерах.

Как известно, язык – это динамическая система, и это должно находить отражение в словарях и грамматиках. В обучении иностранным языкам важную роль играют учебники, учебно-методические пособия, использующиеся преподавателями в качестве основного источника материала. В процессе преподавания иностранного языка многие специалисты сталкиваются с проблемой нехватки актуальных текстовых материалов. Благодаря своей репрезентативности корпус текстов становится незаменимым источником аутентичного материала для лингвистов и преподавателей иностранных языков. Анализ корпусов текстов, методы и наработки корпусной лингвистики являются перспективным направлением в области преподавания иностранных языков.

Подводя итог, следует признать корпус важнейшим инструментом в изучении и обучении иностранных языков. Корпусное преподавание не

только обеспечивает наблюдение за функционированием языка в реальных контекстах. Оно делает обучение открытым творческим процессом, в центре которого стоит студент/ученик, вовлеченный в самостоятельную исследовательскую деятельность. Корпусные методы зарекомендовали себя в мировой практике лингвистических исследований и преподавания иностранных языков как высокоэффективное дополнение к уже имеющимся, традиционным образовательным технологиям.

Следует также отметить большой потенциал корпусов текста в формировании лингвокультурологической компетенции при обучении иностранному языку у студентов. Примеры метафор, фразеологических и паремиологических единиц, реалий, ложных друзей переводчика, взятые из корпуса, являются живым материалом и позволяют студентам наблюдать их реализацию в речи.

Литература

1. Горина О.Г. Использование технологий корпусной лингвистики для развития лексических навыков студентов-регионоведов в профессионально-ориентированном общении на английском языке: автореф. дис. … канд. пед. наук. Москва, 2014. – 24 с.

2. Дикарева С.С., Чернявская О.Г. Корпусная web-дидактика: принципы и перспективы // Ученые записки Таврического национального университета им. В.И. Вернадского Серия «Филология. Социальные коммуникации» Том 26 (65). № 1 – С. 350-355 URL: http://sn-philolsocom.crimea.edu/arhiv/2013/uch_26_1fil/063_dika.pdf

3. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн., - СПб.: СПбГУ. РИО. Филологический факультет, 2013. – 143с.

4. Леденева С.Н.Использование корпуса текстов при обучении английскому языку.//Магия ИННО: новые технологии в языковой подготовке специалистов-международников. Материалы научно-практической конференции к 70-летию факультета международных отношений(Москва, 4-5 октября 2013г.). Т.1/ отв.ред. Д.А.Крячков.- Моск.гос.ин-т междунар. отношений (Ун-т)МИД России- М.: МГИМО - Университет, 2013. - С.287-294 URL: http://mgimo.ru/library/publications/1007792/ (Дата обращения: 15.01.2016)

5. Нагель О.В. Корпусная лингвистика и ее использование в компьютеризированном языковом обучении Текст // Язык и культура. – 2008. – Т.1. №4. – С. 53-59

6. Плунгян В.А. Почему современная лингвистика должна быть лингвистикой корпусов // Публичная лекция, 2009. [Электронный ресурс]. - URL: http://www.polit.ru/lectures/2009/10/23/corpus.html (Дата обращения: 10.01.2016)

7. Подлесская В. И. Современные компьютерные методы в изучении и преподавании лингвистических дисциплин: корпусная лингвистика. URL: http://www.rsuh.ru/print.html (Дата обращения: 10.01.2016)

8. Садовникова О.Э. Прямое и косвенное использование корпусов в зарубежной лингводидактике // Magister Dixit – научно-педагогический журнал Восточной Сибири. 2013 - № 2 (06) URL: http://md.islu.ru/sites/md.islu.ru/files/rar/statya__pryamoe_i_kosvennoe_ispol_zovanie_korpusov_tekst_ov_v_zarubezhnoy_lingvodi_daktike.pdf

9. Сидорова Е.А. Подход к построению предметных словарей по корпусу текстов // Труды международной конференции «Корпусная лингвистика–2008». СПб., 2008. С. 365–372.

10. Соснина Е.П. Корпусная лингвистика и корпусный подход в обучении иностранному языку // Прикладная лингвистика. Статьи. – URL: http://ling.ulstu.ru/linguistics/resourses/literature/articles/corpus_linguistics_language_teaching(Дата обращения: 10.01.2016)

11. Синклер Д. Предисловие к книге «Как использовать корпуса в преподавании иностранного языка»/ Д. Синклер [Электронный ресурс]. – URL: http://www/ruscorpora.ru/corpora-infro.html (Дата обращения: 10.01.2016)

12. Чернякова Т.А. Использование лингвистического корпуса в обучении иностранному языку // Язык и культура, №4, 2014 URL: http://cyberleninka.ru/article/n/ispolzovanie-lingvisticheskogo-korpusa-v-obuchenii-inostrannomu-yazyku (Дата обращения: 10.01.2016)

13. Barbara Seidlhofer. 2002. Pedagogy and Local Learner Corpora: Working with Learner-driven Data. In S. Granger, J. Hung & S. Petch-Tyson.(eds) Computer Learner Corpora, SecondLanguage Acquisition and Foreign LanguageTeaching. John Benjamins: Amsterdam, pp. 213-234.

14. Biber D. (1993), Representativeness in Corpus Design, Literary and Linguistic Computing, Vol. 8, No. 4, pp. 243–257

15. Kohn, Kurt (2009) Computer assisted foreign language learning. In: Karlfried Knapp; Barbara Seidlhofer (Hrsg.) Foreign Language Communication and Learning. Handbooks of Applied Linguistics, volume 6.Berlin: Mouton-de Gruyter, 573-603.

16. Köhler Reinhard, 2005, Korpuslinguistik. Zu wissenschaftstheoretischen Grundlagenund methodologischen Perspektiven, in: LDV-Forum 20/2, S. 1-16.

17. McEnery T., Hardy A. Corpus linguistics, Cambridge University Press, Cambridge, 2011.

18. Nasselhauf, N. (2004): “Learner corpora and their potential for language teaching”, inSinclair, J. (Ed.), How to Use Corpora in Language Teaching. Amsterdam:John Benjamins Publishing, pp. 125-152.

19. Thomas J. Using Corpora in Language Teaching and Learning // Teaching English with Technology:A Journal for Teachers of English. January 2006. Vol. 6.Issue 1. URL: www.iatefl.org.pl/call/j_soft23.htm

20. Simpson, R. &Mendis D. (2003). A corpus-based study of idioms in academic speech. TESOLQuarterly, 27 (3), 419-441.

21. Varley, S. (2008). I’ll just look that up in the concordance: Integrating corpus consultation into the language learning environment. Computer Assisted Language Learning, 22 (2), 133-152.

Интернет – источники

22. URL: http://www.ruscorpora.ru/

23. URL: http://www.ids-mannheim.de/cosmas2/projekt/

24. URL: http://wortschatz.uni-leipzig.de

УДК81'322.2

ЛИНГВИСТИЧЕСКОЕ АННОТИРОВАНИЕ ЗАЛОГОВЫХ ФОРМ ГЛАГОЛА ЯЗЫКА САХА

А.Н. Ноговицына
Северо-Восточный федеральный университет
им. М.К. Аммосова, г. Якутск
erkin2007@mail.ru

В данной статье рассматривается проблема лингвистического аннотирования залоговых форм глагола языка саха. Для отображения залоговых форм якутского глагола автором предлагаются следующие тэги: основной - ACT (Active), побудительный - CAUS (Causative), возвратный - REFL (Reflexive), страдательный - PASS (Passive), совместно-взаимный - RECIPR (Reciprocal).

Ключевые слова: корпусная лингвистика, якутский язык, залоговые формы глагола, морфологическая разметка текстов.

Создание электронных корпусов миноритарных тюркских языков является актуальной задачей современной тюркологии. В последнее время активно обсуждаются проблемы разработки унифицированной системы морфологической разметки текстов на тюркских языках для использования в национальных электронных корпусах и других системах автоматической обработки текста. Создание единого стандарта представления лингвистической информации позволит тюркским языкам войти в единое информационное пространство.

Для разметки залоговых форм глагола на языке саха нами заимствованы тэги из Лейпцигской системы глоссирования, которые используются также при аннотировании электронных корпусов тюркских языков.

В якутском языке (как и в других тюркских языках) существуют следующие залоги глагола: 1) основной, 2) побудительный, 3) возвратный, 4) страдательный и 5) совместно-взаимный [2: 253].

Как известно, формой основного залога в тюркских языках принято называть первичную основу глагола, служащую исходным структурным элементом для образования всех производных форм глагола, в том числе и залоговых [1:13]. Для разметки основного залога языка саха нами предлагается использование пометы: ACT (Active).

Глаголы страдательного залога образуются с помощью аффиксов –н и –ылын [1:104]. От основ с конечным гласным образуются глаголы,

подобно возвратным, посредством аффикса –н, например, эрбэн – ‘пилиться, быть распиленным’от эрбээ – ‘пилить’; от основ с конечным согласным и *й* – посредством аффикса –ылын, например, баайылын – ‘вязаться, быть связываемым, связанным’ от баай – ‘вязать’, образующего 4 алломорфа: -ылын/-илин/-улун/-үлүн (см. табл. 1).

Таблица 1

Сокращения	**Расшифровка сокращений**	**Название категории на русском языке**	**Название категории на якутском языке**	**Алломорфы**	**Морфемы**
PASS	Passive	Страдательный залог (пассив)	Атынтан туһаайыы	-н -ылын/-илин/-улун/-үлүн	-н -ылын

Для разметки возвратного залога нами использована помета REFL (Reflexive). В якутском языке возвратные глаголы образуются посредством аффикса –н (-ын) [1: 76] (см. табл 2).

После основ с конечным гласным или дифтонгом следует аффикс –н, например, анан – ‘назначать себя, себе; быть назначенным’ от анаа – ‘назначать’; после основ с конечным согласным следует аффикс –ын, например, ылын – ‘взяться, брать себе’ от ыл – ‘взять, брать’.

Таблица 2

Сокращения	**Расшифровка сокращений**	**Название категории на русском языке**	**Название категории на якутском языке**	**Алломорфы**	**Морфемы**
REFL	Reflexive	Возвратный залог (рефлексив)	Бэйэни туһаайыы	-н (-ын/-ин/-ун/-үн)	-н (-ын)

Рассматривая аффиксы страдательного и возвратного залогов, мы отмечаем наличие омонимичности аффиксов. Аффикс –н возвратного залога по внешней форме совпадает с аффиксом –н страдательного залога. Например:

Холбоо – ‘соединять’, холбон – ‘соединяться, быть соединенным’. -Страдательный залог.

Холбоо - ‘соединять’, холбон – ‘соединяться, присоединять к себе’. - Возвратный залог.

Как отмечал Л.Н. Харитонов, “во всех этих случаях наличие у данного производного глагола возвратного или страдательного значения устанавливается лишь по контексту, а также по смыслу глагола” [1: 105]. Отсюда омоформы глагола могут вызвать проблемы при автоматической обработке текста, поэтому снятие омонимии - это одна из серьезных проблем лингвистического аннотирования морфологических категорий языка.

В якутском языке побудительные глаголы образуются посредством аффиксов –т, -тар, -ар, -ыар [1: 53]. Как видно, каузативные глаголы отличаются большим разнообразием аффиксов. Здесь уместно отметить, что некоторые каузативные аффиксы якутского языка близки к татарским: “каузативные глаголы … образуются при помощи разнообразных аффиксов: -т, -тыр/-тер, -дыр/-дер, -кар/-кер, -ыр/-ер, -кыз/-кез, -гыз/-гез и др. в зависимости от фонетических особенностей глагольной основы” [3: 67].

В качестве разметки побудительного залога нами использована помета CAUS (Causative) (см. табл.3).

Таблица 3

Сокращения	**Расшифровка сокращений**	**Название категории на русском языке**	**Название категории на якутском языке**	**Алломорфы**	**Морфемы**
CAUS	Causative	Побудительный залог (каузатив)	Дьаһайартуһайыы	-т	-т
				-тар/-тэр/-тор/-төр -дар/-дэр/-дор/-дөр -нар/-нэр/-нөр/-нор -лар/-лэр/-лөр/-лор	-TAp
				-ар/-эр/-ор/-өр	-Ap
				-ыар/-иэр/-уор/-үөр	-ЫAp

Аффикс –т употребляется при следующих условиях:

- при всех глагольных основах с конечным долгим гласным и дифтонгом, например, санат – ‘заставить думать’ от санаа – ‘думать’;

- при основах непереходного значения с конечным *й*, например, байыт – ‘обогащать’ от бай – ‘богатеть’;

- при основах непереходного значения с конечным *р*, например, куурт – ‘сушить’ от куур – ‘сохнуть’;

Аффикс –тар присоединяется к глагольным основам с конечным согласным и обычно непереходного значения, например, тиктэр – ‘заставить шить’ от тик – ‘шить’.

Аффикс –ар встречается при немногих односложных основах непереходного значения с конечными согласными, например, ситэр – ‘доводить до конца, завершать’ от сит – ‘завершаться’.

Аффикс –ыар употребляется при единичных односложных основах непереходного значения с конечным согласным, например, туруор – ‘поставить, заставить стоять, встать’ от тур – ‘стоять, вставать’.

Наиболее широкую сферу применения в первичных глаголах имеют аффиксы –т и –тар. Аффиксы –ар и –ыар встречаются намного реже [1:53-57].

В данном случае мы сталкиваемся с проблемой омонимии аффиксов, выражающих разные грамматические значения: 1) аффиск –т побудительного залога совпадает по форме с аффиксом -т недавнопрошедшего времени; 2) аффикс –ар совпадает по форме с причастием настоящего времени на –ар (положительная форма); 3) аффикс -тар совпадает по форме с аффиксами: –тар условного наклонения (-дар/-лар/-нар); –лар множественного числа (-тар/-тэр/-тор/-төр, -дар/-дэр/-дор/-дөр, -нар/-нэр/-нөр/-нор, -лар/-лэр/-лөр/-лор).

При аннотировании совместно-взаимного залога нами использована разметка RECP (Reciprocal). Совместно-взаимные глаголы в якутском языке образуются посредством аффикса –с (-ыс) [1:19].

- после основ с конечным гласным следует –с, например, ыас – ‘доить вместе’ от ыа – ‘доить’;

- после основ с конечным согласным следует –ыс (-ис/-ус/-үс), например, барыс – ‘уходить, идти вместе’ от бар – ‘уходить, идти’.

Таблица 4

Сокращения	**Расшифровка сокращений**	**Название категории на русском языке**	**Название категории на якутском языке**	**Алломорфы**	**Морфемы**
RECIPR	Reciprocal	Совместно-взаимный залог (реципрок)	Холбуу туһаайыы	-с (-ыс/-ис/-ус/-үс)	-с (-ыс)

Термины **(**Causative, Reflexive, Passive, Reciprocal), используемые в современной лингвистике для обозначения залоговых форм глагола, созвучны с терминами (Causativa, Reflexiva, Passiva, Reciproca), которыми оперирует О.Н. Бётлингк в своем знаменитом труде “О языке якутов”. Данный факт был также отмечен Г.Г. Торотоевым: «В нашей работе, в частности, в идентификации грамматических категорий с соответствующими тэгами очень помог научный труд О.Н. Бётлингка «О языке якутов», который был издан на немецком языке в далеком 1851 г. Все падежи якутского языка именуются согласно принятому тогда в языковедении терминами, имеющими латинские корни. И поэтому нет существенных расхождений между терминами, используемыми О.Н. Бётлингком и современными унифицированными тэгами» [5: 367].

“Создание систем автоматического анализа морфологии тюркских языков – насущная проблема, актуальная вовсе не формально. Задача эта прежде всего прикладная, но тесно связанная с теоретической областью. Во-первых, именно на теоретических описаниях строятся программные средства, а, во-вторых, исследования, которые может провести на уже реализованных автоматических системах, могут серьезно скорректировать наши теоретические представления о языке. То есть иногда язык оказывается не таким, как его описывают лингвисты” [4: 135]. Таким образом, эффективное использование возможностей искусственного интеллекта позволит получить новые знания об устройстве языка, и в дальнейшем усовершенствовать методы лингвистических исследований.

Литература

1. Харитонов Л.Н. Залоговые формы глагола в якутском языке. – Ленинград: Ленинградское отделение Издательства Академии наук СССР. - 1963. – 121 с.

2. Харитонов Л.Н. Грамматика современного якутского литературного языка: Фонетика и морфология. Т.1 / Л. Н. Харитонов, Н. Д. Дьячковский, С. А. Иванов и др.; Отв. ред. Е. И. Убрятова. – М.: Наука, 1982. – 496 с.

3. Галиева А.М.Отражение каузативности глаголов в корпусе татарского языка// Труды Казанской школы по компьютерной и когнитивной лингвистике: материалы междунар. науч. конф. (Казань, 6-9 февраля 2014 г.) - Казань: Изд-во “Фэн”Академии наук РТ, 2014. - Вып. 16. - С. 66-71.

4. Орехов Б.В. Проблемы морфологической разметки башкирских текстов // Труды Казанской школы по компьютерной и когнитивной лингвистике: материалы междунар. науч. конф. (Казань, 6-9 февраля 2014 г.) - Казань: Изд-во “Фэн”Академии наук РТ, 2014. - Вып. 16. - С. 135-139.

5. Торотоев Г.Г. Linguistic annotation of grammatical categories of Sakha language (on example of noun) // Proceeding of tne International Conference on Turkic languages processing (Kazan,17-19 September 2015): Academy of Sciences of the Republic Tatarstan Press, 2015, pp.363-373.

УДК 811.512'37

СЕМАНТИЧЕСКИЕ ПАРАЛЛЕЛИ В АЛТАЙСКИХ ЯЗЫКАХ

Д.Б. Рамазанова
Институт языка, литературы и искусства
им. Г.Ибрагимова АН РТ

В статье рассматриваются параллельные семантические модели в татарском (и в тюркских), тунгусо-маньчжурских, монгольских языках. Основную лексическую базу анализа составляют тематические группы: соматизмы, названия одежды, и др. Определена роль принципов номинации, метафоризации и структурного оформления лексем и развития их семантики.

Ключевые слова: развитие семантики, семантические параллели, семемный состав слова, семантические модели, названия пальцев.

Лексикологическая проблематика, разрабатываемая в кампаративном направлении, в конечном счете приводит к задаче выявления предполагаемого общего исходного словарного фонда генетически родственных языков с характеристикой его и в семантическом отношении [14, 3]. Решение этих задач прежде всего базируется на обзоре основных лексико-тематических групп: термины родства, соматизмы, названия животных и растений, наименования жилища, утвари, обозначения действий и др.

В лексикологических исследованиях часто отмечается семантический сдвиг в словах. При диахроническом анализе приходится обращать внимание на изменение семантической функции слова, метафорический переход значения. При изучении общеалтайской лексики в сравнительном аспекте были обнаружены случаи параллельности развития значения в родственных языках. Так, исследования показали, что слово *баш* еще будучи в корневой форме приобрело ряд значений: голова, ум, разум, глава, ведущий, крыша, конец, верхний конец и вершина, верх, верхушка, верховье, конец, начало, головка, кочан, заголовок, вожак, главный и др. Широкий спектр значений развился у корневого слова *күз*: глаз, зрение, взгляд, в центре внимания, определенное отношение, мнение, пустая дыра, ушко, глазок (в двери), звено(у окна), глазок (у картофеля), наполненная жидкостью углубление: середина лужи, полынья, поры (в выпеченном хлебе) и др. Ряд аналогичных направлений семантического развития известны и в других языках.

Известно, что в ходе развития языка развивалась и их семантика, происходило структурное развитие, образовывались дериваты. Выявилось, что некоторые первичные или вторичные образования оказывались синонимами корневых форм. Например, тат. *баш* – глава, главарь, *башлык* – глава, главарь, тат. *күз* – точка зрения, мнение, *күзлек* – точка зрения, воззрение; *күз* – глазок и *күзчә* – глазок (у картофеля), *күз* – полынья и *күзләwек* (күз+лә+wек) – полынья и т.д. (7, 13-21, 33-49). Названия частей тела относятся древнейшему пласту лексики, подавляющее большинство из них восходят к праалтайскому периоду [10, 203; 7, 4; Рамазанова, 2016: 93-101].

Семантические параллели между алтайскими языками обнаруживаются в способах номинации, при этом основы-лексемы составляют разные слова. Любопытный пример составляет слово *бармак*. Считается, что оно возникло на основе глагола *бар-* (поймать, ловить, хватать рукой), этот корень до сих пор бытует в монгольском языке [3, 172]. Такой же семантический переход присущ и некоторым другим языкам. Так, в маньчжурских языках слово *дала* (рука) также образовалось на основе глагола *дай-* (брать, хватать). Аналогичным способом возникло слово *карак* (бармак) от корня *ка-* (брать). Таким образом, в алтайских языках понятия палец, рука возникли на основе различных глаголов с общим значением, на основе общей семы брать, хватать, ловить и др., на основе единой номинации.

В то же время в некоторых языках алтайской группы слово палец возникло в результате развития другой семы. Например, в прототюркском языке бытовал глагол *дар-* (ветвиться), отсюда *дармак* (палец), в праалтайском языке *чар-* (также: ветвиться), *чарбу* – запястье, часть руки ниже плеча; в монгольском *сарба* – ветвиться, в шорском языке *сарбаш* (палец) и др.

Из вышеизложенного, можно заключить, что лексемы со значением палец взникли в результате одинакового развития различных лексем, близкие по значению и по принципам номинации, что подтверждает общность алтайских языков

В алтайских языках обнаруживается общность и в названиях пальцев. Так, имеется общность между всеми тремя группами алтайских языков в содержании названий безымянного пальца. В татарском *атсыз бармак* (безымянный палец, эвенк. ātэγāk < ātэ – (его) имя + γак (тат. юк, рус. без, нет), эвенк. гэрбйе āчин, эвен. ач гэрбэлэ, манжч. гэбу аку симхун < гэрби/гэрбэ, гэбу – имя + āчин, ач, аку (нет).

Ср.: якут. ātасуох, алт. атыйак < аты + юк, т.е. нет имени, безымянный.

Алтайские языки объединяются и общностью принципов номинации мизинца. Как пишет В.Д. Колесникова [4, 325], в тунгусо-маньчжурских языках этот палец называется словами, объединяющимися в общую сему маленький, младший, последний, одинокий. Сравни в татарском: маленький, младенец, малюсенький, диал. кечр'ук – малюсенький, кәтеки – маленький [7, 113-115; 12].

Видимо, можно предположить и о том, что слово *аяк* (нога) также перенес семантический сдвиг еще будучи в корневой форме. Его семантическое развитие имеет общие черты во всех трех (в тунгусо-маньчжурских, монгольских, тюркских) группах языков: ноги животных, нога человека, ножки мебели и других артефактов [4, 329].

Любопытная общность отразилась в названии подсолнуха. Как известно, татарские названия данного растения представляют собой словосочетания *көн, кояш* (солнце) + *багыш, багар* < *багу* (глядеть, смотреть) либо *әйләну* (поворачиваться). Ср.: письменно-монгольский nara(n) čečeg (солнце + цветок), бурятск. наран сэсэг (так же), калмык, narŋ + өwsŋ < солнце+трава.

Примеры на семантическую параллель можно продолжить. В татарском языке известны названия частей руки, употребляющиеся как название меры длины: *берилле* – толщина, равная толщине одного пальца, *икилле* – равная толщине двух пальцев. Второй компонент у приведенных примеров (ил+ле) восходит к др.-тюркск. *әл*, распространенному в значениях *кул* (рука), *кул чугы* (кисть руки), *бармак* (палец) и др. В маньчжурском языке бытует слово *урхун* – мера длины, равная толщине одного пальца.

Возможно, соматизмы, выражающие разные части рук еще в праалтайский период употреблялись и как названия меры длины. Например, в тунгусо-маньчжурских языках известны слова сӱо, орокском сиро – пядь, вершок; четверть, в эвенкийском сӱм – вершок и др. Ср.: монг. сццм, якут. сӱӧм, тат. сөям, башк. һөйәм – пядь, четверть и др. Как и в татарском языке, в тунгусо-маньчжурском и монгольском языках слова *кар, кары* и их варианты употреблялись как названия меры длины [4, 318, 323].

Семантический сдвиг у соматизмов в метафорическом плане довольно распространенное явление. Примеры общих явлений в номинации, между тюркскими и тунгусо-маньчжурскими языками можно продолжить. Как указали выше, ряд значений татарского слова *баш* обнаруживается и в тунгусо-маньчжурских языках. Слово тат. чырай наблюдается во всех трех (тунгусо-маньчжурских, монгольских, тюркских) группах алтайской семьи языков, причем во всех из них обнаруживаются (кроме основной функции – лицо) переносные его значения: физиономия, черты лица, внешний вид, облик.

Как указано было выше, татарское слово күз еще в корневой форме приобрело большое число переносных, метафоричных значений [7, 31-49]. Среди них: глазки плетеных артефактов: сетей, платков, различных вязаных вещей. Это же значение (глазки сети) известно в южной группе тунгусо-маньчжурских языков. Из этого можно высказать предположение, о том, что такое развитие корневого слова *күз* произошло еще в праалтайский период. Активно употребляемые в татарском языке метафорические выражения «күз алмасы» (глазное яблоко), «күз чокыры» (глазница) известны и в тунгусо-маньчжурских языках.

Как известно в татарском языке соматизм *борын* также имеет значительное число переносных значений [13, 60]. Среди них назовем нос (лодки, самолета, чайника, самовара и других подобных емкостей), мыс и ., которые характерны кроме тюркских и для монгольских, тунгусо-маньчжурских языков [4, 260-283].

Слова *авыз* (тат.), *ам* (тунг.-маньч.), *ам/ама* (монг.) также имеют семантическое развитие в общем направлении: «Помимо значения «рот», «уста», указанные слова почти во всех тунгусо-маньчжурских и монгольских языках означает любое «отверстие», «вход», «проход» в нору, берлогу…» [4, 289].

В последнее время активизировались исследования лексического состава языков в различных (структурных, лексических, синхронных, диахронных) и др. аспектах, на передний план вышел семантический аспект, определение направления семантического развития и др.

В говорах татарского языка для выражения понятия «мой муж», «муж» употребляются слова со значением «хозяин»: байым – мой хозяин, дословно богатый, *хуҗам* – мой хозяин, *татарым* – мой хозяин (досл. татарин), *ир кешем* – мой муж (досл. мужчина) и т.п. [8, 46-47]. Аналогичное явление присуще и тунгусо-маньчжурским языкам, где в указанной функции употребляется слово *иде* (тат. ия – хозяин) в различных фонетических вариантах и значениях (хозяин, царь, правитель, господин, повелитель, государь и др.).

К.Г. Менгес анализирует распространенность слова idi (господин) в различных тюркских языках алтайской семьи, также и в других тюркских языках [5,101-110], аналогичного мнения и К.М.Мусаев [6, 33].

Иследуя параллелей семантического характера необходимо указать перенос значения название части тела → названия одежды и украшений, что, как показали исследования последних лет (7, 2013), особенно активно выступает в татарском языке. Примеры такого явления обнаруживаются и в тунгусо-маньчжурских языках. Например, понятия «воротник», «ожерелье», «шарф», «ошейник» передаются в них соматизмами, обозначающими шею, горло, глотку, горловину [4, 292-293].

В маньчжурском языке понятия «запястье» и «браслет» передаются одним и тем же словом: *сэмкэн*. Сравни также: тат. беләзек (браслет), письменно-монгольские biločüg ~ bilečeg ~ bilisüg ~ bilüčeg ~ bilečeg ~ bilüčug и монг. билжиг (кольцо, перстень).

Общность между родственными языками наблюдается и при семантичес-кой адаптации заимствований. Так, у заимствования рус. вдруг в татарских говорах зафиксированы следующие смысловые оттенки: *дурык* (лаишск.) – быстро, моментально, *дырук* (лаишск., хвал., дрожж.) – неожиданно, *төрөк* (тобол.), *терек* (тевриз.) вместе, сразу, одновременно, *дөрөн* (нагорн., дубъязск., пермск., нижнекамско-кряшенский) – одновременно, вдруг и др. [12]. Ср.: эвенк. дурук – все вместе, всех; эвен. gэγэрэк – неожиданно, вместе, тут же. Считается, что слово вошло из якутского языка, где дурук ~ дӯрук – все вместе, скопом.

Изучение семантики, номинационных принципов, особенностей переноса значений, выявление как направления, так и состава семем, связь всех их с этнокультурным развитием мышления народа имеют большое значение для сравнительно-исторических, сравнительно типологических разработок в общем языкознании.

Исследование семантических параллелей между родственными языками имеет важное значение не только для разработок по языкознании в историческом плане, но также и по прикладному языкознанию, для усовершенствования теории машинного перевода, семиотики и др.

Литература

1. Дыбо А.В. Семантическая реконструкция в алтайской этимологии: соматические термины (плечевой пояс). М.: Языки русской культуры, 1996. – 389 с.

2. Дыбо А.В. Антропоморфная и зооморфная метафора в тюркских языках //Сравнительно-историческая грамматика тюркских языков. – М.: Наука, 2006. С. 648-659.

3. Егоров В.Г. Этимологический словарь чувашского языка. Чебоксары: Чувашкниго изд-во,1964. 368 с.

4. Колесникова В.Д. К характеристике названий частей тела человека в тунгусо-маньчжурских языках // Очерки сравнительной лексикологии алтайских языков. Ленинград: Изд-во «Наука», Ленингр. отделение, 1972. С.257-336.

5. Менгес К.Г. Тюркское idi ‘господин’, некоторые его рефлексы в тюркских языках и параллели в других языковых семьях // Turcologica: к 70-летию А.Н.Кононова. Л., 1976. С.101-110.

6. Мусаев К.М. Лексикология тюркских языков. М.: Наука, 1984. – 228 с.

7. Рамазанова Д.Б. Татар телендә кешегә бәйләнешле лексика. Казан: Татар. кит. нәшр.. 2013. 364 б.

8. Рамазанова Д.Б.Атамаларда гаилә һәм туганлык мөнәсәбәтләре. Казан: Татар. кит. нәшр.. 2014. 287 б.

9. Рамазанова Д.Б. Татар тарихи лексикологиясенә материаллар (ностратик чордан борынгы болгар чорына кадәр). Казан: Школа, 2016.

10. Сравнительно-историческая грамматика тюркских языков. М.: Наука, 1997/2000 г.

11. Сравнительный словарь тунгусо-маньчжурских языков: материалы к этимологическому словарю. – Том I; А-Ң. Л.: Наука, Ленингр. отд-е, 1975. – Т.I. – 672 с.; Т.II. 992 с.

12. ТТЗДС: Татар теленең зур диалектологик сүзлеге. – Казань: Татар. кит. нәшр., 2009. 839 б.

13. ТРС: Татарско-русский словарь. Казань, 1988.

14. Цинциус В.И. Задачи сравнительной лексикологии алтайских языков // Очерки сравнительной лексикологии алтайских языков. Л.: Изд-во Наука. Ленингр. отдел-е, 1972. С.3-14.

ТАТАР ТЕЛЕННӘН БАСМА һӘМ ЭЛЕКТРОН ДӘРЕСЛЕКЛӘР

Р.К. Сәгъдиева
Казан федераль университеты, Казан
ramsag777@rambler.ru

В статье описывается новизна новых печатных и электронных учебников татарского языка, методика работы с этими пособиями, эффективное использование новых информационных технологии.

Ключевые слова: *татарский язык, электронный учебник, задания, учебный процесс.*

РФ Фән һәм мәгариф министрлыгының № 1559 (8.12.2014) “Федераль исемлеккә кергән дәреслекләргә таләпләр үзгәрүе турындагы” әмере нигезендә, 2015 нче елның 1 сентябреннән федераль исемлеккә кергән һәр дәреслекнең басма һәм электрон варианты булырга тиеш. Әлеге таләп, иң беренче чиратта, басма китаплар урынына электрон форманы кулланууга кертә башлауны, укучыларны калын-калын дәреслекләр күтәреп йөртүдән азат итүне, балаларны өстәмә материал белән тәэмин итүне һ.б. күз алдында тота. Россиянең кайбер төбәкләрендә нәкъ менә шундый дәреслекләрне файдалана да башлаганнар. Бүгенге көндә дәреслекнең электрон формасын куллану-кулланмау һәр мәктәп карамагына калдырыла. Мондый төр дәреслекләрнең уңайлыклары белән берәттән четерекле яклары да юк түгел. Мәсәлән, һәр укучының да планшет белән тәэмин ителгән булуы кирәк, техника ватылган очракта нишләргә кебек сораулар, әлбәттә, уйландыра.

Соңгы елларда теге яки бу фәннән, шул исәптән татар теленнән дә электрон кулланмалар эшләнелә, укытучыларга, укучыларга тәкъдим ителә. Татар теле белән бәйле мультимидеялы интерактив Интернет-дәреслекләр татар телендә сөйләшергә өйрәтүне максат итеп куя [3].

Рус телендә төп гомуми белем бирү оешмалары өчен 5-7 нче сыйныф дәреслекләре 2015 нче елда мәктәпләргә таратылды [1, 2, 4]. Әлеге дәреслекләрнең электрон форматы күптән түгел генә диск рәвешендә дөнья күрде. Төп шартларның берсе буларак, басма һәм электрон дәреслекләр эчтәлек, төзелеш һәм бизәлеш ягыннан бер-берсенә тулысынча тәңгәл килә.

Укучының белемнәрен системага салу, тирәнәйтү, яңа мәгълүмат бирү, укучыларның сөйләм һәм язу культурасын үстерү, аларны тәрбияләү – татар теле дәресләренең төп максаты. Бу максатка ирешүдә, укытучыларның иң беренче ярдәмчесе — дәреслек. Рус мәктәпләрендә укучы татар балалары өчен эшләнгән әлеге дәреслекләр Федераль дәүләт белем стандартлары (ФГОС) таләпләренә җавап бирә.

Теге яки бу дәрәҗәдә таныш теманы өйрәнгәндә, укучыларның белгәннәрен искә төшерүдән башлау яхшырак. Уку мәсьәләсен кую нәкъ менә шуны күздә тотып башкарыла.

Яна материалны да укучыларның элекке белемнәренә таянып аңлату уңай нәтиҗә бирә. Болай эшләү яңа темага җиңелрәк кереп китәргә, аны аңлап үзләштерергә ярдәм итә.

Дәресне проблемалы оештыру, кагыйдәләрне укучыларның үзләреннән чыгарту яки аларны шуңа якын китерү өчен дә дәреслектә материал бар. Грамматик формаларның телдәге кулланылышы да шулай мөстәкыйль күзәтүләр ярдәмендә үзләштерелә. Нәтиҗәләрнең дөреслеген күнегү азагында китерелгән аңлатма яки кагыйдә буенча тикшереп була. Укучыларның эзләнү һәм акыл эшчәнлеген укытучы тиешле юнәлештә алып барырга гына тиеш.

Дәреслектә билгеле бер күләмдә теоретик материал, күнегүләр, төрле төр иҗади эшләр, материалны кабатлау һәм ныгыту өчен сорау һәм биремнәр, схема һәм таблицалар бирелде. Теоретик белемнәрне үзләштерү һәм аларны ныклы күнекмәгә әверелдерү максатыннан, дәреслектә күнегүләрнең, иҗади эш төрләренең саны шактый арттырылды. Күнегүләр күп булган очракта, алар арасыннан иң кирәклеләрен сайлап, укучыларга тәкъдим итү укытучы карамагына калдырылды. Күнегүләр өчен текстлар мөмкин кадәр оригиналь әсәрләрдән алынды, аларда язучыларның тел үзенчәлекләре тулаем диярек сакланды.

Кагыйдәләр, язма биремнәр, сөйләм үстерү белән бәйле биремнәр, парлап эшләү өчен эш төрләре, өстәмә биремнәр шартлы билгеләр белән күрсәтелде. Дәреслек ахырында орфографик-орфоэпик, синонимнар, антонимнар һәм аңлатмалы сүзлекчәләр бирелде. Шулай ук бәйләнешле сөйләм үстерү дәресләрендә сочинениеләр язу өчен рус һәм татар рәссамнарының картиналары урнаштырылды.

Дәреслектә тел күренешләренең фәнни яктан дөрес яктыртылуына әһәмият бирелде, теоретик аңлатмаларны җиңеләйтүгә дә игътибар ителде. Укучыларның бирелгән материалны кабул итә алу мөмкинлекләре дә исәпкә алынды.

Укыту процессында дәвамчанлык мәсьәләсе дә зур әһәмияткә ия. Шуны истә тотып, дәреслектә алдагы сыйныфларда өйрәнелгән темалар буенча кыскача гына күзәтү ясала, укучыларның белемнәре искә төшерелә.

Рус һәм татар телләрен өйрәнү нәтиҗәсендә, укучылар телнең аралашу коралы булуына төшенәләр, аны милли мәдәни күренеш буларак аңлыйлар. Туган телне өйрәнү укучы өчен белем алуның төп нигезен тәшкил итә, аны уйлау һәм күзаллауга өйрәтүдә, иҗади мөмкинлекләрен үстерүдә төп чараларның берсе булып тора. Шул ук вакытта укучылар, телдән һәм язма формада аралашу өчен, төрле мәгълүмати мөмкинлекләрдән киңрәк файдалана белергә дә өйрәнәләр. Укучылар телдән һәм язма сөйләмнең дөреслеге кешенең гомуми культурасы үсеше дәрәҗәсен билгеләвен дә аңлыйлар. Алар туган телдәге башлангыч орфоэпик, лексик һәм грамматик төшенчәләрне, сөйләм әдәбе кагыйдәләрен үзләштерәләр.

Һәр уку елы ахырында укучылар хатасыз яза белүне гомуми үсеш дәрәҗәсенең күрсәткече буларак кабул итә; орфографик һәм орфоэпик кагыйдәләрне, тыныш билгеләрен урынлы куллана белү үз җөмләләрен төзегәндә һәм бирелгән җөмләләрне тикшергәндә кирәк булуына ышана; тел белеменең фонетика, лексикология, сүз ясалышы, морфология һәм синтаксис бүлеге буенча башлангыч мәгълүмат ала; һәр тел материалы алдагы сыйныфларда үтелгән фонетик, лексик, морфологик һәм синтаксик берәмлекләр белән үрелеп бара. Бу исә, үз чиратында, укучыга алга таба катлаулырак төшенчәләр белән эш итүдә таяныч була. Нәтиҗәдә, укучыда танып-белү эшчәнлеге белән кызыксыну барлыкка килә һәм ул алдагы сыйныфларда татар теле буенча алачак белемнәрнең нигезен тәшкил итәчәк.

Хәзерге көндә барлык укучылар да татар теленнән электрон дәреслеккә күчәргә әзер түгел. Шуңа күрә югарыда аталган дәреслекләрнең электрон варианты, иң беренче чиратта, татар теле укытучылары өчен тәкъдим ителә. Ул, укытучыларга куллану өчен, өстәмә материал белән тулыландырылды. Укытучы бер УМК белән эшләгән очракта, һәр дәрестә басма һәм электрон дәреслектән уңышлы файдалана алачак. Анда дүрт төрле бирем урын алды: аудиоязма, диктант, тест һәм физкультминутлар өчен видеороликлар.

Уку елы дәвамында күп кенә шигъри һәм чәчмә әсәрләрнең диктор тарафыннан укылышын тыңларга мөмкин. Әлеге төр биремнәр, беренче

чиратта, укучыларга авазларны дөрес әйтергә өйрәтсә, икенче чиратта, аларны әсәрне сәнгатьле итеп укырга да өйрәтә. Укытучы электрон дәреслекне экранга чыгармыйча, әлеге төр биремне кулланып, хәтер диктантлары да яздыра ала. Кайбер очракларда шул ук текстларны изложение яздыру максатыннан да файдаланырга мөмкин. Укучы бер генә укытучының түгел, ә диктор сөйләмен дә отып калырга күнегергә тиеш. Басма дәреслектә кайбер чәчмә текстлар ике өлешкә бүленеп бирелде. Алар барысы да укучыны тәрбияләүне максат итеп куялар. Укытучы электрон дәреслекнең билгеле бер күнегүендәге текстның алдагы өлешен тыңлата, һәм укучыга әсәрнең дәвамын сөйләп карарга, теге яки бу ситуациягә фикерен белдерергә кушыла. Укытучы укучыларга сораулар биреп, баланы сөйләштерә, аларның фикерләрен тыңлый. Шуннан соң басма дәреслекне ачып, укучылар язучының фикере, ягъни әсәрнең дәвамы белән танышалар. Ахырдан нәтиҗә ясау: укучының фикерен дөрес юнәлешкә бору бик мөһим.

Электрон дәреслектә аерым темалардан соң тестлар тәкъдим ителә: алар 23 темадан соң урын алган. Һәр сорауга дүрт төрле җавап варианты күрсәтелә, укучы шуннан дөрес җавапны сайлап ала һәм җавапны саклый. Алдагы сорау автомат рәвештә үзеннән үзе ачыла һәм шул тәртиптә эш дәвам итә. Ахырдан соңгы биттә җавапларның нәтиҗәләре чыга, ягъни максимум ничә балл җыеп була иде, укучы күпме балл туплаган, ничә процент нәтиҗәлелеккә ия, барысы да ачык чагыла. Шунда ук җавапларны, хаталарны карау мөмкинчелеге тудырылган: бала һәр биремнең дөрес җаваплары белән таныша, анализ ясала, онытылган яисә начар үзләштерелгән теманы ныгыту эше алып барыла.

Укытучыларга технологик картаны үз эченә алган методик әсбаплар да тәкъдим ителә. Алар Федераль дәүләт белем стандартлары (ФГОС) таләпләре буенча төзелгән программаны җентекләбрәк аңлату һәм укытучыга дәреслек белән эшләүне җиңеләйтү максатларын күздә тотып төзелде. Ул 2 зур бүлектән тора. Беренче бүлек аңлатма язуын, укытуның көтелгән нәтиҗәләрен, укучыларда формалаштырылырга тиешле универсаль уку-укыту гамәлләрен, эчтәлекнең темаларга һәм сәгатьләргә бүленешен, ел ахырында укучылардан таләп ителгән күнекмәләрне, үтәлергә тиешле язма эшләрнең күләмен, аларны бәяләү нормаларын үз эченә алган эш программасыннан тора. Икенче бүлектә дәрес эшкәртмәләре Федераль дәүләт белем стандартлары (ФГОС) таләбе буенча технологик карталар формасында бирелә. Әлеге дәрес эшкәртмәләре үрнәк буларак китерелә: укытучы эш алымнарын үзенчә үзгәртә, төрләндерә ала. Анда һәр дәреснең технологик картасы урын алган. Методик яктан, һәр дәрестә динамик пауза булу шарт. Методик әсбапта кайсы күнегүне эшләгәннән соң физкультминут үткәрелергә

тиешлек күрсәтелеп барыла. Әлбәттә, бу авторларның тәкъдиме генә, укытучы һәр дәрескә технологик картаны яисә дәрес планын үзенчә эшләргә хокуклы. Ләкин электрон дәреслектә нәкъ менә методик әсбапта күрсәтелгән күнегүләрнең һәрберсе янында билгеле бер тамга белән, физкультминутлар ясату максатыннан, видеороликлар урын алды. Дәреслектә 35 төрле видеоролик кулланылды, шуларның кайберләре берничә урында кабатланырга мөмкин. Бу очракта диктор сәнгатьле итеп шигырьне укый, бер укучы (егет яисә кыз) шундый төр күнегүләрне башкаралар. Укытучы һәр дәреснең уртасында видеороликны күрсәтеп, укучыларга физкультминут эшләтә ала.

Электрон дәреслекнең тагын бер уңай ягы: укытучы һәр дәрес ахырында сүзлек яисә сүзтезмә диктантлары яздыру мөмкинлегенә ия. Технологик картада күрсәтелгән һәр дәрес ахырында диктантлар бар. Алар һәрвакыт 6 сүзне яисә сүзтезмәне үз эченә алган. Дәрес дәвамында кулланылган текстларда очраган авыр сүзләрне укучы исендә калдырып, диктант язып карый ала. Аларны диктор укый, укучы әйтелешне тыңлый һәм яза. Уң яктагы түгәрәккә басып, җавабын тикшерә. Дөрес язылган очракта яшел дөрес дигән тамга чыга, әгәр хата киткән булса, кызыл х тамгасы пәйда була. Бу очракта укучы яңадан дикторны тыңлый ала һәм кабаттан яза, үз-үзен тикшерә. Берничә дәрес дәвамында укытучы шундый төр биремне эшләткәннән соң, укучылар алга таба дәрес барышында текстларның һәр сүзен игътибар белән укып, язылышын исендә калдырып барырга омтылачак. Бер дәрестә язган сүз алдагы дәресләрдә очрамый. Һәр дәрес ахырында укучы яңа сүзнең язылышын исендә калдыра.

Шунысы кызык: теге яки бу дәрес вакытында диктант яздырганнан соң яисә тестларны эшләткәннән соң, электрон дәреслек кире шул биткә әйләнеп кайта. Укытучы да, укучыда яңадан әлеге битне эзләп утырмый. Бу бик уңайлы, укытучының эшен нык җиңеләйтә, системаның уңайлы һәм төгәл эшләнгәнлеге хакында сөйли.

Электрон дәреслек Интернет белән бәйләнмәгән, ягъни аны ачып эшләткәндә, Интернет кирәк түгел.

Алда әйтелгәнчә, электрон дәреслек басма дәреслекне сүзгә сүз, биткә бит кабатласа да, анда укытучы һәр дәрестә куллана алырлык кызыклы биремнәр һәм эш төрләре чагылыш тапкан. Яңа теманы аңлату өчен дә әлеге төр дәреслекләр уңайлы, тактада ачып укучыга юнәлеш бирелә, бала эзләнә, күзәтә, уйлана һәм нәтиҗә ясый. Аннары гына укытучы теге яки бу кагыйдәне экранда күрсәтә, һәм укучы үзенең нәтиҗәсе белән грамматик кагыйдәләрне чагыштыра, фикер алышуда катнаша.

Заман үзгәрә, үсә, шуның белән беррәттән һәр нәрсә яңалык белән баетыла. Басма дәреслекләрнең электрон төре укучыларга да,

укытучыларга да файдалы булыр дип өметләнәсе килә. Әлеге яңа төр форма җәмгыятебездә урнашып кына бара. Вакыт узу белән электрон дәреслекләр тагын да камилләшәчәк, укучының белем дәрәҗәсен арттыруда мөһим роль уйнаячак.

Әдәбият

1. Сәгъдиева, Р.К. Татар теле: рус телендә төп гомуми белем бирү оешмалары өчен дәреслек (татар телен туган тел буларак өйрәнүче укучылар өчен) 6 нчы с-ф. / Р.К.Сәгъдиева, Р.М.Гарәпшина, Г.И.Хәйруллина. – Казан: “Мәгариф-Вакыт” нәшр., 2015. – 191 б.

2. Сәгъдиева, Р.К. Татар теле: рус телендә төп гомуми белем бирү оешмалары өчен дәреслек (татар телен туган тел буларак өйрәнүче укучылар өчен) 7 нче с-ф. / Р.К.Сәгъдиева, Г.Ф.Харисова, Л.К.Сабиржанова, М.Ә.Нуриева – Казан: “Мәгариф-Вакыт” нәшр., 2015. – 215 б.

3. Сулейманов Д.Ш., Гильмуллин Р.А., Хасанова Л.Р. Интерактивный Интернет-учебник по татарскому языку «Татар теле онлайн» // Эл. журнал "Образовательные технологии и общество", № 1, 2011. Спец. раздел выпуска под ред. акад. АН РТ, дир. НИИ «Прикладная семиотика» АН РТ, проф. КФУ Д.Ш. Сулейманова / http://ifets.ieee.org/russian/depository/v14_i1/pdf/10r.pdf

4. Шәмсетдинова, Р.Р. Татар теле: рус телендә төп гомуми белем бирү оешмалары өчен дәреслек (татар телен туган тел буларак өйрәнүче укучылар өчен) 5 нче с-ф. / Р.Р.Шәмсетдинова, Г.К.Һадиева, Г.В.Һадиева – Казан: “Мәгариф-Вакыт” нәшр., 2015. – 175 б.

РАЗМЕТКА МОРФОЛОГИЧЕСКИХ КАТЕГОРИЙ В НАЦИОНАЛЬНОМ КОРПУСЕ КЫРГЫЗСКИХ ТЕКСТОВ

Ташполот САДЫКОВ, Бакыт ШАРШЕМБАЕВ
Бишкекский гуманитарный университет имени К. Карасаева,
tash_sadykov@mail.ru
Кыргызско-турецкий университет “Манас”, Кыргызстан,
bakyt101@mail.ru

Идентификация категорий частей речи и их распределения в текстах является одной сложнейших проблем тюркологии. В отличие от флективных языков в тюркских языках слова и словоформы чаще подвергаются явлениям конверсии и омономии. Для разметки морфологических категорий слов и словоформ в корпусе, а также для снятия текстовой омонимии, предлагается система морфологической разметки, совместимой с унифицированной системой для национального корпуса тюрских текстов.

Аанализируя существенные единицы (слово, словоформа, корень, приставка) используемые в тексте, *морфологические теги* (*morphological tegs*) являются условной системой обозначения морфологических категорий. Руководством для обозначения разметок морфологических тегов национального корпуса кыргызского языка является стандарт обозначений, принятый ассоциацией построения национальных корпусов тюркоязычных стран (г.Казань) [Садыков, Шаршембаев 2014: 140-147; http://www.eva.mpg.de/lingua/resources/glossingrules.php; http://ips.antat.ru/page.php].

Тексты, размеченные системой тегов, проходят сначала морфологический, а затем синтаксический, семантический, прагматический анализ и преобразуются в корпус понятный компьютерному языку. Говоря точнее, такой национальный корпус формализованный для удобства автоматической обработки текстов компьютером, для пользователя будет являться *национальным корпусом* или *базой знаний* (*knowledge base, intelligent database)* извлечения ценной лингвистической, когнитивной, этнологической, лингвокультурной инофрмации.

Так как в родственных тюрских языках работа по унификации не разаработана, одинаковые морфологический категории в разных языках обычно обозначаются по разному. Ученые-языковеды, лингвисты, инженеры программисты занятые составлением национальных корпусов хорошо понимают необходимость унификации тегов используемых для разметок текстов. И в то же время появляется необходимость унификации тегов для обозначения одинаковых явлений не только родственных, но и неродственных языков.

Созданная по инициативе Национальной Академии Республики Татарстан и Казанского федерального университета ассоциация построения национальных корпусов тюрских языков (Казань, 2014) приняла стандарт тегов соотвествующая типологическому стандарту Лейпцига, которая помечает каждое слово текста принадлежащей той или иной части речи и обозначается следующими тегами [http://www.eva. mpg.de/ lingua/ resources/ glossingrules.php].

Теги **tegs**	**Название** **full term**	**Часть речи** **parts of speech**
N	noun	существительное
ADJ	adjective	прилагательное
V	verb	глагол
ADV	adverb	наречие
NUM	numeral	числительное
PN	pronoun	местоимение

CNJ	conjunction	союз
POST	postposition	предлог
PART	particle	частица
INTRJ	interjection	междометие
MOD	modal word	модальное слово
IMIT	imitative word	подражательное слово

Наряду с этим для обозначения морфологических категорий каждого слова и словоформы текста предлагается следующая система тегов:

Числительные категории – Number
1. Единственное число – singular
2. Множественное число – plural
Теги:
1. **SG <=> ∅**
∅
2. **PL <=> ЛАр**
-лар -дар -тар
-лер -дер -тер
-лор -дор -тор
-лөр -дөр -төр.

Притяжательные категории – Possessive
Единственное число – singular:
1. первое лицо единственного числа - 1st person singular possessive (‘my’),
2. второе лицо единственного числа - 2nd person singular possessive (‘your’),
3. второе лицо единственного числа ласк. - 2nd person sing. po**ss.** formal (‘your’),
4. третье лицо единственного числа - 3nd person singular possessive (‘his/her/its’),
Множественный падеж – plural:
5. первое лицо множественного числа - 1st person plural possessive (‘our’),
6. второе лицо множественного числа - 2nd person plural possessive (‘your’),
7. второе лицо множественного числа ласк.- 2nd person pl. po**ss.** formal (‘your’),
8. третье лицо множественного числа - 3nd person plural possessive (‘their’),

Теги:

1. **POSS_1SG <=> [Ы]м**
   -ым -им -ум -үм
   -м;
2. **POSS_2SG <=> [Ы]ң**
   -ың -иң -уң -үң
   -ң;
3. **POSS_2SGF <=> [Ы]ң[Ы]з**
   -ыңыз -иңиз -уңуз -үңүз
   -ңыз -ңиз -ңуз -ңүз;
4. **POSS_3SG <=> [с]Ы[н]**
   -ы -и -у -ү -ын -ин -ун -үн
   -сы -си -су -сү -сын -син -сун -сүн;
5. **POSS_1PL <=> [Ы]б[Ы]з**
   -ыбыз -ибиз -убуз -үбүз
   -быз -биз -буз -бүз;
6. **POSS_2PL <=> [Ы]ң[А]р**
   -ыңар -иңер -уңар -үңөр
   -ңар -ңер -ңар -ңөр;
7. **POSS_2PLF <=> [Ы]ң[Ы]зд[А]р**
   -ыңыздар-иңиздер -уңуздар -үңүздөр
   -ңыздар -ңиздер -ңуздар -ңүздөр;
8. **POSS_3PL <=> [с]Ы[н]**
   -ы -и -у -ү
   -ын -ин -ун -үн
   -сы -си -су -сү
   -сын -син -сун -сүн;

Падежные категории - Noun Cases

1. Именительный падеж – nominative,
2. Родительный падеж – genitive,
3. Дательный (направительный) падеж – dative,
4. Винительный падеж – accusative,
5. Местный падеж – locative,
6. Исходный падеж – abblative.

Теги:

1. NOM <=> ∅
∅

2. GEN <=> [н]Ын
-нын -нин -нун -нүн

-дын -дин -дун -дүн
-тын -тин -тун -түн
-ын -ин -ун -үн;

3. DAT <=> [Г]А

-га -ге -го -гө
-ка -ке- ко –кө
-а -е- о -ө;

4. ACC <=> [н][Ы]

-ны -ни -ну -нү
-ды -ди -ду -тү
-ты -ти -ту -тү
-ы -и -у -ү;

5. LOC <=> ДА

-да -де -до -дө
-та -те -то -тө;

6. ABL <=> [Д]Ан

-дан -ден -дон -дөн
-тан -тен -тон -төн
-ан -ен -он -өн;

Личные категории - Personal

Единственное число – singular:
1. первое лицо единственного числа - 1st person singular,
2. второе лицо единственного числа - 2nd person singular,
3. второе лицо единственного числа вежл. - 2nd person singular formal,
4. третье лицо единственного числа - 3nd person singular,

Множественное число – plural:
5. первое лицо множественного числа - 1st person plural,
6. второе лицо множественного числа - 2nd person plural,
7. второе лицо множественного числа - 2nd person plural formal,
8. третье лицо множественного числа - 3nd person plural,

Теги:

1. **1SG <=> м[Ы]н**

-мын -мин -мун -мүн;

2. **2SG <=> с[Ы]ң**

-сың -сиң -суң -сүң;

3. **2SGF <=> с[Ы]з**

-сыз -сиз -суз -сүз;

4. **3SG <=> ∅**

∅;

5. **1PL <=> б[Ы]з**
-быз -биз -буз -бүз;
6. **2PL <=> с[Ы]ң[А]р**
-сыңар -сиңер -суңар -сүңөр;
7. **2PLF <=> с[Ы]зд[А]р**
-сыздар-сиздер -суздар -сүздөр;
8. **3PL <=> [с]Ы[н]**
∅;

Имя прилагательное – adjective: сравнительная степень – comparative

Теги:
COMP <=> [Ы]рААК
-ыраак -ирээк -ураак -үрөөк
-раак -рээк -раак -рөөк;

Числительное – Numeral
1. Порядковое числительное - ordinal numeral
2. Собирательное числительное - collective numeral,
3. Приблизительное числительное1 – approximate numeral1,
4. Приблизительное числительное2 – approximate numeral2,
5. Приблизительное числительное3 – approximate numeral3,

Теги:
1. NUM_ORD <=> [Ы]нчЫ
-ынчы -инчи -унчу -үнчү
-нчы -нчи -нчу -нчү;
2. NUM_COLL <=> ОО[н]
-оо -өө
-оон -өөн;
3. NUM_APPR1 <=> чА
-ча -че -чо -чө;
4. NUM_APPR2 <=> ДАй
-дай -дей -дой -дөй
-тай -тей -той -төй;
5. NUM_APPR3 <=> ДАгАн
-даган -деген -догон -дөгөн
-таган -теген -тогон -төгөн.

Глагол – Verb: Залоговое категория - Voices

1. Основной залог – active,
2. Страдательный залог – passive,
3. Рефлексивный залог – reflexive,
4. Посредственный залог – causative,
5. Возвратный залог – reciprocal.

Теги:

1. ACT <=> ∅

∅

2. ACT <=> [Ы]л||н

-ыл -ил -ул -үл

-л;

-ын -ин -ун -үн

-л;

3. REFL <=> [Ы]н

-ын -ин -ун -үн

-л;

4. CAUS <=> Д[Ыр]

5. RECP <=> [Ы]ш

-ыш -иш -уш -үш

-ш;

Повелительное наклонение - Imperatives

1. Первое лицо единственного числа – Hortative: 1st person singular – ‘let me’,
2. Первое лицо множественного числа - Hortative:1st person plural – ‘let’s’,
3. Второе лицо единственного числа вежл. – Imperative: 2nd person singular,
4. Второе лицо множественного числа вежл. - Imperative: 2nd person plural,
5. Второе лицо единственного числа форм.– Imperative: 2nd person singular formal,
6. Второе лицо множественного числа форм. - Imperative: 2nd person plural formal,
7. Третье лицо единственного числа – Jussive: 3rd person singular – ‘let him/her/it’,
8. Третье лицо множественного числа - Jussive: 3rd person plural – ‘let them,
9. Просительный вежл. - precative (‘please’).

Теги:

1. HOR_SG <=> [А]йЫн

-айын -ейин -ойун -өйүн

-йын -йин -йун -йүн;

2. HOR_PL <=> [А||й]лЫ[к]

-алык -елик -олук -өлүк

-йлык -йлик -йлук -йлүк;

-алы -ели -олу -өлү

-йлы -йли -йлу -йлү;

3. IMP_SG <=> ГЫн

-гын -гин -гун -гүн

-кын -кин -кун -күн;

4. IMP_ PL <=> ГЫлА

-гыла -гила -гула -гүла

-кыла -кила -кула -күла;

5. IMP_SGF <=> [Ы]ңЫз

-ыбыз -ибиз -убуз -үбүз

-быз -биз -буз -бүз;

6. IMP_PLF <=> [Ы]ңЫздАр

-ыңыздар-иңиздер -уңуздар -үңүздөр

-ңыздар -ңиздер -ңуздар -ңүздөр;

7. JUS_SG <=> сЫн

-сын -син -сун -сүн;

8. JUS_PL <=> [Ыш]сЫн

-ышсын -ишсин -ушсун -үшсүн

-сын -син -сун -сүн;

9. PREC_1 <=> чЫ

-чы -чи -чу -чү;

Категория времени - Verb tenses

1. Настоящее время – present,
2. Определенное прошедшее время - past definite,
3. Неопределенное прошедшее время - past indefinite,
4. Прошедшее неожиданное время – past evidentiality,
5. Обыкновенное прошедшее время – past iterative,
6. Определенное будущее время - future definite,
7. Неопределенное будущее время - future indefinite,
8. Неопределенное отрицательное будущее время - future indefinite negative.

Теги:

1. **PRES <=> [А||й]**
   -а -е -о -ө
   -й;
2. **PST_DEF <=> ДЫ**
   -ды -ди -ду -дү
   -ты -ти -ту -тү;
3. **PST_INDF <=> ГА[н]**
   -ган -ген -гон -гөн
   -кан -кен -кон -көн;
   -га -ге -го -гө
   -ка -ке -ко -кө;
4. **PST_EVID <=> чУ**
   -чу -чү;
5. **PST_ITER <=> [Ы]п[тЫр]**
   -ыптыр -иптир -уптур -үптүр
   -птыр -птир -птур -птүр;
   -ып -ип -уп -үп
   -п;
6. **FUT_DEF <=> [А||й]**
   -а -е -о -ө
   -й;
7. **FUT_INDF <=> [А]р**
   -ар -ер -ор -өр
   -р;
8. **FUT_INDF_NEG <=> БАс**
   -бас -бес -бос -бөс
   -пас -пес -пос -пөс;

Аспект – aspect

1. Отрицательный – negative,
2. Вопросительный – interrogative.

Теги:

1. **NEG <=> БА**
   -ба -бе -бо -бө
   -па -пе -по -пө;
2. **INT <=> БЫ**
   -бы -би -бу -бү
   -пы пи -пу -пү.

Причастия – Participles

1. Причастие настоящего времени – present participle
2. Причастие прошедшего времени – past participle,
3. Причастие будущего времени - future participle,
4. Причастие отрицания будущего времени - future participle negative.

Теги:

1. PCP_PR <=> [УУ]чУ
-уучу -үүчү
-чу -чү;

2. PCP_PS <=> ГАн
-ган -ген -гон -гөн
-кан -кен -кон -көн;

3. PCP_FUT_DEF <=> [А]р
-ар -ер -ор -өр
-р;

4. PCP_FUT_ NEG <=> БАс
-бас -бес -бос -бөс
-пас -пес -пос -пөс;

Деепричастия - Converbs

1. Сопровождающие деепричастия - Adverbial verb (accompanist),
2. Долгие деепричастия - Adverbial verb (continuing),
3. Целевые деепричастия - Adverbial verb (Intentional),
4. Отрицательная форма деепричастия - Adverbial verb (negative form),
5. Последовательные деепричастия - Adverbial verb (successive meaning),
6. Ограничительные деепричастия - Adverbial verb (limiting).

Теги:

1. ADVV_ACC <=> [Ы]п
-ып -ип -уп -үп
-п;

2. ADVV_CONT <=> [А||й]
-а -е -о -ө
-й;

3. ADVV_INT <=> ГАнЫ
-ганы -гени -гону -гөнү
-каны -кени -кону -көнү;

4. ADVV_NEG <=> мАйЫн[чА]
-майынча -мейинче -мойунча-мөйүнчө;
-майын -мейин -мойун -мөйүн;

5. ADVV_SUC <=> ГЫчА

-гыча -гиче -гуча -гүчө

-кыча -киче -куча -күчө;

6. ADVV_ SUC <=> ГАнчА

-ганча -генче -гончо -гөнчө

-канча -кенче -кончо -көнчө.

Отглагольное существительное – Verbal nouns (masdars)

1. Отглагольное существительное -оо – infinitive 1,
2. Отглагольное существительное -уу – infinitive 2,
3. Отглагольное существительное -ыш – infinitive 3,
4. Отглагольное существительное -мак – infinitive 4,
5. Отглагольное существительное -гы – infinitive 5.

Теги:

1. INF_1 <=> ОО

-оо -өө;

2. INF_2 <=> УУ

-уу -үү;

3. INF_3 <=> [Ы]ш

-ыш -иш -уш -үш

-ш;

4. INF_4 <=> мАГ

-мак -мек -мок -мөк

-маг -мег -мог -мөг;

5. INF_5 <=> ГЫ

-гы -ги -гу -гү

-кы -ки -ку -кү;

Модальные формы - Modal forms

1. Условно модальные – conditional,
2. Модальности намерения (желательность) - desiderative (intention),
3. Желательно модальные – optative1,
4. Желательно модальные – optative2,
5. Сомнительные модальные – premonitive (warning).

Теги:

1. COND <=> сА

-са -се -со -сө;

2. DESIDE <=> мАк[чЫ]

-макчы -мекчи -мокчу -мөкчү

-мак -мек -мок -мөк;

3. ОРТ <=> ГЫ+POSS келет||келди
-гы -ги -гу -гү
-кы -ки -ку -кү;
4. ОРТ <=> ГАй эле+PERS
-гай -гей -гой -гөй
-кай -кей -кой -көй;
5. PREM <=> БАГАй эле+PERS
-багай -бегей -богой -бөгөй
-пагай -пегей -погой -пөгөй;

Литература

Садыков Т., Шаршембаев Б. Система морфологической разметки для корпуса кыргызских текстов // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. - Казань: Изд-во Фан АН РТ, 2014, с. 140-147.

http://ips.antat.ru/page.php;http://www.eva.mpg.de/lingua/resources/glossingrules.php]. [http://www.eva. mpg.de/lingua/resources/glossingrules.php].

ФОРМИРОВАНИЕ КОРПУСА С РАЗМЕТКОЙ СУЩНОСТЕЙ В НОВОСТНЫХ МЕДИА РЕСУРСАХ ДЛЯ КАЗАХСКОГО ЯЗЫКА

З.Н. Садыкова, В.В. Иванов
Казанский федеральный университет, Казань
Sadykovazn@gmail.com

В статье описывается процесс построения корпуса с разметкой именованных сущностей в новостных медиа-ресурсов для казахского языка. Рассмотрены основные признаки для извлечения именованных сущностей в тексте. Описаны правила выделения объектов с использованием онлайн-инструмента для разметки brat. В настоящее время ведется аннотирование коллекции новостных материалов казахского языка, собранных сотрудниками Назарбаев Университета, однако работа по завершению сбора размеченного корпуса еще не окончена.

Ключевые слова: корпус, разметка, именованные сущности, распознавание именованных сущностей.

Введение

За последние годы растет интерес к формированию корпусов текстов на национальных языках [5]. Интенсивно идет разработка корпусов

английского и русского языка, содержание текстов таких корпусов достигает примерно миллионов словоупотреблений [4]. Созданные массивы текстов могут неоднократно использоваться многими исследователями в решении различных задач. На сегодняшний день сформированные и аннотированные текстовые массивы казахского языка имеются в ограниченном количестве. Перечислим корпуса казахских текстов, существующие на данный момент:

3. Казахский национальный корпус (сайт: http://dawhois.com/www/til.gov.kz.html). Данный корпус является одним из первых, но не размечен и очень маленький.
4. Алматинский корпус казахского языка (сайт: http://web-corpora.net/KazakhCorpus/search/?interface_language=ru). Размер корпуса составляет около двух миллионов словоупотреблений. Считается первой версией Национального корпуса казахского языка - НККЯ как справочно-информационная система на основе обширного фонда размеченных текстов литературного казахского языка, государственного языка Республики Казахстан. Размещен на сайте в открытом доступе.
5. KLC (сайт: http://link.springer.com/chapter/10.1007%2F978-3-642-54903-8_44). Корпус казахского языка, созданный сотрудниками Назарбаев Университета, содержит более 135 миллионов слов. Makazhanov, Aibek, et al. "Syntactic annotation of Kazakh: following the universal dependencies guidelines. A report." *TurkLang-2015*: 338.

Предоставлен не полный список всех собранных текстовых массивов на казахском языке. Имеются ряд мелких специализированных корпусов. Нашей конечной целью является формирование аннотированного множества казахских текстов для дальнейшего использования в научно-исследовательской работе.

Формирование размеченного корпуса

Под разметкой понимается извлечение структурированной информации из неструктурированных массивов текста. Поскольку наиболее информативными являются именно размеченные корпуса, то при формировании изначально большую роль играет именно разметка. Создание аннотированного корпуса - трудоемкий процесс, который занимает большое количество времени и человеческих ресурсов. В настоящее время наиболее распространено ручное аннотирование, поскольку оно обладает высоким качеством результирующей разметки.

Существует несколько видов разметки [3]:

- морфологическая;
- лемматизация;
- синтаксическая;
- дискурсивная;
- семантическая;
- морфо-синтаксическая.

В данной работе выполнена семантическая разметка текста для выделения именованных сущностей на базе новостных статей. Именованные сущности представляют собой объекты и факты, наделенные определенной значимой информацией. Коллекция исходных документов, предоставленная сотрудниками Назарбаев Университета (Nazarbaev University), собрана из 8 официальных новостных сайтов Республики Казахстан. Данная коллекция содержит 56000 текстов на казахском языке. Типы именованных сущностей, использовавшиеся в разметке, определены в Таблице 1.

Таблица 1

Типы именованных сущностей

№	Тип	Описание
1	Per	Ф.И.О. человека, имена вымышленных персонажей, прозвище, псевдоним
2	Org	Организации, компании, общественные объединения.
3	Loc	Природные георгафические объекты.
4	GPE	Геополитические объекты.
5	Event	Общественные события, акции, мероприятия.
6	Award-name	Награды, степени,звания.
7	Tender	Конкурс, тендеры.

Разметка

В рамках выполненных работ были созданы правила для аннотирования и сформирована инструкция по разметке текстов. Разметка именованных сущностей проводилась в три этапа. Работа осуществлялась с помощью онлайн-инструмента для разметки письменных текстов brat (brat rapid annotation tool).

Первый этап состоял из обработки более 1000 документов одним аннотатором. В ходе выделения именованных сущностей, встречались объекты, которые относились к нескольким типам сущности. Выбор

правильного типа определялся по контексту, а так же от наличии вспомогательных слов. Например, в предложении *“Еуразия” бірінші телеарнасы* (первый канал “Евразия”) слово “Евразия” относится к ORG (организация), а в предложении *Еуразия құрлығында өтті (проводилось на материке Евразия)* слово *“Евразия”* к *GPE (*геополитический объект*)*. В случаях, когда по контексту выбор не определялся, то объекту одновременно присваивались разные категории сущностей.

Второй этап заключался в проверке выборочных текстов независимым экспертом. Из 100 документов были найдены и исправлены ошибки в 20 статьях. Эта статистика показывает, что процент ошибок был сравнительно небольшим.

На третьем этапе необходимо было произвести оценку качества размеченных текстов. К решению задачи был подключен еще один аннотатор, которому предстояло осуществить поиск именованных сущностей в исходной коллекции документов. На текущий момент работа над третьим этапом продолжается. Результат оценки меры согласия двух аннотаторов за последние две недели представлен в Таблице 2.

Таблица 2

Оценка качества разметки

	Коэффициент согласия между аннотаторами
1 неделя	0,86
2 неделя	0,89

Анализ разметки

На рис.1. показано соотношение каждой категории сущности ко всем выделенным объектам в размеченном корпусе. В новостных статьях наиболее часто встречались имена персон (1084), наименование географических объектов (974), организаций (973). События, связанные с присуждением наград и объявлением тендеров появлялись во всем корпусе редко.

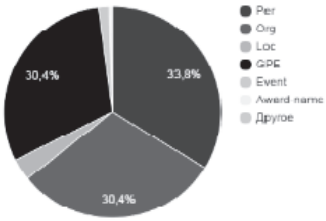


рис.1

Заключение

В работе представлены результаты формирования корпуса новостных текстов с разметкой именованных сущностей на казахском языке. Размеченное множество состоит из более чем 1000 документов и содержит около 3000 объектов. Результат может быть востребован в дальнейших научных исследованиях и системах интеллектуального анализа неструктурированной информации.

ТАРИХИ ЭЧТӘЛЕКЛЕ ФРАЗЕОЛОГИК БЕРӘМЛЕКЛӘР (Н. Исәнбәтнең “Татар теленең фразеологик сүзлеге” материалында)

Ф.Р. Сибгаева
Казан федераль университеты

Идел буе Болгар дәүләтендә ислам динен кабул иткән вакытлардан алып XX гасырның утызынчы елларына кадәр барлыкка килгән фразеологик берәмлекләр татар теленең фразеологик берәмлекләрендә күпләп урын алганнар. Әлеге фразеологик берәмлекләр теге яки бу тарихи вакыт, факт, шәхес турында күзалларга ярдәм итә. Тел кануннарына, сөйләмгә ят фразеологизмнар кулланыштан төшеп калган, бары тик халык иләгеннән иләнгәннәре генә сүзлекләргә теркәлеп калган.

Тарихны чагылдырган фразеологизмнар тарихи чор, вакыйгалар, тарихта булган шәхесләр исемнәре белән бәйле. Ул шәхесләрнең исемнәре я турыдан-туры, я берәр конкрет вакыйга белән бәйле булган.

Тарихи вакыйгаларга караган фразеологиянең бер өлеше берәр билгеле булмаган урын (шәһәрме, авылмы, илме) белән бәйләнгән. Борынгы мәшһүр шәһәрләр исемнәре белән бергә хәзерге Татарстан, яки башка өлкәләрдәге авыл исемнәре дә очрый. Билгеле бер урын исеме белән бәйле берәмлекләр бер авыл яисә шәһәрдә җәмәгатьчелек игътибарын уяткан берәр хәл булганнан соң туган. Башта ул, күбесенчә мәзәкле хәлләр, башта шул авылдагы кешеләр арасында, тора-бара күршедәге авылларга да таралып, бөтен өлкә яисә ил яңалыгына әверелгән. Шулай итеп, ул тедән-телгә, кешедән-кешегә, буыннан-буынга күчеп, тел җәүһәренә әйләнгән.

Шәм шәриф (2 т., 249 б.). Шәм шәриф дип, хәзерге Дәмәшк шәһәрен атап йөрткәннәр. Дәмәшк – 661-750 нче елларда Әмәвиләр (Омейяды) гарәп хәлифәлегенең башкаласы булып торган. Дәмәшк шәһәрендә борынгы заманнарда ук өтен Шәрык дөньясына һәм Көнбатышта дан казанган корыч коелган. Көнбатышта ул – «булат», «булатная сталь» исеме астында билгеле булган. Бу хезмәтләрендә язып калдырган. Әлеге корычны кою сере урта гасырларда югалган.

Фразеологизмнар арасында Казан ханлыгы чорына караганнары да бар. Мәсәлән, *калага урыс керде, чыкты* (1т., 340 б.). Аның тарихы әлеге хәлләргә карый: Касыйм ханлыгы 1450 нче елларда Василий Икенче тарафыннан рус дәүләтенә хезмәт күрсәткән татар бәкләренә бирелгән. Касыйм тәхетенә ханнар, Казан, Кырым, Себер ханлык ларыннан чакырылып, руслар тарафыннан билгеләнгән. Касыйм ханнары еш кына Казанга яу чабуларда Мәскәү ягында сугышканнар (1467-1469, 1487, 1552 еллар). Касыйм ханлыгыннан Мәскәүгә яраклы Шаһгали кебек ханнарны Казан тәхетенә утырталар (Шаһгали 3 тапкыр хан булган). Мәскәү 1552 елдан соң Казан ханлыгын үзенә кушкач, Касыйм ханлыгы белән идарә итү тулысы белән рус гаскәр башлыкларына күчкән, ә Касыйм бәкләре җир биләүчеләргә әйләнгәннәр.

Әби патша заманында (2т., 305 б.). Фразеологизм «бик күптән булган хәлләр» дигән мәгънәне аңлата. Әби патша дип татарлар 1762 нче елда Россия тәхетенә утырган императрица Екатерина Икенчене яратып атап йөрткәннәр. Татар халкының хәтерендә Әби патша күп яхшылык эшләгән патшабикә булып калган. Ул идарә иткән дәвердә татарлар күп кенә уңай үзгәрешләр кичергән: 1783 нче елда Екатерина дин иреге турындагы законга кул куйган, 1788 нче елда Татар диния нәзарәте ачылган, 1784 нче елдан алып татар морзалары рус дворяннарында булган бар хокукларга ия була алганнар, татарларга икътисад, промышленность белән шөгыльләнү иреге бирелгән.

Татарның хәрби хезмәттә булуына карый торган фразеологик берәмлекләр дә лексикографик чыганакларда күпләп теркәлгән. Бу

фразеологизмнар патша Россиясендә хәрби эшнең ничек баруын берникадәр күзалларга ярдәм итә.

Кара яу (1 т., 355 б.) бик күп булып басып килгән дошман явын аңлаткан фразеологизм.

Тел алу. Тел тотып китерү (2 т., 123 б.). Әлеге фразеологизм сугышта дошман гаскәреннән кеше тотып алып кайтып сөйләшеп, хәрби серләрен алу дигән мәгънәне белдерә. «Борынгы төрки халыкларда бик борынгыдан калган тәгъбир. Мәсәлән, VIII гасырдан калган Төньюкук язма ташында «Тылыг келтерде» (телне тотып китерде) дип язылган» [1: 176].

Угылың Ырумга, кызың Кырымга [китсен]. Әлеге теләкнең мәгънәсе: угылың Ырумга — Рим империясенә, тәре походларына каршы сугышка китсен, кызың Алтын Урданың резиденциясе булган Кырым бәкләренә кияүгә китсен. Тәре походларын XII гасыр урталарында Көнчыгышта мәмлүкләр туктаталар. Тәре походлары христианнарның изге җирләрен мөселманнардан азат итүгә сылтау итеп оештырыла. Мөселманнар кулына Акра шәһәре күчкәч, тәре походларында катнашкан гаскәр Көнчыгышта үзенең көчен бөтенләй югалта. Мәмлүкләр чирүенең нигезен, күпчелеген төрки чыгышлы сугышчылар тәшкил иткән. Мәмлүкләр XIII гасырдан алып XIX гасыр башына кадәр Мисыр Сирия җирләре белән идарә иткәннәр. Тәре йөртүчеләргә каршы көрәшнең башында мәшһүр мәмлүк солтаны Би Барс тора.

Кырым чирүе күк (1 т., 433 б.). Әлеге фразеологизм кеше күплекне, ишле гаиләне аңлата. Фразеологизмның тарихы XVI гасырда Казан ханлыгында барган хәлләр белән бәйле. 1518 нче елда Мәскәү князьлегендәге касыйм татары Шаһгали хан Казан тәхетенә утыртыла. Казанлылар аны яратмый. Нурсолтан бикәнең Кырым ханы Миңлегәрәйдән туган улы Сәхибгәрәйне Казанга хан итеп чакырып китерәләр. Аны туганы Мөхәммәтгәрәй хан күпсанлы

Кырым чирүе белән Казанга китереп куйган, һәм аларның зур өлеше Казан ханлыгында хезмәттә калган. Сөембикә ханбикә заманында да кырым чирүе зур роль уйнаган. Әлеге чирү бик зур булып, аның турында халык тапкыр чагыштыру да чыгарган.

Гүәрдин кебек [таза] (1 т., 228 б.). Патшага хезмәт иткән иң шәп солдатлардан сайлап алынган отрядлар исеменнән алынган. Гүәрдиннәр эре-эре гәүдәле, озын буйлы, таза гаскәриләр булганнар. Беренче гүәрдин гаскәрләрен XVII гасырның 90 нчы елларында Петр I оештыра.

Киткән баш киткән (1т., 384 б.). Галим Н. Исәнбәт бу фразеологизмның мәгънәсен чарасыздан баш иргә, күнәргә риза булу дип аңлата. XVII гасыр ахырларыннан Петр I заманнарында татардан никрутларны ала башлыйлар. Патша чыгарган канун буенча һәрбер крестьян җыены карышмыйча үзләреннән никрутлар (рекрутларны)ны

гаскәрдә хезмәт итәргә җибәрергә тиеш булган. Шул никрутлардан Рәсәй гаскәре тупланган. 1874 нче елда никрутка бару йөкләмәсе хәрби хезмәткә (воинская повинность) алыштырыла. Ягъни билгеле бер яше җиткән яшь кешене хәрби хезмәткә ала алганнар. Солдатка барырга теләмәгән яшьләр төрле юллар белән хезмәттән качарга теләгәннәр. Күбесе *казакъ* булып киткәннәр (1 т., 336 б.), алар казакъ халкы арасына күчеп киткәннәр, үзләре дә вакыт үтү белән казакъка әйләнеп калганнар.

Хәрби эш белән бәйле башка, күчерелмә мәгънәне аңлаткан фразеологизмнар да бар.

Дарысы җитмәде (1 т., 230 б.) фразеологизмы бәхәс-көрәш тотуда гыйльми багажы җитмәүне, позициясен бирүче, югалтучы кешегә әйтелә. XVI йөздән алып, сугышларда дарылы туп-мылтык-лар кулланыла башлый. Әлеге әйтелмә шул заманнарда туган. Дары кулланылышка кергәнгә кадәр гаскәрләр җәя кулланганга, көчен, позицияне югалтуны «җәясе булмады» дип атап йөрткәннәр.

Гаскәрдәге солдатларны командаларны үтәргә тиз өйрәнсен өчен төрле ысуллар кулланылган. Патша заманында солдат уң белән сулны аерсын өчен, уң кулбаш погонына печән, сул кулбаш погонына салам кыстырганнар. Шуннан *«печән дигәч – уңга, салам дигәч – сулга»* дигән фразеологизм туган[1].

Мәкәрҗә ярминкәсе XVI гасыр урталарыннан 1816 нчы елга кадәр эшләгән. Рус һәм татар эшмәкәрләреннән кала Мәкәрҗәдә Урта Азия, Кавказ, Һиндстан, Иран сәүдәгәрләре мех, тире, металл, тукымалар һ.б. белән сату-алу иткәннәр.

Базарларда сату-алуны оештыру, аны тиз һәм табышлы итеп алып барыр өчен, эшкуарларга «унике тел белгән» тылмачлар кирәк булган. Безнең Урта Идел болгарларының Ага базарында, Казан, Мәкәрҗә ярминкәләрендә Азия һәм Европа сәүдәгәрләре белән бик күп аралашырга туры килгәнгә, «унике тел белү» бер дә артык булмаган.

Тамга җыю (2 т., 108 б.). Элек сайлауларда бер кешене авыл халкы исәбенә кертү вакытында аны (тамга) тавыш бирү нәтиҗәсендә ачыклаганнар. Исем, фамилияләрен яза белмәгән кешеләр нәсел тамгаларын куйганнар. Фразеологизмның төбендә төрки халыклардагы һәр нәселнең, гаиләнең тамгасы булуы ята. Борынгы заманда «тамганы» эш кәгазьләренә ханнар, бәкләр куйган. «Тамга термины беренче тапкыр борынгы төрки, уйгур язмаларында очрый. XIII гасырдан «тамга» сүзе Алтын Урданың бар өлкәләрендә дә таралган була. Тамга эчендә хан исеме, дәүләт башында торган ыруның гербы ясалган булган. Гадәттә тамгалар алтын, алсу, зәңгәр төсләр белән куелганнар. Тамга хәзерге замандагы печать, мөһернең беренче формасы булган.

Милләт мәҗлесе (2 т., 19 б.) дип милли парламентны атап йөрткәннәр. 1917 нче елгы Февраль революциясеннән соң, Керенский заманында Уфада айлар буенча барган милли идарәнең бөтен Россия төрле груһ татар депутатлары мәҗлесенә шулай әйткәннәр.

Тарта торгач он булыр (2 т., 212 б.) дип хаксыз репрессияләргә зарланучыларны юату өчен XX гасырның 30 нчы елларында әйтелгән [1:212]. 1919 нчы еллардагы коллектив лаштыру барышында репрессияләр массакүләм төс алган. 1929-1938 нче еллар репрессияләре халыкның барлык социаль классларына кагылган. Сәяси нигездә Татарстанның күп кенә политиклары, фән һәм мәдәният эшлеклеләре иза чиккәннәр. Репрессияләр татар элитасының 150-200 мең кешесенең язмышын изеп, таптап узган. 1930 нчы елларда ОГПУ-НКВД органнары берничә дистә криминал эш ача, мәсәлән: «Антисовет милли оешма» эше, «Атласов» эше, «Солтангалиев» эше, «Контрреволюцион милли баш күтәрүчеләр» эше һ.б. Күргәнебезчә, дәүләтчелеген югалткан халыкның язмышы 1552 нче елдан башлап күбесенчә канлы, бәхетсез көннәр белән тулы булган.

Фразеологияне тарихи планда өйрәнү белем киңлеген, халык, дөнья, дин тарихы, төрле милләтләр менталитетын тирән аңлауны сорый. Әгәр татар фразеологиясе бу һәм башка факторларны исәпкә алынып тикшерелсә, ул, һичшиксез татар теленең, милли психологиясен, глобаль тарихи вакыйгаларның асылына төшенергә ярдәм итәр иде. Киләчәктә компьетер, яңа технологияләр ресурсларын кулланып өйрәнүләрне дәвам итү кирәк.

Әдәбият

1. Исәнбәт Н. Татар теленең фразеологик сүзлеге. Ике томда. – Т.1. – Казан: Таткитнәшр., 1989. – 495 б.

2. Исәнбәт Н. Татар теленең фразеологик сүзлеге. Ике томда. – Т.2. -Казан: Таткитнәшр., 1990. – 365 б.

УДК 81.32

ЛЕКСИЧЕСКИЙ ПОДХОД К ОЦЕНКЕ ДИНАМИКИ УРОВНЯ БЛАГОПОЛУЧИЯ В ОБЩЕСТВЕ

В.Д. Соловьев, В.В. Бочкарев
Казанский федеральный университет, Казань
maki.solovyev@mail.ru, vbochkarev@mail.ru

В статье приводится краткий обзор работ посвященных изучению уровня благополучия с помощью лексических методов. Эти методы основаны на коллекции Google Books Ngram и массовых опросах информантов и отражают восприятие жизни людьми, зафиксированное в миллионах изданных книг.

Ключевые слова: *счастье, благополучие, доходы, эмоции, лексика*

Благополучие – то, к чему стремятся люди во всем мире. Определению факторов, влияющих на благополучие, посвящено множество работ, в частности, исследования Э. Дитона – лауреата Нобелевской премии по экономике 2015 г. [1-3]. Э. Дитон исследовал в первую очередь экономические факторы. Недавно появились интересные новые альтернативные методы изучения благополучия. С созданием в 2011 г. коллекции Google Books Ngram (сокращенно GBN) (https://books.google.com/ngrams) появилась возможность проследить отражение уровня благополучия в сознании людей через тексты миллионов книг. Другим интересным источником данных являются массовые опросы, которые теперь легко проводить с помощью сервиса Amazon's Mechanical Turk (https://www.mturk.com/mturk/welcome). В данной статье приводится обзор ряда ключевых исследований благополучия с помощью этих средств.

Наиболее прямо уровень благополучия (или, наоборот, нищеты) с его отражением в литературе изучался в работе [4]. В статье вводятся понятия индекса экономической нищеты и литературной нищеты. Первый вычисляется по государственным статистическим данным на основе инфляции и безработицы. Индекс литературной нищеты вычисляется по GBN как разность встречаемости синонимов слов joy и синонимов слова sadness. Оказалось, что эти два параметра коррелируют друг с другом.

В работе [5] анализируется большое число определений счастья в словарях разных веков, упоминания счастья в речах президентов Америки, других исторических письменных источниках. В итоге делается несколько выводов. Существуют большие культурные вариации

концепции счастья. Основное различие – между счастьем как внутренним состоянием и счастьем как следствием внешних факторов – удачи. В статье анализируются причины смещения в США ориентации с понятия “счастье народа” на “счастье личности”, происшедшем по данным GBN около 1920 г. Это экономический подъем, развитие масс-медиа, массовая доступность автомобилей.

Кросс-культурное сравнение корреляций восприятия счастья и удачи, также счастья и уровня дохода на материале представленных в GBN немецком, французском, итальянском, испанском, английском и русском языках проведено в [6]. Естественная корреляция счастья с уровнем доходов обнаружена для всех языков, кроме русского. Для русского языка в 20 веке частота употребления слова *счастье* и разность частот слов *богатство* и *нищета* остаются примерно на одном уровне, как и для слова *удача*, при некотором нарастании трудностей. См. рис. 1. Для 19 века же имеет место антикорреляция: частота употребления слова *счастье* растет, в то время как разность частот слов *богатство* и *нищета* падает.

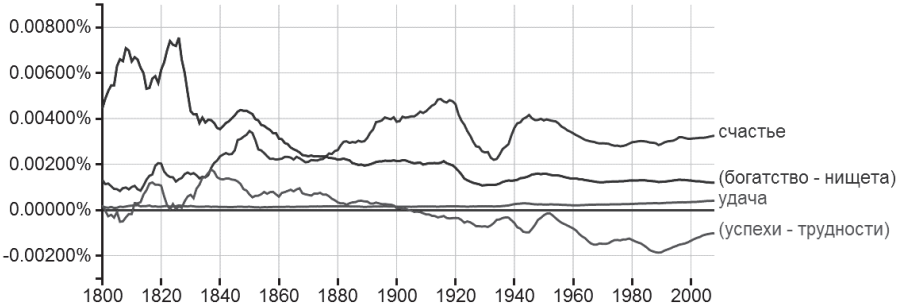


Рис.1. Частота слова *счастье* и сопоставляемых лексем в русском языке

Динамика общего ощущения благополучия на материале тех же 6 языков изучалась в [7]. В этой работе для каждой из базовых эмоций: ‘удовольствие’, ‘страх’, ‘гнев’, ‘печаль’, ‘отвращение’ с помощью словарей для каждого из языков были найдены синонимы этих слов, затем найдены их суммарные частоты и вычислена величина (удовольствие) – (страх+гнев+печаль+отвращение). Следуя работе [4], считаем, что эта величина отражает общее удовлетворение жизнью, и назовем ее индексом удовлетворенностью жизнью. Результаты представлены на рис. 2.

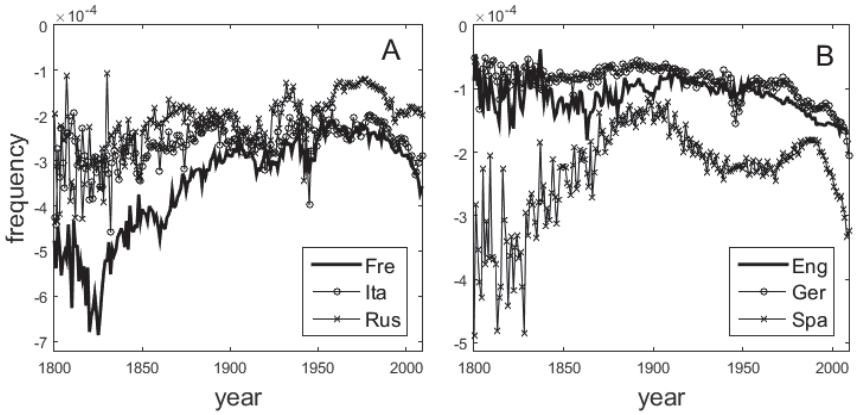


Рис. 2. Индекс удовлетворенностью жизнью

Графики оказались существенно различны для разных языков, причем эти 6 языков можно разделить на две группы со схожей картиной: русский, французский, итальянский и английский, немецкий, испанский. В первой из групп наблюдается тенденция к повышению уровня удовлетворенности, во второй – к снижению. Испанский язык демонстрирует смешанную тенденция: в 19 веке наблюдался рост удовлетворенностью жизнью, как для языков первой группы, но в 20 веке имеет место общая тенденция к снижению, как для языков второй группы. Объяснение обнаруженных феноменов выходит за рамки принятых методологий исследований.

Поскольку оценка благополучия дается через язык, то естественным представляется постановка вопроса – является ли сам язык в целом позитивно или негативно ориентированным. В работе [8] приведены аргументы, что английский язык является позитивным, в том смысле, что число слов, имеющих позитивную оценку, превосходит число негативных слов. Результат получен на основе массовых опросов носителей языка посредством сервиса Amazon’s Mechanical Turk. Оценивалось 10 000 наиболее частотных слов. В [9] этот результат перенесен на 10 языков и трактуется как языковая универсалия. Однако в [10] эти результаты поставлены под сомнение, обращено внимание на недостаточно обоснованную методологию исследований, примененную в [8, 9], и необходимость продолжения исследований, поскольку эти результаты могут иметь ключевую роль для дальнейших исследований эмоций и уровня счастья.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (гранты № 15-29-01173, 15-06-07402).

Литература

1. Kahneman D., Deaton A. High income improves evaluation of life but not emotional well-being // Proc Natl Acad Sci USA. 2010. Vol. 107, №. 38. P. 16489–16493

2. Aghion Ph., Akcigit U., Deaton A., Roulet A. Creative Destruction and Subjective Well-Being // NBER Working Paper №. 21069. 2015. http://www.nber.org/papers/w21069

3. Steptoe A., Deaton A., Stone A. Subjective wellbeing, health, and ageing // Lancet. 2015. Vol. 385. P. 640–648.

4. Bentley R.A., Acerbi A., Ormerod P., Lampos V. Books Average Previous Decade of Economic Misery // PLoS ONE. 2014. 9(1): e83147.

5. Oishi Sh., Graham J., Kesebir S., Galinha I.C. Concepts of Happiness Across Time and Cultures // Pers Soc Psychol Bull. 2013. 39: 559. DOI: 10.1177/0146167213480042.

6. Solovyev V., Batyrshin I. What does happiness depend on? Quantitative comparative analysis of various cultures // Advances in Social and Behavioral Sciences. 2015. Vol. 10. P. 3-9.

7. Solovev V. D., Bochkarev V. V., Bayrasheva V. R. Dynamics of emotions in European languages // Proceedings 7th international conference on cognitive science. Svetlogorsk: Association of Cognitive science. 2016.

8. Kloumann I.M., Danforth C.M., Harris K.D., Bliss C.A., Dodds P.S. Positivity of the English Language // PLoS ONE. 2012. 7(1): e29484.

9. Dodds P. S., Clark E. M., Desu S., Frank M. R., Reagan A. J., Williams J. R., … Danforth C. M. Human language reveals a universal positivity bias // Proc Natl Acad Sci USA. 2015. Vol. 112, № 8. P. 2389–2394.

10. D. Garcia, A. Garas, F. Schweitzer. The language-dependent relationship between word happiness and frequency // Proc Natl Acad Sci USA. 2015. Vol. 112, № 23. E2983.

ПОКА, КАК И ЭКСПЛЕТИВНОЕ ОТРИЦАНИЕ

С.Г. Татевосов
Московский государственный университет имени М.В. Ломоносова, Москва
tatevosov@gmail.com

В статье предлагаются наброски к анализу союза *пока*, которые показывают, что отрицание, часто описываемое в литературе как эксплетивное, в действительности интерпретируется строго композиционально. Важный аргумент в пользу такого анализа дают временные придаточные с союзом *как*, в которых отрицание проявляет сходные семантические свойства.

Ключевые слова: *темпоральная семантика, временные модификаторы, отрицание*

Семантика сентенциальных сирконстантов с союзом *пока* привлекала значительное внимание исследователей ([1], [2], [3]), однако,

как кажется, в понимании того, как она устроена, можно добиться большего. Анализируя *пока*, мы сталкиваемся с тем, что его значение сложным образом взаимодействует с временем и видом обеих предикаций, а главное — с наличием отрицания в зависимой предикации. Сравним:

http://webcache.googleusercontent.com/search?q=cache:tz8X-J_KxMAJ:yaoi-club.flydes.ru/shkola_uke.php%3Fid%3D72%26str%3D20+&cd=50&hl=en&ct=clnk&gl=ru

(1) Организм же не автомат: включил-выключил. Пока он придет в норму, пройдет несколько часов [sovet.kidstaff.com.ua].

(2) Давайте я его в комнату отведу. И побуду рядом, пока он не придет в норму [yaoi-club.flydes.ru].

Благодаря примерам такого рода мы видим поразительное свойство предложений с *пока* и *пока не*: отрицание как будто не оказывает обычного влияния на их условия истинности. В (1) сообщается об интервале в несколько часов, где правой границей выступает момент, когда ребенок придет в норму. Аналогично в (2) мы говорим об интервале (в течение которого субъект главного предложения («я») проведет время в месте с объектом («с ним»)), который завершается в тот момент, когда объект приходит в норму. Семантический вклад отрицания в (2) не просматривается. Некоторые исследователи, в частности, Л.Н. Иорданская и И.А. Мельчук [2], говорят, что отрицания в таких примерах имеет эксплетивный характер и не интерпретируется.

Цель этих заметок — предложить набросок альтернативной теории, исходящей из того, что значение *пока*-предложений с отрицанием и без полностью композиционально и что отрицание имеет в них регулярную интерпретацию.

В качестве первого шага полезно сопоставить (квази-)синонимичные предложения с *пока* и *пока не* с другим случаем «эксплетивного» отрицания, иллюстрируемым в (3)-(4).

(3) Завтра будет двадцать три года, как я виделся с Лениным.
(4) Завтра будет двадцать три года, как я не виделся с Лениным.

Предложения в (3)-(4) различаются на отрицание, однако описывают ровно одно и то же положение вещей: ближайшая к моменту речи встреча говорящего с председателем Совнаркома находится в прошлом на расстоянии 23 года. (Предложения имеют, конечно и другие различия. (3), в частности, содержит пресуппозицию уникальности, а (4) —

импликатуру множественности, которые мы обсудим в более полной версии этой статьи.)

Как возможно, что (3) и (4) имеют одинаковые условия истинности? Мы предлагаем следующий ответ на этот вопрос. (3)-(4) сообщают об одном и том же интервале, но **описывают** его разными способами. (4) подает его как интервал длительностью двадцать три года с правой границей ‘завтра’, в течение которого **не было ни одного события** вида ‘говорящий встречался с Лениным’. (3) говорит о минимальном интервале, отделяющем завтра от **ближайшего события** в прошлом вида ‘говорящий встречается с Лениным’. Эти описания экстенсионально эквивалентны: они задают один и тот же интервал. В более формальных терминах:

(5) $\lambda t. \mid t \mid = 23 \text{ года} \wedge RB(t) = \text{завтра} \wedge \forall t' \exists e [\text{встреча}(\text{Ленин})(\Gamma)(e) \wedge t' = \tau(e) \rightarrow \neg\, t' \subseteq t]$

(6) $\lambda t. \mid t \mid = 23 \text{ года} \wedge RB(t) = \text{завтра} \wedge \forall t' [\, t' \subseteq t \rightarrow \neg\exists\, e [\text{встреча}(\text{Ленин})(\Gamma)(e) \wedge t' = \tau(e)]$

(5) — предикат над временными интервалами. Он содержит в своем экстенсионале интервалы длительностью двадцать три года, правая граница которых совпадает с ‘завтра’. Любой интервал, на котором происходит встреча говорящего с Лениным, находится за пределами этого интервала. Это то прочтение, которое соответствует интуитивном пониманию ‘ближайшая встреча — на расстоянии 23 года’. (6) также обозначает временные интервалы с правой границей завтра. На этот раз о них сообщается, что на любом подынтервале этого интервала отсутствует событие встречи говорящего с Лениным. Это прочтение ‘нет ни одной встречи в течение 23 лет’. Первый предикат тем самым соответствует предложению с утвердительным зависимым в (3), а второй — с отрицательным зависимым в (4).

Формулы в (5)-(6) эквивалентны. Если присмотреться к ним внимательнее, можно заметить, что они представляют собой частный случай логической эквивалентности в (7):

(7) $P \rightarrow \neg Q \equiv \neg P \vee \neg Q \equiv Q \rightarrow \neg P$

В (5)-(6) переменной P соответствует формула $\exists e [\text{встреча}(\text{Ленин})(\Gamma)(e) \wedge t' = \tau(e)]$, а переменной Q — формула $t' \subseteq t$. (7) объясняет видимую синонимичность предложений типа (3)-(4) принципиальным образом.

Логическая структура этих предложений в действительности сводится к общей формуле с дизъюнкцией:

(8) $\lambda t. \mid t \mid = 23$ года $\wedge$ RB(t) = завтра $\wedge \forall t' [\neg\exists e$ [встреча(Ленин)(Г)(e) $\wedge$ $t' = \tau(e)] \vee \neg\, t' \subseteq t]]$

Аналогичная линия рассуждений выстраивается и в случае с *пока*. Тот класс употреблений, который представлен примерами типа (1)-(2) можно анализировать следующим образом. *Пока*-клаузы обозначают временной интервал *t*, который модифицирует время ассерции главной клаузы. А именно, они сообщают, что время ассерции главной клаузы находится внутри *t* и имеет общую с ним правую границу. Информации о длительности интервала клаузы с *пока*, в отличие от предложений (3)-(4) с *как*, не передают. Описание *t* осуществляется теми же двумя способами, что и в предыдущем случае. Первый способ, который мы видим в (2), — это говорить о *t* как об интервале, на котором отсутствуют события вида ‘они пришел в норму’, а второй, (1), — как об интервале, отделяющем нас от ближайшего такого события.

Логическая форма предложений в (1)-(2) показана в (9)-(10):

(9) $\lambda t. \exists e$ пройдет(неск.часов)(e) $\wedge \exists t' [t \subseteq t' \wedge$ RB(t)=RB(t′) $\wedge \mid t' \mid = h \wedge$
$\forall t'' \exists e' [$придет.в.норму(он)(e′) $\wedge\ t'' = \tau(e) \rightarrow \neg\, t'' \subseteq t']$

(10) $\lambda t. \exists e$ побыть.рядом(Г)(e) $\wedge \exists t' [t \subseteq t' \wedge$ RB(t)=RB(t′) $\wedge \mid t \mid = h \wedge$
$\forall t'' [t'' \subseteq t' \rightarrow \neg\exists e' [$придет.в.норму(он)(e′) $\wedge\ t'' = \tau(e')]$

(9) обозначает временные интервалы, в течение которых пройдет несколько часов. Их правая граница — это одновременно правая граница любого интервала такого, что события прихода в норму находятся за его пределами. (10) вводит в рассмотрение интервалы, в течение которых Г побудет рядом с субъектом. Правая граница этих интервалов совпадает с правой границей интервалов, на которых отсутствуют события прихода в норму.

Таким образом и в (5)-(6) и в (9)-(10) семантическая игра строится вокруг эквивалетности описаний ‘нет внутри X’ — ‘есть за пределами X’, где X — это интервалы соотнесенные с *как*-клаузами в (5)-(6) и с *пока*-клаузами в (9)-(10). Появление в них отрицания означает, что мы описываем интервал через отсутствие соответствующих событий внутри него. Если отрицания нет, мы толкуем о наличии таких событий исключительно за его пределами. Отрицание, таким образом, вносит

в интерпретацию свой обычный семантический вклад. Близкую линию рассуждений предлагает и Е.В. Падучева [3]: с ее точки зрения, отрицание интерпретируемо, а его использование обусловлено тем, что под отрицанием создаются гомогенные временные дескрипции.

Разумеется, изложенное выше — это лишь предварительные наброски теории, объясняющей дистрибуцию отрицания в *пока*- клаузах и в *как*-клаузах. Они не учитывают, что тот и другой случай связаны с различающимися пресуппозициями и импликатурами, которые более полная теория должна, разумеется, описать и объяснить. Отдельный вопрос, — возможно ли с помощью логической формы типа (9)-(10) (возможно, с некоторыми минимальными модификациями) описать все употребления придаточных с *пока*, например, такие, как (11).

(11) Пока Иван работал, Маша читала.

Не пытаясь решить этот вопрос здесь, отметим еще раз то, что кажется нам главным результатом этого очерка. Семантическая теория должна относиться серьезно к имеющемуся в ее распоряжении морфосинтаксическому материалу. Теория, которая, не найдя грамматическом элементу места в семантической деривации, объявляет его эксплетивным, существенно проигрывает альтернативам не только в концептуальной привлекательности, но и в эмпирической адекватности.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проект № 14-06-00435).

Литература

[1] Барентсен А. Проблемы описания союза пока // Nomachi M., Danylenko A., Piper P. (eds). Grammaticalization and lexicalization in the Slavic languages. Proceedings from the 36th Meeting of the commission on the grammatical structure of the Slavic languages of the International committee of Slavists. München: Otto Sagner Verlag, 2014.

[2] Iordanskaja L., Mel'čuk I. Semantics of the Russian conjunction POKA 'while, before, until' // Von grammatischen Kategorien und sprachlichen Weltbildern — Die Slavia von der Sprachgeschichte bis zur Politsprache. Festschrift für Daniel Weiss zum 60 Geburtstag. Berger T. (Hrsg.). Wien: Verlag Otto Sagner, 2009.

[3] Падучева Е.В. Акциональная классификация глаголов и семантика союза пока // ВЯ. 2015. №5.

УДК 81’33

ТАТАР ТЕЛЕ КОРПУСЫНДА ГРАММАТИК ОМОНИМИЯНЕ ЧИШҮНЕҢ КАЙБЕР НӘТИҖӘЛӘРЕ

А.Р. Фазлыева
Б.Э. Хәкимов
Казан федераль университеты, Казан
ТР ФА “Гамәли семиотика” фәнни-тикшеренү институты
khakeem@yandex.ru, afalina.year@gmail.com

В статье представлены некоторые результаты исследования избран-ных типов грамматических и функциональных омонимов в татарском языке на основе контекстных правил. Рассматриваются контекстные характеристики частотных функциональных омонимов и омоморфем. Результаты исследования могут быть использованы для разработки формальных правил автоматического разрешения омонимии в корпусах татарского языка.

Ключевые слова: *омонимия, многозначность, снятие многозначности, корпус текстов, грамматический омоним, татарский язык*

Татар лингвистикасында соңгы елларда телнең функциональ, территориаль һәм башка вариантларын чагылдырган корпуслар – нәзари һәм гамәли тикшеренүләр өчен хезмәт итүче махсус информацион-эзләү системалары үсеш кичерә. Корпус лингвистикасы өчен омонимияне һәм күпмәгънәлелекне чишү проблемасы бик актуаль.

Әлеге тикшеренүдә без «Туган тел» татар гомумтел корпусының тамгалау (разметка) системасы нигезендә татар телендәге кайбер грамматик омоним типларының контекст чолганышын өйрәндек.

Омонимия – бер үк авазлар комплексының төрле төшенчәләрне белдерүе [1: 9]. Аңа лексик берәмлекләр, фонетик структуралары охшаш, мәгънә ягыннан үзгә грамматик формалар керә. Омонимия күренешенә кагылышлы фәнни хезмәтләрдә омонимнарның ясалыш үзенчәлекләре, аларны классификацияләү принциплары карала. Мисал өчен, лексик-семантик һәм грамматик мәгънәләре ягыннан омонимнарның 3 төркемен күрсәтергә була:

- лексик (сүзләр бер сүз төркеменә карый),
- лексик-грамматик (төрле сүз төркеменә карыйлар) яки функциональ,
- грамматик (сүздә грамматик формаларның омонимлашуы) типларга бүлү.

Грамматик омонимнарның 2 төре аерыла: 1) омоморфема – аваздаш, әмма мәгънәләре төрле морфемалар - *-чык: җыерчык (сүз ясый) – кызчык (стилистик кушымча);* 2) омомодель – төзелеше, формасы бертөрле, грамматик мәгънәсе төрлечә булган җөмлә яки җөмлә кисәге: *мамык шәл – мамык үсә* [2: 20]. Аффикс күпмәгънәле булып, бер үк сүзформа төрле грамматик мәгънәләргә ия була ала. Бу грамматик омоним.

Лингвистикада "функциональ омонимия" төшенчәсе дә билгеле. Ул яңгырашлары бер үк, этимологик яктан кардәш, ләкин төрле сүз төркемнәренә караган сүзләрне белдерә [3: 14]. Бу күренеш «Туган тел» корпусында еш очрый. Ул мөстәкыйль һәм ярдәмче сүз төркемнәренә хас. Татар тел белемендә бу темага аерым хезмәтләр багышланган [1, 4, 5, 6].

Соңгы елларда, корпус лингвистикасы үсеше белән бәйле рәвештә, омонимияне автоматик ысуллар белән чишү мәсьәләсенә дә игътибар бирелә башлады. Шул рәвешле, [9] хезмәтендә:

1) грамматик омонимияне һәм күпмәгънәлелекне чишүне автоматизацияләү мөмкинлекләрен тикшерү мәсъәләләре карала;

2) морфологик анализ (автомат рәвештә) барышында альтернатив тикшерү релевантлыгы билгеләнә;

3) еш очраган омоформаларны классификацияләү һәм омонимияне чишү методлары тәкъдим ителә;

4) классификация һәм контекст кагыйдәләре нигезендә эшләнә торган лингвистик чыганаклар һәм программа модульләре татар теле корпусында күпмәгънәлелекне чишәргә мөмкинлек бирәчәк дигән фикер әйтелә.

Хәзерге вакытта омонимияне автомат рәвештә чишү эшләрендә контекст методы, статистик һәм гибридлы методлар кулланыла. Омонимнарның һәр тибына тәфсилле лингвистик анализ үткәрүне таләп итсә дә, татар теле өчен кулай вариант – контекст кагыйдәләренә нигезләнгән метод, чөнки:

1) татар теленең статистик һәм гибрид модельләрне кулланырга мөмкинлек биргән һәм тиешле дәрәҗәдәге күләме булган электрон корпусы юк;

2) грамматиканың даимилеге һәм телнең кагыйдәләргә буйсынуы төгәл контекст чикләүләрен табарга һәм тасвирларга ярдәм итә.

«Туган тел» татар милли корпусы ТР Фәннәр академиясенең «Гамәли семиотика» фәнни-тикшеренү институты белән КФУ галимнәре тарафыннан эшләнгән. Корпус татар әдәби теле текстлары тупланмасыннан тора. Биредә морфоанализатор нигезендә автомат рәвештә эшли торган морфологик тамгалау системасы кулланыла. Системаның максаты – барлык грамматик формаларны күрсәтү.

Корпуста грамматик омонимия чишелмәгән һәм, тамгалау барышында омонимик сүзләр булган очракта, альтернатив тикшерү юлы кулланыла. Бу проблема проектта морфологик күптөрлелекне чишү моделен файдалану нәтиҗәсендә хәл ителә. Текстны автоматик эшкәртү системаларында «омоним» һәм «күпмәгънәлелек» төшенчәләре арасында чик югала. Алар еш кына икече бергә омоним дип атала. Корпуста, статистик корпус мәгълүматларын кулланып, грамматик омонимнарның төрле типларын контекст кысаларында тикшерү эше бара, аларны автомат рәвештә чишү методлары тәкъдим ителә.

Әлеге эшебезнең беренче өлеше берничә функциональ омоним тибын тикшерүгә багышланган. Шул максатка ирешү өчен түбәндәге эшләр башкарылды:

1) Корпуста берничә омоним сүз кергән җөмләләр тикшерелде *(соң, без, ит, ук, бит)*.

2) Аларның морфологик характеристикасы билгеләнде (**POST** – бәйлек, **PART** – кисәкчә, **Adv** – рәвеш, **Adj** – сыйфат, **N** – исем, **V** – фигыль, **PN** – алмашлык).

3) Сүзнең типик варианты, үзенчәлекләре ачыкланды.

Таблица 1

Соң сүзе

Сүз төркеме	**Контекст үзенчәлекләре**	**Мисаллар**
бәйлек	1) гадәттә сорау җөмләләрдә очрый	
	а) күбесенчә хәбәрдән соң килә	Нишләргә **соң** , егетләр ?
	б) сорау һәм зат алмашлыкларыннан соң килә ала	— Син үзең **соң** күпме өмет иткән идең ?
	2) тойгылы җөмләләрдә очрый:	
	а) җөмлә башында килә ала	**Соң** инде белгән өстенә сорамасаң .
	ә) сүз җөмлә була ала	— **Соң** !
	б) хәбәр + да/дә/та/тә кисәкчәләреннән соң килә ала	— Һай , хәйләкәр дә **соң** үзең!
бәйлек/ рәвеш	Чыгыш килешендәге исем/алмашлык/ фигыль (хикәя, сыйфат, исем) сорый	Байтак барганнан **соң** , Хафиз : – Кайтыйк , Галим , вакыт соң инде , – диде .
бәйлек/ рәвеш	үзеннән соң кисәкчә (да/дә, гына/генә) ияртә ала	хәл булганнан **соң** дамы?
рәвеш	1) Чыгыш килешендәге сүздән ераклаша ала	1) Аннан ул бик **соң** гына кайтты.

	2) аның алдыннан килүче сүз фигыльнең хикәя фигыль төркемчәсе булып килми	
	3) аны җөмләдән алып куеп карап була	Әнием кичтән дә **соң** ятып , бүген дә юл килеп арганга күрә, ... йокларга ятты. – Әнием кичтән дә ятып, бүген дә юл килеп арганга күрә, ... йокларга ятты.
	4) гадәттә, җөмлә ахырында хәбәр алдыннан яисә хәбәр составында килә	Минем хәтта , **соң** булса да , нидер эшлисем , ниндидер чара да күрәсем килеп киткән иде .
сыйфат	соң ~ соңгы	Картлар гына түгел , **соң** заманда яшьләр дә килеп карый башлады.

Таблица 2

Без сүзе

Сүз төркеме	**Контекст үзенчәлекләре**	**Мисаллар**
исем	кайбер бәйлекләр (өчен, кебек, белән һ.б.) белән бергә килгәндә башлангыч формаларын үзгәртми (аларга кушымча өстәлми)	
алмаш-лык	киресенчә	— **Без** , әбекәем , без !

Таблица 3

Ит сүзе

Сүз төркеме	**Контекст үзенчәлекләре**	**Мисаллар**
фигыль	гадәттә, җөмлә ахырында исем, сирәгрәк башка сүз төркемнәре белән бер мәгънә белдереп килә	Үзең киңәш **ит** инде, җаным .
исем	җөмләдә мөстәкыйль рәвештә урнаша	... хезмәт көненә **ит** , сөт биреп баруны ферма алдына бурыч итеп куясы килә .
	гадәттә хәбәрдән алда килә	Иртән балалар уянуга өстәлдә парланып **ит** тора .

Тикшеренүнең икенче өлешендә без *-лык/-лек* күпмәгънәле һәм күпфункцияле аффиксының контекстта кулланылу үзенчәлекләренә игътибар иттек. “Туган тел” корпусында бу аффикс ике тэг (шартлы тамга) белән билгеләнергә мөмкин: NMLZ (‘исемләшү’ мәгънәсендә) һәм PSBL (‘ихтималлылык’, ‘мөмкинлек’). Тикшерү барышында:

1) *-лык/-лек* күпмәгънәле кушымчалары кергән җөмләләр алынды (Excel форматында).

2) Аларның морфологик характеристикасы билгеләнде.

3) Сүзләрнең үзенчәлекләре ачыкланды, алар классларга бүленде. Анализ нәтиҗәләре 4 нче таблицада күрсәтелгән.

Таблица 4. *-лык/-лек* кушымчасының контекстуаль мәгънәләре.

№	Мәгънә	Контекст үзенчәлекләре	Мисаллар
11.	сүзъясагыч кушымча, абстракт һәм гомумиләштерү мәгънәләрендә килә торган исем ясала.	1) исем (N), сыйфат (Adj), сирәгрәк – фигыль (V) сүз төркеменнән ясала	Мәсәлән, нигә әле мәхәббәтне мәхәббәт яки җаныңның матурлыгы белән яуламаска?
		2) аңа исемгә хас булган барлык кушымчалар ялгана ала (килеш, тартым, сорауны белдерүче (INT, INT_MIR), ихтималлыкны белдерүче кушымчалар (PROB) һ.б.)	Мине ниндиләр исемлегенә кертергә уйлыйсыз инде?
		3) 2, 3, 4 классларга карамаса, ул 1 класска кертелә	
		4) сыйфат (ADJ, ATTR_MUN, ATTR_ABES) + бер лексемасы (NUM)+ …+.-лЫК формуласына туры килә	
22.	“... өчен җитәрлек” мәгънәсен белдергән атрибутив мәгънә	1) 1. –лЫК кушымчасы ялганган сүздән сул якта микъдар саны килә (NUM)	Аның суфичылык фәлсәфәсе рухында язылган берничә шигырен, мәдхия, мөнәҗәтләрен, дини-әхлакый эчтәлекле тәрҗемә хикәяләрен, риваятьләрен һәм юлъязмаларын эченә алган “Тәрҗемәй хаҗи Әбелмәних Әл-бистәви Әл-сәгыйди” исемле 51 битлек мәҗмугасы 1845 елны Казанда басылып чыга.
		2) –лЫК кушымчасы ялганган формалар темпораль (вакыт берәмлекләре) мәгънәле исемнәрдән ясалган.	Аның ата-анасы Гәрәй мирзаның атасы Һибәтулла мирзада гомерлек ялчы булып торган иде.
		3) *яшьлек* сүзе 2 класска керсен өчен, аның алдыннан микъдар саны килергә тиеш. Башка очракта аны 1 класска кертеп карау кирәк.	Шул унтугыз яшьлек тәүфыйклы егет уналты яшьлек Көлемсәргә гашыйк.
		4) *гомерлек* сүзе һәрвакыт 2 класска керә.	Аның ата-анасы Гәрәй мирзаның атасы Һибәтулла мирзада гомерлек ялчы булып торган иде.

33.	Эшне башкарырга мөмкинлек булу (мөмкинлек булмау) мәгънәсен белдерә торган фигыль формасы.	-лЫК кушымчасы ялганган сүз аффиксаль формулаларда күренә	Анларда язмышның бу агымына каршы барырлык куәт юк.
44.	-Дан аблятив (чыгыш килеше) кушымчасы ялганган иярчен сәбәп кисәкләр булган очракларда синтаксик номинальләштерү.	Аффиксаль чылбыр түбәндәгечә языла: фигыль нигезе (V) + үткән заман сыйфат фигыль кушымчасы (PCP_PS -ган, -гән, -кан, -кән алломорфлары) +лЫК +чыгыш килеше кушымчасы (ABL, -тан/ -тән алломорфлары)	Берәүләр, авыру булганлыктан, имтиханда дүртле алганнар, бишлесез – ышаныч кечкенә.

Тикшерелгән омонимнарның морфологик үзенчәлекләре ачыкланды. Алар нигезендә “Туган тел” корпусында омонимияне чишү өчен формаль кагыйдәләр төзеләчәк. Әлеге кагыйдәләрнең дөреслегенә аларны тикшерү эшен башкарганнан соң гына ышанып булачак.

Омонимияне контекстлар ярдәмендә абсолют төгәллек белән чишеп булмый. Бу әлеге күренешнең лексикографик чыганакларда каршылыклы төс алуы һәм грамматик омонимиянең катлаулы очракларын бары катлаулы синтаксик анализ методлары ярдәмендә генә чишү мөмкинлегенә бәйле. Шулай да, семантик контестларны санау методын кулланып, чагыштырмача төгәл нәтиҗәләргә ирешергә мөмкин. Шуның өчен омонимнарның тулы классификациясен булдыру һәм төрле типтагы омонимнарның грамматик үзенчәлекләрен төгәлләштерү кирәк. Шулай итеп, киләчәктә без, табылган үзенчәлекләрне формаль рәвешкә китереп, махсус алгоритмик кагыйдәләр төзергә, омонимнарның калган очракларын контекст эчендә тикшерергә һәм аларның морфологик үзенчәлекләрен билгеләргә тиешбез.

Әдәбият

1. Салахова Р.Р. Омонимичные суффиксы татарского языка: Дис. ...канд. филол. наук. Казань, 2004, С.9.

2. Галиуллина Г.Р. Татар теле. Лексикология: таблицалар, схемалар, анализ үрнәкләре, күнегүләр, сүзлекчә: Югары сыйныф укучылары һәм студентлар өчен. – Казан: Мәгариф, 2007. – 20 б.

3. Бабайцева В.В. Переходные конструкции в синтаксисе. – Воронеж, 1967. – С. 14.

4. Курбатов Х.Р. Грамматические омонимы в татарском языке // Татар теле һәм әдәбияты. – Казан: Татар кит. нәшр., 1959. – С.307-311.

5. Сәлимгәрәева Б. С. Хәзерге татар телендә омонимнар.—Уфа,1982.

6. Ахтямов М.Х. Проблемы омонимии в современном башкирском литературном языке: Автореф. дис. канд. филол. наук. – М., 1966. – С. 20.

7. Сулейманов Д.Ш., Хакимов Б.Э., Гильмуллин Р.А. Корпус татарского языка: концептуальные и лингвистические аспекты // Вестник ТГГПУ. – 2011. – № 4 (26). – С. 211-216.

8. Невзорова О.А., Салимов Ф.И., Хакимов Б.Э., Гатиатуллин А.Р., Гильмуллин Р.А. Галиева А.М., Якубова Д.Д. Аюпов М.М. Семантико-грамматическая аннотация в русско-татарской лексикографической базе данных // Филологические науки. Вопросы теории и практики. – Тамбов: Грамота, 2012. №7 (18): в 2-х ч. Ч.1. – С.141-146.

9. Хакимов Б.Э. Гильмуллин Р.А., Гатауллин Р.Р. Разрешение грамматической многозначности в корпусе татарского языка // Ученые записки Казанского университета. Сер. Гуманит. Науки. – 2014. – Т. 156, кн. 5. – С. 236-244.

10. Интернет-ресурс: http://tatcorp.antat.ru/about

11. Интернет-ресурс: http://corpus.antat.ru

12. Зиляева Р.А. Абдрахманова Г.Г. Татар теленең аңлатмалы сүзлеге. Өч томда. I, II, III том. Казан: Татарстан китап нәшрияты, 1977.

13. Сафиуллина Ф.С., Ризванова Л.М. Татар теленең омонимнар сүзлеге.—К., 1997.

РАЗДЕЛ III | **МОДЕЛИ И ТЕХНОЛОГИИ ДЛЯ ЛИНГВИСТИЧЕСКИХ ПРИЛОЖЕНИЙ**

УДК 811.512.145; 81'366.5

РАЗРАБОТКА КОНТЕКСТНЫХ ПРАВИЛ ДЛЯ РАЗРЕШЕНИЯ МОРФОЛОГИЧЕСКОЙ МНОГОЗНАЧНОСТИ В КОРПУСЕ ТАТАРСКОГО ЯЗЫКА

Р.Р. Гатауллин, Р.А. Гильмуллин
НИИ «Прикладная семиотика» АН РТ
ramil.gata@gmail.com, rinatgilmullin@gmail.com

Данная работа является продолжением работ по подготовке размеченного корпуса татарского языка. Ранее был представлен проект (http://tatcorp.antat.ru) по ручному снятию морфологической многозначности в корпусе татарского языка, основной целью которого является использование краудсорсингового подхода для накопления данных. Следующим шагом стала разработка инструментария для создания, тестирования и апробации контекстных правил разрешения морфологической многозначности.

Ключевые слова: *корпус татарского языка, разрешение морфологической многозначности, контекстный метод.*

Татарский язык относится к агглютинативному типу языков, т.е. словоформы в татарском языке образуются последовательным присоединением аффиксов к основе слова. Аффиксы имеют жесткий порядок следовании, но некоторые последовательности могут повторяться, усложняя смысл словоформы и образуя таким образом теоретически бесконечно длинные словоформы (например, «урманнардагылардагылар» - те, что (кто) у тех, кто в лесах) [1]. Но на практике, по статистическим данным корпуса, длина аффиксальной цепочки в среднем не превышает 5-6 аффиксов, а максимальная длина аффиксальной цепочки равна 12. Но и такая ситуация приводит к большому разнообразию типов морфологической многозначности.

Из предварительного анализа корпуса татарского языка выявлено, что почти треть всех словоформ (~31%) корпуса в той или иной мере имеют

более одного разбора [2]. Количество типов морфологической многозначности превышает 10000 для корпусной выборки в 21 млн. словоупотреблений [3], что вместе с агглютинативностью татарского языка теоретически приводит к бесконечному многообразию таких типов. Практически же должна быть возможность свести их к конечным классам, для которых правила разрешения будут одними и те же. На данный момент было выявлено порядка 400 таких классов, но данная гипотеза еще требует подтверждения и дальнейших исследований.

Для автоматического разрешения морфологической многозначности предлагается применять гибридный метод, включающий метод, основанный на контекстных правилах, и статистико-вероятностный метод. Такой выбор обусловливается, тем, что, несмотря на то, что метод контекстных правил показывает достаточно хорошую точность при разрешении, сама по себе разработка таких правил достаточно трудоемкая задача, требующая тщательного лингвистического анализа контекстных ограничений для каждого типа многозначности. Статистико-вероятностные методы и методы машинного обучения хорошо выявляют и успешно используют при работе скрытые закономерности и взаимосвязи в контексте, но минусом их является плохая обучаемость на разреженных данных. Для некоторых типов, действительно, в корпусе имеется мало примеров.

Таким образом, предлагается перед использованием статистико-вероятностного метода, как первоначальный этап, применять контекстный метод. Также кроме случаев с разреженными данными, контекстный метод подходит для случаев, когда многозначность возникает вследствие избыточности словаря основ морфоанализатора, и других исключительных случаев, когда легче прописать контекстные ограничения, чем готовить обучающую выборку по данному типу или случаю [2].

В настоящее время подготовлено веб-приложение для разметки корпуса татарского языка. Основной упор делается на «кроудсорсинговый» аспект, т.е. использование усилий большого количества людей. С помощью данного инструмента, во-первых, будет подготовлен вручную размеченный подкорпус. Во-вторых, будет получено достаточное количество данных для обучения статистико-вероятностного метода. Следующим шагом в развитии этого проекта стал инструментарий разработки контекстных правил для разрешения, использующий ранее полученные данные для тестирования и апробации разрабатываемых правил.

В ранних публикациях [4] описывались идея и архитектура инструментария, представлен прототип, который показал работоспособность. В текущей работе инструментарий был доработан и реализован в виде веб-приложения. В работе [5] подробно описаны основные

достоинства и недостатки метода контекстных правил, приведены конкретные структуры обобщенных правил для разрешения функциональной омонимии некоторых типов.

Обобщенный метод контекстного разрешения функциональной омонимии для татарского языка включает несколько этапов [4]:

1. построение полной классификации типов функциональных омонимов;

2. выделение минимального множества разрешающих контекстов для каждого типа. Минимальность множества означает, что для каждого типа функционального омонима следует оценить сложность распознавания каждой части речи, принадлежащей данному типу. Затем необходимо построить множество разрешающих контекстов (МРК), имеющих минимальную сложность распознавания. В алгоритмической записи данное требование выражается следующим правилом: если для функционального омонима X, имеющего тип T1 или T2, подошло правило из МРК, то тип омонима X определяется примененным правилом, иначе приписывается альтернативный тип;

3. построение управляющей структуры обобщенного правила, обеспечивающего максимальную точность распознавания.

Для решения поставленных задач разработана соответствующая архитектура программного инструментария, включающая следующие базовые объекты и понятия:

- *Омоформа* (или функциональный омоним) – слова, совпадающие в своем звучании лишь в отдельных формах (той же части речи или разных частей речи);
- *База типов омоформ* (или База контекстных правил) – иерархически упорядоченный список типов омоформ; для каждого типа определено множество разрешающих контекстов. На основе этих правил происходит разрешение многозначности для отдельно взятого типа омоформ;
- *Обобщенное правило разрешения* (ОПР) – правило, на основе контекстной информации определяющее актуальный вариант структуры омоформы. Для каждого типа функционального омонима следует оценить сложность распознавания каждой части речи, принадлежащей данному типу;
- *Множество разрешающих контекстов* (МРК) – совокупность минимальных разрешающих контекстов, достаточных для распознавания функционального омонима как определенного варианта структуры омоформы;
- *Управляющая структура* обобщенного правила обеспечивает и контролирует порядок применения правил;

•*Минимальный разрешающий контекст* – неделимое в данном контексте простое условие, имеющее минимальную сложность распознавания.

Процесс распознавания омонимии происходит следующим образом:

1. У анализируемого слова определяется тип функциональной омонимии, и в соответствии с этим типом из Базы контекстных правил находится обобщенное правило разрешения;

2. Управляющая структура задает порядок применения правил;

3. При применении каждого правила, проверяется каждый минимальный контекст разрешении этого правила;

4. Если при проверке правила получили подтверждение о его истинности, то функциональная омонимия распознается в соответствии с этим вариантом структуры омоформы;

5. Иначе, если есть другое правило, осуществляется переход к следующему правилу и выполняется то же самое;

6. И если нет другого правила, то в качестве структуры выбирается тип по умолчанию;

7. Если нет такого типа, то многозначность помечается как неразрешенная.

В некоторых источниках «краудсорсинг» (англ. crowd – толпа, народ; source - ресурс) трактуется как мобилизация ресурсов людей посредством информационных технологий с целью решения задач, стоящих перед бизнесом, государством и обществом в целом [5]. Действительно, для решения некоторых задач такой подход полностью оправдывает себя. Примером служат всемирно-известный ресурс Википедия, программистский ресурс http://stackoverflow.com/ и другие разного рода форумы, где сбором информации и наполнением сайта занимаются обычные пользователи. В сфере NLP можно отметить Открытый корпус русского языка http://opencorpora.org/, где с помощью пользователей происходит разметка корпуса [6]. Как уже отмечалось ранее [3], у нас также имеется опыт в таком роде проекте: с помощью пользователей ресурса http://tatcorp.antat.ru разрешается морфологическую многозначность в корпусе татарского языка.

Основная идея подхода состоит в разбиении задачи на мелкие подзадачи, которые достаточно легко решаются и не сильно затрудняют пользователя. Другой важной частью является мотивация пользователей. Для одних это всевозможные “ачивки” (англ. achievement достижение), для других развитие opensource проектов (англ. open - открытый; source - ресурс).

Применения данного подхода для разработки контекстных правил вполне возможно, но в отличие от случая ручного снятия

морфологической многозначности, где требуется простое знание языка, разработка контекстных правил требует определенных знаний в области языкознания и лингвистики, что сильно ограничивает круг возможных пользователей. Но несмотря на это, есть возможность привлечения для работы студентов-лингвистов и учителей, которые занимаются данной проблематикой.

Для того, чтобы начать разработку контекстных правил необходимо зарегистрироваться на сайте http://tatcorp.antat.ru и перейти на вкладку http://tatcorp.antat.ru/disam/rules/, где представлен список всех омоформ, для которых уже имеются правила разрешения. Имеется возможность как улучшать уже имеющиеся правила, так и создавать новые правила для не имеющихся в списке типов омоформ (см. Рис.1).

Так как в разработке участвует не один пользователь, появляется необходимость разграничения доступа пользователей, а также необходимость своего рода “песочницы”, где происходит разработка и тестирование правил. После этого правило может быть добавлено в основную базу правил.

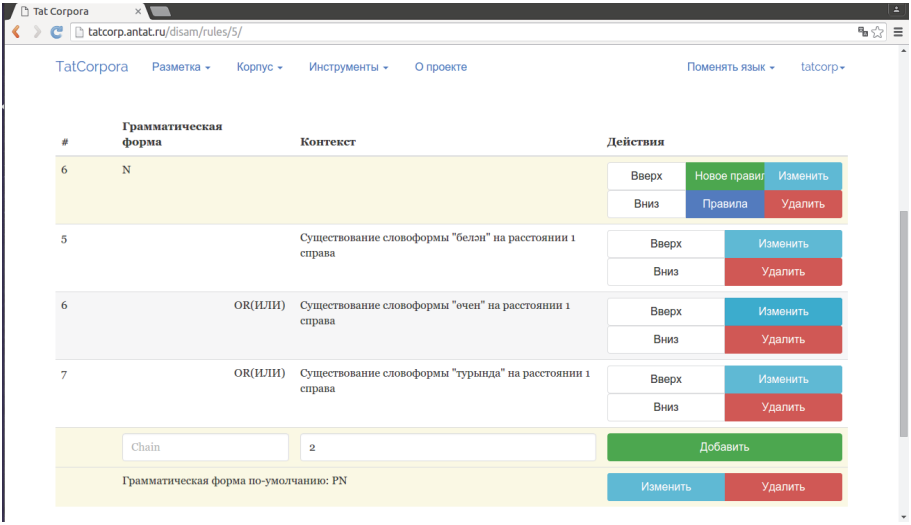


Рис. 1. Основная страница веб-приложения
по разработке контекстных правил

Исходя из этих соображений, был приняты простые правила:

- нельзя удалять и редактировать правила других пользователей. При желании доработать правило, нужное правило дублируется с припиской к текущему пользователю и все улучшения делаются там;

•пока правило не добавлено в основную базу правил, она находится в зоне “песочницы”, где она может редактироваться и тестироваться;

•перед добавлением (либо обновлением) в основную базу, правило тестируется на корпусных данных, и при условии успешного прохождения тестов, правило добавляется в основную базу и может быть применено в процессе разрешения многозначности.

Как уже было описано ранее, разработка правил состоит из нескольких этапов:

•сначала выбирается тип омоформ, для разрешения которых разрабатываются правила;

•потом выбирается управляющая структура, т.е. в процессе лингвистического анализа выявляются множества минимальные разрешающие контексты и упорядочиваются по частотности так, что самые частые случаи будут рассматриваться первыми;

•для каждого множества минимальных контекстов определяется порядок минимальных контекстов (см. Рис.2).

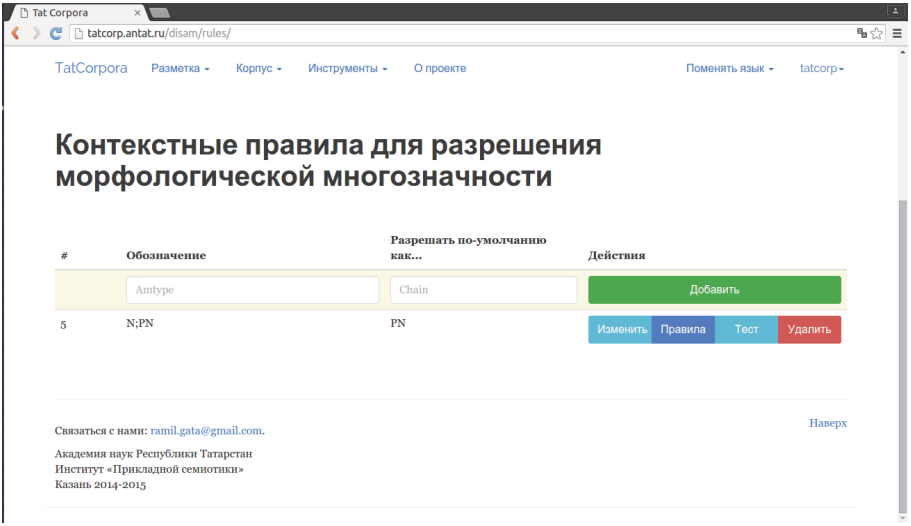


Рис. 2. Страница управляющей структуры правила
с определенным множеством минимальных контекстов

Кроме самой разработки имеется возможность тестирования правил на корпусных данных Очевидно, таким образом, легко выявляются исключительные и ошибочные случаи распознавания, что намного облегчает процесс разработки.

Заключение

В данной работе представлен инструментарий разработки контекстных правил для разрешения морфологической многозначности в корпусе татарского языка. При разработке особый упор делается на “краудсорсинговый” аспект приложения. На данный момент приложение на этапе тестирования. В разработке участвуют 4 человека, было разработано 9 тестовых правил, которые по большей части покрывают исключительные случаи морфологической многозначности.

Следующим шагом планируется реализация статистико-вероятностных методов для разрешения и компоновка их с методом контекстных правил и их совокупности на материале корпусных данных; получены предварительные результаты по разработке методики автоматической классификации татарских текстов из корпусной коллекции по набору морфологических признаков.

Литература

1. Сулейманов, Д.Ш. Двухуровневое описание морфологии татарского языка / Д.Ш. Сулейманов, Р.А. Гильмуллин // Тезисы Международной научной конференции "Языковая семантика и образ мира". Казань: Изд-во Казан. гос. ун-та., 1997. Книга 2. С. 65-67.

2. Разрешение грамматической многозначности в корпусе татарского языка / Б.Э.Хакимов, Р.А.Гильмуллин, Р.Р.Гатауллин // Учен. зап. Казан. ун-та. Сер. Гуманит. науки. - 2014. - Т. 156, кн. 5. - С. 236-244.

3. Веб-инструментарий для снятия морфологической многозначности в текстовом корпусе татарского языка / Р. Р. Гатауллин // Сохранение и развитие родных языков в условиях многонационального государства: проблемы и перспективы: материалы V Международной научно-практической конференции (Казань, 19-22 ноября 2014 г.). – Казань: Отечество, 2014. - С. 71-73

4. Программный инструментарий для разрешения морфологической многозначности в татарском языке / Р.Р. Гатауллин, Д.Ш. Сулейманов, Р.А. Гильмуллин // Открытые семантические технологии проектирования интеллектуальных систем OSTIS-2014 Open Semantic Technologies for Intelligent Systems МАТЕРИАЛЫ IV МЕЖДУНАРОДНОЙ НАУЧНО-ТЕХНИЧЕСКОЙ КОНФЕРЕНЦИИ (Минск, 20-22 февраля 2014 года), - Минск. : БГУИР, 2014. - С. 503-508.

5. Ю.В. Зинькина, Н.В. Пяткин, О.А. Невзорова, Разрешение функциональной омонимии в русском языке на основе контекстных правил. // Труды межд. конф. Диалог'2005.– М.: Наука, 2005. С. 198-202.

6. Crowdsourcing morphological annotation / Bocharov V.V., Alexeeva S.V., Granovsky D.V., Protopopova E.V., Stepanova M.E., Surikov A.V. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая–2 июня 2013 г.). Вып. 12 (19). — М.: РГГУ, 2013.

УДК 81’33

МНОГОФУНКЦИОНАЛЬНАЯ МОДЕЛЬ ТЮРКСКОЙ МОРФЕМЫ: ОТДЕЛЬНЫЕ АСПЕКТЫ

Д.Ш. Сулейманов,
НИИ «Прикладная семиотика» АН РТ
dvdt.slt@gmail.com
А.Р. Гатиатуллин
НИИ «Прикладная семиотика» АН РТ
agat1972@mail.ru
А.Б. Альменова,
НИИ «Прикладная семиотика» АН РТ
almen_akmaral-baijan@mail.ru
А.М. Баширов
ООО "ТемирТех"
a.basheerov@gmail.com

Статья содержит описание многофункциональной модели тюркской морфемы, которая заполненная соответствующим контентом, может иметь различное практическое применение, прежде всего, как ресурсная база для программных продуктов, осуществляющих компьютерную обработку тюркских языков. Эта модель представляет собой эффективный инструмент в сравнительных исследованиях ученых-тюркологов, в частности, для сравнительного анализа тюркских языковых единиц.

Основным компонентом многофункциональной модели является реляционно-ситуационная модель, которая используется для описания целого ряда аспектов модели.

Ключевые слова: *тюркская морфема, многофункциональная модель тюркской морфемы, реляционно-ситуационная модель, лингвопроцессор.*

В настоящее время, для повышения эффективности разработок по компьютерной обработке тюркских языков, необходима единая технологическая база, которая может быть использована в многоязычной системе машинного перевода для тюркских языков, а также системе многоязычного информационного поиска в сети Интернет на тюркских языках. Для разработки такой технологической базы требуется целый ряд лингвистических ресурсов, моделей и словарей. В данной статье описывается один из таких ресурсов - многофункциональная модель тюркской морфемы.

Многофункциональная модель тюркской морфемы построена на основе структурно-функциональной модели татарской аффиксальной мор-

фемы [1], которая расширена для работы с несколькими языками, а также дополнена описанием корневых морфем. Многофункциональная модель представляет собой систему из нескольких лингвистических моделей, основным элементом которых является морфема. Все эти модели объединены в единую систему с помощью информационно-программной оболочки, которая представляет собой технологический инструментарий, как для заполнения базы данных, так и использования базы данных в практических приложениях. Эта программно-информационная оболочка реализована в виде Web-интерфейса, что должно позволить участвовать в процессе заполнения модели специалистам по разным тюркским языкам. Программно-информационная оболочка с заполненной базой данных практически будет готовой информационно-справочной системой о морфемах тюркских языков.

Многофункциональная модель тюркской морфемы имеет иерархическую структуру, состоящую из множества подмоделей. Описание свойств морфем каждого из тюркских языков является отдельной подмоделью общей модели. Общая схема модели представлена на рис. 1.

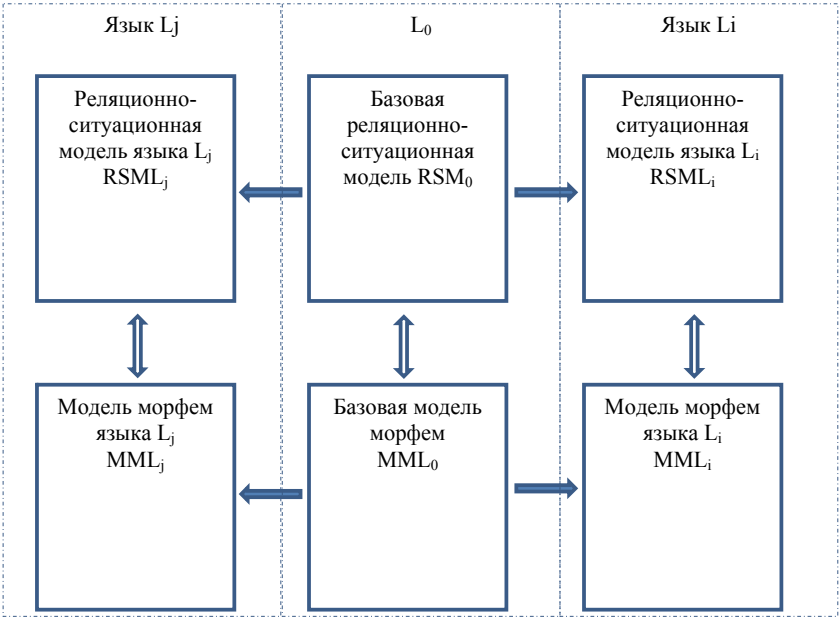


Рис.1. Общая схема модели

На рис.1 представлено, что в основе модели лежит базовая модель, которая является общей для всех языков тюркского семейства и на ее ос-

нове производится наполнение подмоделей каждого конкретного языка. На рис. 1 тюркские языки показаны всего двумя представителями Li и Lj, в реальной для каждого тюркского языка реализуется своя подмодель.

На данной схеме отношения, показанные с помощью стрелок от базовой модели MML_0 к конкретной языковой модели MML_i являются отношениями типа класс-подкласс.

Как показано на рис.1. модели морфем MML_i каждого из языков связаны со своей реляционно-ситуационной подмоделью $RSML_i$. Реляционно-ситуационная подмодель [2] языка L_i является наполнением базовой реляционно-ситуационной модели RSM_0 информацией из языка L_i. Базовая реляционно-ситуационная модель представляет собой систему семантических универсалий, с помощью которых описываются значения морфем. В модели морфем семантические универсалии реализуются с помощью реляционно-ситуационных фреймов.

Реляционно-ситуационный фрейм представляет собой реализацию типовой ситуации, состоящей из названия ситуации и набора слотов, которые являются ролями конструктивных элементов этой ситуации.

Система слотов, заполненная примерами из языка L_i, образует реляционно-ситуационную модель языка L_i.

На рис.2. представлен пример реляционно-ситуационного фрейма, описывающего действия по пространственному перемещению объекта.

```
Situation 7.3.: action_local
    Object:       S1;
    Old_local:    S2;
    New_local:    S3;
    Direction:          S4;
    Route:        S5;
    Interval:           S6;
    Time:         S7;
    Period:       S8;
    Instrument:   S9;
End_Situation
```

Рис.2. Пример реляционно-ситуационного фрейма

В многоязычной модели тюркской морфемы реляционно-ситуационная система необходима для описания разных аспектов модели, но в первую очередь она позволяет производить сопоставление значений аффиксальных морфем разных тюркских языков. Так если аффиксальные морфемы именных категорий в разных тюркских языках совпадают и по форме и значению, то глагольные аффиксы в разных тюркских языках мо-

гут не совпадать. Отдельному аффиксу на одном тюркском языке может соответствовать целая конструкция из нескольких морфем на другом языке. В эту цепочку могут входить как аффиксальные, так и корневые морфемы. Например, казахской аффиксальной морфеме -АтЫн в татарском языке соответствует сочетание: -Й тор-ГАн

баратын (казах.) = бара торган (татар.).

Также в одних тюркских языках есть такие аффиксы, которым на других языках в разных контекстах соответствуют разные морфемные цепочки. Для описания таких контекстов необходима система описания семантических контекстов, реализуемая в нашей модели в виде реляционно-ситуационной модели.

Заключение

В статье дается концептуальное описание многофункциональной лингвистической модели тюркских морфем для которой идет процесс уточнения структуры, создания информационно-программной оболочки и заполнения базы данных для разных тюркских языков. Весьма конструктивным и продуктивным представляется использование данной многофункциональной и многоязычной модели тюркских морфем в качестве одного из центральных, ядерных, модулей в едином веб-портале для тюркских языков. Авторы статьи выражают также надежду, что данный проект послужит интеграции усилий ученых-тюркологов для расширения базы данных описаниями различных тюркских языков, что обеспечит эффективное использование многофункциональной модели в качестве технологического инструментария и межязыкового модуля в системах компьютерной обработки тюркских языков.

Литература

1. Сулейманов Д.Ш., Гатиатуллин А.Р. Структурно-функциональная компьютерная модель татарских морфем. - Казань: Фэн, 2003. - 220с.

2. Сулейманов Д.Ш., Гатиатуллин А.Р., Вагапов Д.Р. Семантико-синтаксическая модель татарского предложения в контексте реляционно-ситуационной системы // В сб. Трудов: Откр-е семант-е техн-ии проект-ия инт-х систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2013): матер. III Межд. научн.-техн. конф. (Минск, 21-23 февраля 2013г.) / редкол.: В.В. Голенков (отв. ред) [и др.]. - Минск: БГИУР, 2013. - С.329-332.

УДК 004.89

ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ПРОГНОЗИРОВАНИЯ ФИНАНСОВЫХ РЫНКОВ

Ф.М. Гафаров, З.Т. Галимханова
Казанский федеральный университет, Казань
fgafarov@yandex.ru, zuhra-1996@mail.ru

Прогнозирование финансовых временных рядов – это важный элемент любой инвестиционной деятельности. Для успешной игры на финансовом рынке необходимо разработать систему, которая на прошлом поведении временных рядов позволит прогнозировать динамику в последующие моменты времени. В данной работе исследовано возможности применения нейронных сетей с прямой связью для прогнозирования динамики курсов валют по отношению к американскому доллару.

Ключевые слова: *финансовые рынки, прогнозирование временных рядов, валюта, валютный рынок, валютный курс, искусственный интеллект, нейронные сети.*

Одним из перспективных научных направлений применения искусственного интеллекта (ИИ) является анализ прогнозирование динамики временных рядов. Среди важнейших инструментов в области ИИ следует особо выделить нейронные сети. Основные преимущества нейронных сетей -это: приспособленность к работе с зашумленными данными; возможность к обучению и адаптации в автоматическом режиме; учет качественных данных, плохо поддающихся формализации; способность учитывать произвольно большое количество факторов; универсальность (широкий класс задач решается с использованием, в большинстве случаев, 5-7 стандартных архитектур). Несмотря на то, что нейронные сети хорошо зарекомендовали себя в решении задач во многих областях человеческой деятельности, вопросы их эффективного применения для прогнозирования на финансовых рынках недостаточно изучены [1].

Финансовый рынок - это сложная система, находящееся в постоянном движении, в которой происходят процессы распределении и перераспределении под воздействием меняющегося соотношения спроса и предложения на эти ресурсы со стороны экономических субъектов. Финансовый рынок представляет механизм, который обеспечивает мобилизацию всех свободных денежных капиталов и денежных средств и

распределение этих средств по отраслям сфер хозяйства на условиях срочности, платности, возвратности. Валюта — это денежные средства иностранных государств, национальные денежные средства, обслуживающие международные операции, а также межнациональные расчетные денежные единицы. Валютный рынок — это место, где осуществляется купля-продажа иностранной валюты. Мировой валютный рынок — это особый, организационно оформленный механизм, обслуживающий и регулирующий отношения по переходу права собственности на валютные ценности на основе закона спроса и предложения. Валютный курс — это цена денежной единицы одной страны, выраженная в денежных единицах других стран [2].

Под нейронными сетями подразумеваются вычислительные структуры, которые моделируют простые биологические процессы, обычно ассоциируемые с процессами человеческого мозга [3]. Они представляют собой распределенные и параллельные системы, способные к адаптивному обучению. Элементарным элементом в данных сетях является искусственный нейрон или просто нейрон, названный так по аналогии с биологическим прототипом. Нейрон представляет с собой элемент, который вычисляет выходной сигнал (по определенному правилу) из совокупности входных сигналов. У нейронных сетей много важных свойств, но ключевое из них – это способность к обучению. Обучение нейронной сети в первую очередь заключается в изменении «силы» синаптических связей между нейронами. Существуют множество применений нейронных сетей для решения задачи прогнозирования временных рядов. Обычно при прогнозировании временных рядов используются многослойные, чаще всего трехслойные, нейронные сети прямого распространения. Класс задач, которые можно решить с помощью нейронной сети, определяется тем, как сеть работает и тем, как она обучается. При работе нейронная сеть принимает значения входных переменных и выдает значения выходных переменных. Таким образом, сеть можно применять в ситуации, когда имеется определенная известная информация, и необходимо из нее получить некоторую пока не известную информацию. Как правило, нейронная сеть используется тогда, когда неизвестен точный вид связей между входами и выходами, - если бы он был известен, то связь можно было бы моделировать непосредственно. Другая существенная особенность нейронных сетей состоит в том, что зависимость между входом и выходом находится в процессе обучения сети.

Для обучения нейронных сетей применяются алгоритмы двух типов: с учителем и без учителя [4]. Для обучения учителем пользователь должен подготовить набор обучающих данных. Эти данные представляют собой примеры входных данных и соответствующих им выходов. Сеть учится устанавливать связь между первыми и вторыми. Обычно обучающие

данные берутся из исторических сведений. Затем нейронная сеть обучается с помощью того или иного алгоритма управляемого обучения, при котором имеющиеся данные используются для корректировки весов и пороговых значений сети таким образом, чтобы минимизировать ошибку прогноза на обучающем множестве. Если сеть обучена хорошо, она приобретает способность моделировать функцию, связывающую значения входных и выходных переменных, и впоследствии такую сеть можно использовать для прогнозирования в ситуации, когда выходные значения неизвестны. В этой работе мы использовали это алгоритм обучения. Наиболее распространённым и эффективным методом обучения многослойных нейронных сетей прямого распространения является алгоритм обратного распространения. Обучение алгоритмом обратного распространения ошибки предполагает два прохода по всем слоям сети: прямого и обратного. При прямом проходе входной вектор подается на входной слой нейронной сети, после чего распространяется по сети от слоя к слою. В результате генерируется набор выходных сигналов, который и является фактической реакцией сети на данный входной образ. Во время прямого прохода все синаптические веса сети фиксированы. Во время обратного прохода все синаптические веса настраиваются в соответствии с правилом коррекции ошибок, а именно: фактический выход сети вычитается из желаемого, в результате чего формируется сигнал ошибки. Этот сигнал впоследствии распространяется по сети в направлении, обратном направлению синаптических связей. Синаптические веса настраиваются с целью максимального приближения выходного сигнала сети к желаемому. Нейронные сети могут работать с числовыми данными, лежащими в определенном ограниченном диапазоне. Это создает проблемы в случаях, когда данные имеют нестандартный масштаб, когда в них имеются пропущенные значения, и когда данные являются нечисловыми. Необходимо чтобы значения входов и выходов находились в приделах области значений функции активации. Для этого применяются нормировка и предобработка данных.

Существует два основных метода прогнозирования на финансовом рынке: фундаментальный и технический. В отличие от этих методов анализа, основанных на общих рекомендациях и опыте трейдера, нейросети способны строить оптимальную модель прогнозирования. Другим преимуществом является, то что нейросетевая модель адаптивна и меняется вместе с рынком, что особенно важно для современных финансовых рынков. Нейросеть пытается распознать в текущем состоянии рынка ранее встречавшуюся ситуацию и максимально точно воспроизвести реакцию рынка [5]. Векторы из обучающей выборки лучше подавать на вход в случайном порядке. Количество входов и выходов обычно диктуются условиями задачи, а количество нейронов скрытого

слоя экспериментально. Обычно число нейронов в нем составляет 30-50% от числа входов. Слишком большое число нейронов скрытого слоя приведет к тому, что сеть теряет способность к обобщению (она просто досконально запоминает элементы обучающей выборки и не реагирует на схожие образы). Если же число нейронов скрытом слое слишком малое, сеть оказывается не в состоянии обучаться.

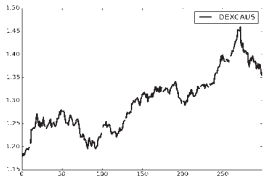
В данной работе были использованы дневные обменные курсы 13 валют по отношению к американскому доллару (DEXUSEU, DEXCHUS, DEXJPUS, DEXCAUS, DEXUSUK, DEXBZUS, DEXMXUS, DEXUSAL, DEXKOUS, DEXSZUS, DEXINUS, DEXTHUS, DEXVZUS) полученные из базы данных Федеральной резервной системы Federal Reserve Economic Data (FRED)[6]. FRED - это база данных подразделения Исследования Федерального резервного банка Сент-Луиса, у которого есть больше чем 237 000 экономических временных рядов из 68 источников. Анализ был проведен на двух интервалах данных: первый- с 2015-01-03 по 2016-02-27, второй- с 2015-01-03 по 2016-02-27, т.е. брался интервал в один и два года. Данные были разделены на обучающую (80%) и тестовую (20%) выборку.

Для построения и обучения нейронных сетей использовалась библиотека PyBrain [7]. PyBrain — одна из лучших Python библиотек для изучения и реализации большого количества разнообразных алгоритмов, связанных с нейронными сетями. Она предназначена для исследователей, студентов, лекторов, разработчиков. Это модульная библиотека, предназначенная для реализации различных алгоритмов машинного обучения на языке Python. Основной его целью является предоставление исследователю гибких, простых в использовании, но в то же время мощных инструментов для реализации задач из области машинного обучения, тестирования и сравнения эффективности различных алгоритмов.

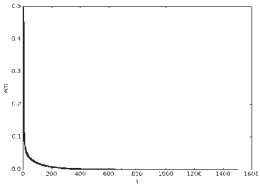
В данной работе использовалась трехслойная нейронная сеть с прямой связью, созданная на основе класса FeedForwardNetwork библиотеки PyBrian со следующей архитектурой: 1 слой – линейная функция активации, 2 слой – сигмоидная функция активации, 3 слой - выходной слой, гиперболический тангенс. Количество нейронов на скрытом слое двадцать четыре, на выходном один., для входного слоя количество нейронов менялось от четырех до двадцати с шагом два. В целях получения более обширной статистической картины для каждого типа нейронной сети мы создали и обучали по пять экземпляров сети. Значимыми для предсказаний являются изменения котировок, поэтому на вход нейронной сети подавались изменения котировок. Для обучения сети использовали стохастический метод градиентного спуска в пространстве весовых коэффициентов (BackpropTrainer библиотеки Pybrain). Нейронная

сеть обучалась прогнозировать изменения курса на N+1 день при подаче на вход изменений курса за N предыдущих дней.

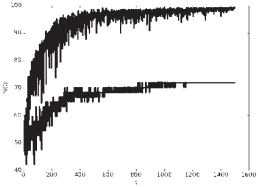
На рисунке 1 представлен обменный курс Канадского доллара по отношению к Американскому доллару с 1 февраля 2015 года по 1 февраля 2016 года. На рисунке 2 показано изменение ошибки нейронной сети в процессе обучения, видно, что в процессе обучения ошибка сети быстро уменьшается. В процессе обучения, мы параллельно анализировали динамику точности прогноза сети на обучающем и тестовом наборах. Если нейронная сеть правильно определяла направление движения рынка (рост или падение курса валюты) на следующий день, то считается что нейронная сеть выдает верный прогноз. На рисунке 3 показано как изменяется процент точности прогноза нейронной сети на обучающем (верхний график) и тестовом наборе данных. Видно, что на наборе обучения нейронная сеть обучается прогнозировать почти до 100%, а на тестовом наборе приблизительно 72%.



(рис.1)



(рис.2)



(рис.3)

Выводы

Нейронная сеть является очень мощным инструментом для анализа и прогнозирования на финансовых рынках. Плюсом нейросетей является объективность при принятии стратегических решений, а минусом – то, что решение принимает фактически черный ящик. Мы проанализировали, как влияет длина входного вектора на точность прогнозов. Оказалось, что для разных валют эти значения разные. Например, для прогноза курса английского фунта наиболее точные результаты показали нейронные сети из 10 входов, а для прогноза швейцарского франка наиболее эффективным оказалось 8 входов. Анализ результатов обучения на всех имеющихся наборах данных для всех валют показал, что наиболее точные прогнозы получаются при длине входных векторов от 8 до 12. Необходимо отметить что финансовый рынок является очень сложной системой, динамика которой зависит от множества факторов. Выявить закономерности в такой системе чрезвычайно сложно, и поэтому нейронные сети показывают

недостаточно высокие результаты (в основном 55%-65% точности). Поэтому нейронные сети должны быть использованы совместно с другими методами прогнозирования фундаментального и технического анализа для повышения точности прогноза.

Литература

1. Бэстенс Д-Э., Ван Ден Берг В.-М., Вуд Д. Нейронные сети и финансовые рынки. Издательство: ТВП, Москва. 1997 год, -254 с.
2. Энг М.В., Лис Ф.А., Мауер Л.Дж., Мировые финансы, ДеКА - 1998, 768 -с.
3. С.Хайкин. Нейронные сети. Полный курс. Издательство: Вильямс. 2006 год, -1104 с.
4. А. И. Галушкин. Нейронные сети. Основы теории. Издательство: Горячая Линия — Телеком. 2010 год, 496 -с.
5. В.И.Ширяев. Финансовые рынки. Нейронные сети, Хаос и нелинейная динамика. М.: Красанд, 2011. — 232 с. — ISBN 978-5-396-00388-0.
6. https://research.stlouisd.org/fred2/ -Federal Reserve Economic Data
7. http://pybrain.org – The Python Machine Learning Library.

УДК 004.91

МОРФОЛОГИЧЕСКИЙ АНАЛИЗАТОР КАК DLL В КРОССПЛАТФОРМЕННОЙ СИСТЕМЕ JAVA

В.П. Желтов, А.Р. Губанов
Чувашский государственный университет им. И.Н. Ульянова, Чебоксары
chnk@mail.ru, AlexGubM@gmail.com

В современных поисковых системах модуль морфологического анализа используется при индексировании текстов, дает возможность формировать поисковый запрос на естественном языке. В качестве аппарата для построения лингвистических моделей, для моделирования взаимодействия компонентов системы, т. е. в качестве мета-аппарата в МА применяются сети Петри.

***Ключевые слова**: морфологический анализатор, DLL, кроссплатформенная система Java.*

В современных поисковых системах модуль морфологического анализа используется при индексировании текстов, дает возможность формировать поисковый запрос на естественном языке(Апресян, Богуславский 1992, Гатиатуллин 1998, Желтов, Желтов, Губанов 2015).

При создании морфологических библиотек(DLL) следует обратить внимание на адаптации одной системы с другой, чтобы программные продукты, связанных с автоматическим морфологическим анализом текстов, были доступны различным приложениям в любой прикладной области.

В современном компьютерном мире предлагаются для разработчиков различные морфологические модули, предназначенные для решения широкого класса задач. В частности, продукт PCO Morphology поставляется в виде динамической библиотеки для Widows, предназначенный для разработчиков информационно-поисковых и аналитических систем. С компьютерным морфологическим анализом связана также, система Mystem, имеющая непосредственное отношение к интересующей нас системе Java. Морфологический анализатор Mystem в системе Java реализован в DLL - JavaMystem. Модуль же татарского МА реализован на базе программного инструментария PC-Kimmo, использующую файл фонологических правил (файл правил), сгенерированный PC-KIMMO, а именно: в начале работы программа считывает из файла правил алфавит, специальные символы, множества, все возможные соответствия и правила; на вход поступает слово в лексической форме, инициализируются состояния автоматов каждого правила. Модуль распознавания словоформ – это морфологический анализ словоформ, по сути, есть обратная функция генерации; модуль распознавания использует файл морфотактических правил, а также файл аффиксальных и корневых лексем. Платформа Java, используя JKimmo, обменивается с PC-Kimmo на основе структуры данных Kimmo Data и Kimmo Result.

В качестве аппарата для построения лингвистических моделей, для моделирования взаимодействия компонентов системы, таких как формальные модели, структуры данных и алгоритмы, т. е. в качестве мета-аппарата в МА чувашского исследователя П.В. Желтова(Желтов 2002) применяются сети Петри, где процесс моделирования(а), и процесс реализации полученной модели(б) разделяются: а) реализуется с помощью сетей Петри и абстрагирован от средств программной реализации; б)проводится по имеющейся сетевой модели в терминах и теории, близких к средствам современных программных продуктов. Кстати, объектно-ориентированные сети Петри реализуются в Java Eclipse при помощи обертки – DLL J pipe-alpha-2.0.

Все вышеперечисленные модели и алгоритмы, как видно, традиционные. Современные системы в своей структуре используют заранее заготовленные фреймы согласно той гипотезе, что мышление человека оперирует фреймовыми структурами знаний разной организации — планами, сценариями, схемами. Аналоги этих структур, как показывают наши наблюдения, используются в системах АОТ и

искусственного интеллекта. Одним из новшеств морфологического анализатора в этой области является является казахский интеллектуальный морфологический анализатор, основанный на семантический сетях, т.е. для формализации правил добавления суффиксов и окончаний предлагается использовать семантическую нейронную сеть, с помощью которой генерируются словоформы казахского языка, и порождается структура словаря начальных форм в виде синхронизированного линейного дерева; нейроны распознают отдельные символы входной символьной последовательности, а на выходе генерируется сигнал, означающий наличие или отсутствие соответствующего символа в анализируемом тексте (нейроны выдают результат распознавания отдельных фрагментов входной символьной последовательности).

МА, как известно, рассматривается как сбалансированный комплекс аппаратных, программных, лингвистических, а иногда и лингводидактических средств, взаимодействующих с мощной базой лингвистических данных и знаний (БДЗ). Важное значение в конкретной реализации модели морфологии большую роль играет реляционный аппарат, потому что позволяет представить данные в удобной форме в виде отношений (таблиц), легко реализуемых средствами проектирования БД, и тем самым облегчается извлечение данных из БД, и отпадает необходимость создания сложных процессоров для их извлечения, как в случае их записи в виде продукционных правил в файле. В частности, в анализаторах шорского, тувинского, якутского словари основ автоматически извлекаются с помощью СУБД STARLING, где словарь основ представляет собой размеченную базу данных, содержащую слова в начальной форме (леммы) и не восстановимые из начальной формы варианты чередований (с использованием технологий той же системы конвертирован в базу данных инвентарь морфем хакасского языка).

В отличие от вышеназванных словарей чувашский словарь основ носит словообразовательный характер, а схема следования аффиксов и словарь морфем носят как словоизменительный, так и словообразовательный характер (словарь основ относится к шаблонам, так как в нем не описываются какие-либо морфологические характеристики: все морфологические характеристики лексемы можно определить по найденным в ней аффиксам, причем в отличие от русского языка одна морфема определяет одну морфологическую характеристику (редко 2 или 3) и никогда не встречаются случаи, когда несколько морфем определяют одну морфологическую характеристику, т.е. в чувашском языке отношение между морфемами слова и его морфологическими характеристиками однозначно. Поэтому, зная из словаря основ словообразовательные аффиксы, а из схемы следования – словоизменительные, МА из словаря морфем точно может определить морфологические характеристики всей

словоформы. Однако проблемным для БД чувашского языка остается та часть списка лемм, которая включает в себя собственные имена — антропонимы, топонимы (проблему следует постепенно решать в процессе совершенствования программы по принципу «обучения с учителем»).

Нельзя обойти здесь вопрос, связанный с оценкой качества морфологического анализа. В литературе в настоящее время в основном используется подход к оценке качества морфологического анализа на основании количества форм слов, неправильно отнесенных к леммам, на основе чего выделяют следующие виды ошибок лемматизации: 1. Under-stemming, когда морфологические формы одного слова относят к разным леммам; 2. Over-stemming, когда разные слова ошибочно относят к одной лемме. Для оценки этих видов ошибок, вводят две метрики UI – under-stemming index - процент терминов, для которых данный модуль лемматизации совершил ошибку under-stemming и OI – over-stemming index – процент слов, для которых морфологический модуль совершил ошибку over-stemming. В основном, соответствующие используемые методы имеют следующие недостатки: качество работы морфологических алгоритмов проверяется на относительно небольшой выборке слов, участвовавших в тестируемых запросах, и для заданной коллекции документов, и при этом возможны ошибки, связанные с особенностями определенного морфологического модуля для данной конкретной группы слов или документов.

Как лучше использовать программный продукт МА в кроссплат-форменной среде. На наш взгляд, универсальным программным продуктом могут выступать так называемые DLL, имеющая такие преимущества, которые предоставляет их использование разработчику: 1) повторное использование кода(функции, хранящиеся в библиотеке, могут быть вызваны на выполнение из приложений, разработанных в системах MS VS Delphi Java и др..; 2) возможность использования загруженного в оперативную память кода несколькими приложениями; 3) при выделении общих для нескольких приложений данных в DLL может привести к экономии как дискового пространства, так и оперативной памяти, иногда очень даже существенному.

Важно, с одной стороны, представить МА как DLL, а с другой стороны, как вызвать эту DLL из другой программы, с другой платформы, в нашем случае из Java. Для этого в программе на Java можно использовать native-методы, расширяющих стандартные возможности Java. Подсистема Java, реализующая эту возможность, называется JNI (Java Native Interface – интерфейс языка Java, позволяющий обращение к native-методам), т.е. «среда» JNI представляет собой массив указателей на методы, используя которые можно осуществлять доступ к полям класса Java и осуществлять вызов методов класса Java (передаваемый же native-

методу объект позволяет осуществить привязку методов JNI непосредственно к конкретному объекту и классу объектов, с которыми может работать native-метод. Анализируемая нами технология включает следующие моменты: а) создание собственного JAVA-метода (для того чтобы передать управление C/C++-коду из JAVA-программы; разработка функций, в которые будет передаваться управление, и оттранслировать их, поместив в библиотечный файл(после создания библиотеки ее можно загружать из JAVA-программы для последующего вызова собственных методов; б) создание заголовочного файла (можно написать вручную или воспользоваться утилитой JAVAH; в) правила формирования имени C/C++-функции(JNI определяет 210 прикладных функций; использование JNI функций необходимо в том случае, если C/C++-функция осуществляет какое-либо взаимодействие с JVM: вызов JAVA-методов, доступ к данным, создание JAVA-объектов и т.д.

Вышеуказанный комплекс действий можно использовать для создания адаптированного МА на основе расширения системы Hunspell для чувашского анализатора и Java обертки для данной системы МА- Java master Hunspell. Разработанная на этой основе концептуальная модель позволяет создавать морфологический анализатор нового типа, представляющий морфологические характеристики текста на чувашском языке. Обоснованным представляется не первом этапе в морфологической разметке указать лишь грамматические категории, явно выраженных аффиксами(эффективность автоматической разметки можно достичь путем введения дополнительных фонологических и морфотактических правил для морфологического анализатора. В этой области следующие исследовательские задачи представляются перспективными: анализ возможностей и разработка правил автоматического определения значений полифункциональных и омонимичных аффиксов, "нулевых" форм, правил автоматического распознавания однородных групп; исследование возможностей создания фонологических и морфотактических правил для специфичных классов лексически, не подчиняющихся общим закономерностям, в первую очередь, заимствований и некоторые другие.

На рынке программных средств существует многообразие средств компьютерного моделирования. МА. адаптированный МА в системе Java, на наш взгляд, можно представить: а) как апплет; б) как плагин (в системеJava Eclipse RCP можно создавать свои плагины; пользовательский интерфейс RCP приложений основан на визуальных компонентах фреймворков SWT и JFace, а также на собственных Eclipse виджетах).

Созданный МА можно использовать также как OLE (Object Linking and Embedding) и как COM объект. OLE позволяет создавать объекты в одном приложении, а затем отображать эти объекты в других

приложениях. Основной задачей OLE-автоматизации является обеспечение взаимодействия компонентов и приложений независимо от языков программирования и средств разработки.OLE (OLE Automation) позволяет программно управлять другими приложениями, вызывая их методы, доступные через интерфейс OLE. При этом, конечно, требуется, чтобы приложение поддерживало автоматизацию OLE.

COM с самого начала проектировалась для того, чтобы обеспечить функциональность взаимодействия приложений и предоставить возможность дальнейшего развития этой функциональности за счет расширений. Прелесть двоичного COM-сервера в том, что способ доступа к нему не зависит от языка. COM-объект представляет собой либо DLL-библиотеку, либо приложение Windows, которые можно создавать в любой системе программирования, способной поддерживать нужный формат представления. COM-объект (а также объект любой из рассматриваемых далее схожих технологий распределенного взаимодействия, например CORBA или CGI) отличается от таких объектов. Существует JNA (Java Native Access) как COM-объект, применяется JNA в Java-проектах для доступа к функциональности и объектам, так называемых «нативных» библиотек — COM-DLL Microsoft Windows. Такая технология на данном рынке представляет собой большой интерес. Основным преимуществом является сокращение времени разработки проекта, если вся необходимая функциональность уже содержится в какой-то стандартной библиотеке Microsoft Windows, либо есть сторонняя COM-DLL с необходимым набором решений, либо это уже применяемая клиентом COM-DLL бизнес-логики. Вторым, но не меньшим по значимости преимуществом является то, что в отличие от рассмотренной выше технологии JNI, здесь не придется писать библиотеку-оболочку на C.

Благодарности. Публикация подготовлена в рамках поддержанного РГНФ научного проекта №15-04-00532.

Литература

1. Лингвистический процессор для сложных информационных систем / Ю.Д. Апресян, И. М. Богуславский, Л. Л. Иомдин и др. М.: Наука, 1992. 256 с.

2. Васильев В.В., Кузьмук В. В. Сети Петри, параллельные алгоритмы и модели мультипроцессорных систем. Киев: Наукова думка, 1990. 216 с.

3. Вудс В.А. Сетевые грамматики для анализа естественного языка // II Кибернетический сб. М.: Наука, 1976. Вып. 13. С. 121-158.

4. Гатиатуллин А. Р. Интегрированный программно-информационный комплекс «Морфема» // Сб. трудов Междунар. семинара по компьютерной лингвистике и ее приложениям «Диалог-98». Казань, 1998. С.453-466.

5. Григорьев А.В. Представление генетических алгоритмов сетями Петри в задаче размещения: автореф. дис. ... канд. техн. наук // Чувашский гос. ун-т. Чебоксары, 2002. 22 с.

6. Димитриев А.П. Стохастическая оптимизация в задаче размещения на сетях Петри: автореф. дис. ... канд. техн. наук // Чувашский гос. ун-т. Чебоксары, 2001. 23 с.

7. Желтов В.П., Желтов П. В., Губанов А. Р. Морфологический стандарт национального корпуса чувашского языка // Современные проблемы науки и образования. 2015. № 2-1.; URL: http://www.science-education.ru/ru/article/view?id=20578 (дата обращения: 30.03.2016).

7. Желтов П.В., Желтов В. П. Разработка лингвопроцессора для анализа естественного языка // Вестник Чувашского университета. Естественные и технические науки. № 2. Чебоксары, 2002.

8. Захаров В.М., Желтов П. В. Моделирование обработки информации в лингвопроцессорах сетями Петри. Казань: Изд-во Казан. гос. техн. ун-та, 2004. 20 с.

9. Нариньяни А.С. Автоматическое понимание текста - новая перспектива // Труды междунар. семинара «Диалог-97» по компьютерной лингвистике и ее приложениям. Москва, 1997, С. 203-208.

10. Сулейманов Д.Ш., Гатиатуллин А. Р. Формальное описание значений аффиксальных морфем // Сб. трудов междунар. семинара по компьютерной лингвистике и ее приложениям «Диалог-98». Казань, 1998. С. 713-725.

11. Сулейманов Д.Ш., Гатиатуллин А. Р. Интегрированный программно-информационный комплекс «Морфема» // Сб. трудов шестой национальной конференции с международным участием КИИ-98 в трех томах. Т.I. Пущино, 1998. С. 208-214.

12. Шаров С.А. Средства компьютерного представления лингвистической информации. URL: http//nl-web (дата обращения: 30.03.2016).

13. Шевченко Я. В., Желтов П. В. Визуальная среда моделирования различных систем // Вестник Чувашского университета. 2011. № 3. С. 141-145.

УДК 37.022

РАЗРАБОТКА ВИРТУАЛЬНОЙ КЛАВИАТУРЫ ДЛЯ ТАТАРОЯЗЫЧНЫХ ПОЛЬЗОВАТЕЛЕЙ НА БАЗЕ МОБИЛЬНОЙ ОПЕРАЦИОННОЙ СИСТЕМЫ ANDROID

А.В. Данилов, Т.А. Ильясов
Казанский федеральный университет, Казань
tukai@yandex.ru

В статье представлена разработка виртуальной клавиатуры для татарского языка на базе мобильной операционной системы Android. Авторами проанализированы особенности, повлиявшие на процесс разработки. Описаны основные компоненты клавиатуры и их функции. Подробно рассматривается процесс разработки предиктивной системы ввода. Данная разработка позиционируется как пример информа-

ционного решения, направленного на сохранение и развитие татарского языка в социально - гуманитарной сфере.

Ключевые слова: *татарский язык, локализация, Android, клавиатура, предиктивный словарь, мобильная коммуникация*

Республика Татарстан является многонациональной, в ней проживают представители различных культур. По данным Всероссийской переписи населения 2010 года 53,15% населения Республики Татарстан составляют татары [1]. Для сохранения и развития татарского языка необходимо, чтобы он стал языком общения в инфокоммуникационной среде, в частности, чтобы интерфейс программных продуктов и приложений были локализованы. Данная проблема постепенно решается, однако она остается острой для мобильных технологий. Отсутствие необходимого качественного программного обеспечения привело к тому, что многие татароязычные пользователи для мобильной коммуникации используют русскую раскладку клавиатуры. В данной статье представлен процесс разработки и популяризации виртуальной клавиатуры как пример информационного решения, направленного на сохранение и развитие татарского языка в социальной сфере.

Для разработки виртуальной клавиатуры для татарского языка необходимо было решить следующие задачи:

1) проанализировать рынок приложений со схожим функционалом, обобщить полученные данные и выявить положительные и отрицательные стороны существующих программных продуктов;

2) используя имеющиеся положительные наработки, разработать новый программный продукт;

3) опубликовать готовый программный продукт в магазине мобильных приложений Google Play Market.

Для решения первой задачи были найдены и изучены виртуальные клавиатуры для татарского языка, расположенные в сети Интернет [2-5]. Функционал программ в целом аналогичен, так как в них применяется один и тот же принцип соответствия букв и символов определенным клавишам. Анализ эргономических свойств клавиатур проводился по нескольким критериям, а именно:

1) *расположение татарских символов на клавиатуре;*

2) *наличие предиктивного словаря.*

Рассмотрим результаты анализа, полученные с опорой на первый критерий. От расположения символов зависит скорость набора символов и удобство использования той или иной клавиши. Расположение символов на разных программных продуктах представлено на рисунках 1-4.

В проанализированных виртуальных клавиатурах использовались два основных подхода для расположения татарских символов:

- первый – выделение символов в отдельную строку (рис. 1,2),
- второй – расположение символов на клавиатуре аналогично стандарту ЙЦУКЕН-раскладки на татарском языке (рис. 3,4).

Оба имеют недостатки: так, использование первого подхода приводит к недостатку места на дисплее мобильного устройства, поэтому процесс ввода становится неудобным.

При использовании второго подхода, можно сделать вывод, что клавиатуры недостаточно удобны, так как в татарском алфавите на 6 букв больше, чем в русском, приходится одной клавише ставить в соответсвие 2 буквы, например, в раскладке ЙЦУКЕН над буквой “Ц” располагается буква “Ң”.

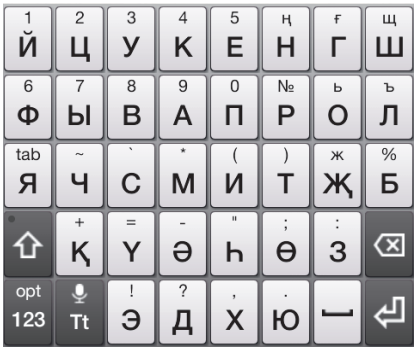


Рис 1. Клавиатура JBak keyboard для татарского языка

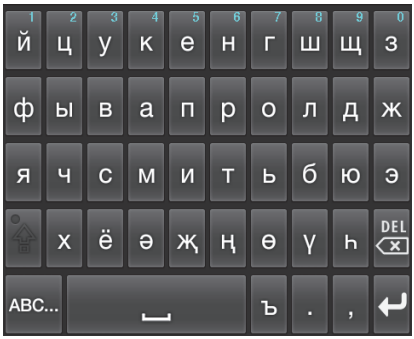


Рис. 2. Клавиатура TatarKey

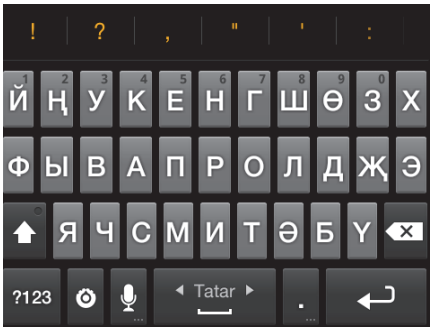


Рис. 3. Клавиатура GingerBread Keyboard для татарского языка

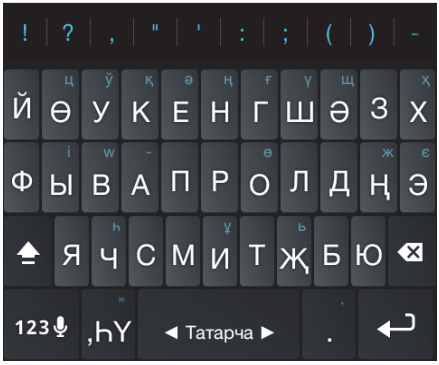


Рис. 4. Клавиатура Multilang Keyboard для татарского языка

Рассмотрим результаты анализа, полученные с опорой на второй критерий. *Предиктивный словарь* существенно влияет на скорость набора, позволяя пользователю вводить слова со специальной панели, не набирая их полностью. Принцип работы клавиатуры с использованием предиктивного словаря сводится к тому, что программа анализирует введенные пользователем буквы и предлагает подходящие слова из предиктивного словаря, что значительно ускоряет ввод слов. Объем и принцип построения словаря определяют его качество. В изученных программных продуктах такой словарь был разработан лишь для одной клавиатуры. Предиктивный словарь состоял из 60 000 слов, и их частотные характеристики не соответствовали тому лексикону, который обычно применяется в мобильном общении. Так, часто используемое в общении слово «сәлам» появляется на панели лишь при вводе с клавиатуры четвертого символа, т.е. при вводе «сәла» (Рис. 5). По

мнению автора, оптимизировав предиктивный словарь, ориентировав его на мобильную речь, можно добиться существенного его улучшения.

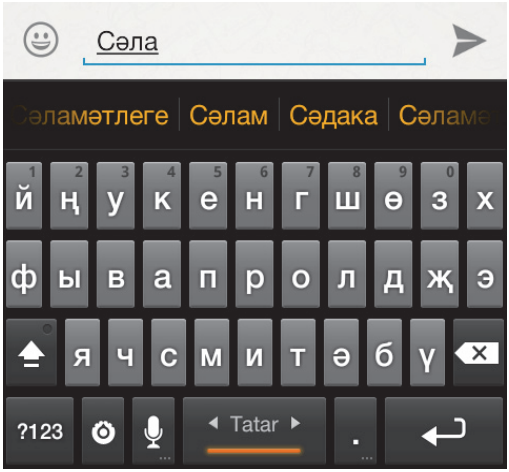


Рис. 5. Работа предиктивной системы на клавиатуре GingerBread Keyboard для татарского языка.

В результате были определены следующие направления разработки программного продукта. Во-первых, была поставлена задача оптимизации расположения символов на виртуальной клавиатуре. Во-вторых, возникла необходимость создания предиктивного татарского словаря, ориентированного на мобильную речь. Авторская идея состояла в разделении системы предиктивного ввода на две подсистемы. Первая подсистема, которая в процессе разработки была названа система предиктивного ввода один (СПВ-1), работает аналогично предиктивному словарю, т.е. при вводе символов система предлагает быстрый ввод для самых часто употребляемых слов. Вторая подсистема (СПВ-2) анализирует введенное слово, далее предлагает устоявшиеся языковые выражения (клише) татарской речи, которые включены в состав предиктивного словаря.

Вторая задача включала в себя разработку приложения и предиктивного словаря, для чего была создана команда разработчиков из сотрудников кафедры математической лингвистики и информационных систем в филологии Казанского федерального университета и Института прикладной семиотики Академии наук РТ. Команда была разделена на группы, исходя из специфики работы. Первая группа работала над программным кодом и дизайном программного продукта, вторая группа составляла предиктивный словарь.

Разработка виртуальной клавиатуры для татарского языка на базе мобильной операционной системы Android включала в себя следующие этапы:

1) разработку дизайна и определение расположения клавиш;
2) построение алгоритма работы клавиатуры и предиктивной системы ввода;
3) составление словарей в соответствии с разработанным алгоритмом;
4) запись кода, включение словарей и компиляция приложения;
5) тестирование приложения.

В результате первого и второго этапов было выбрано расположение татарских символов на виртуальной клавиатуре, приведенное на рисунке 6, а также выработаны базовые принципы работы клавиатуры.

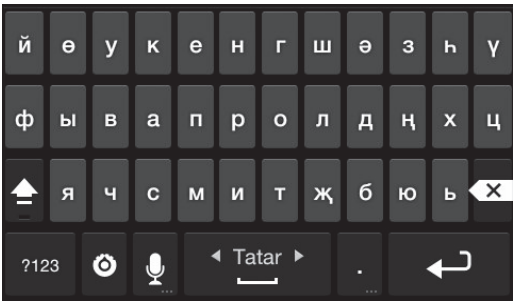


Рис. 6. Финальный вариант раскладки клавиатуры на татарском языке

В результате добавления татарских букв и нехватки места на экране пришлось скрыть некоторые символы (Таблица 1).

Таблица 1

Список скрываемых символов на панели клавиатуры

Символ на клавиатуре	Скрываемый символ
Ә	Э
Ш	Щ
Ь	Ъ
Е	Ё
Җ	Ж

Данный принцип замещения символов родился как компромисс между экономией места на экране и частотой употребления символов: те символы, частота употребления которых меньше остальных, были скрыты

за теми символами, очертание которых наиболее близко к скрываемым символам.

Кроме того, в клавиатуру были включены русская ЙЦУКЕН-раскладка и английская раскладка QWERTY. Данный шаг обоснован особенностями работы мобильной операционной системы Andriod. ОС воспринимает клавиатуру как автономное приложение, которое она запускает или активирует автоматически. Так, если бы в разрабатываемой клавиатуре была только раскладка для татарского языка, то пользователю пришлось бы заходить в меню настроек для того, чтобы выбрать другую клавиатуру, поддерживающую русскую или английскую раскладку. Так как данные раскладки используются часто, было решено включить их в разрабатываемый продукт.

В результате третьего этапа работы был разработан алгоритм работы предиктивной системы. Система разделена на две подсистемы СПВ-1 и СПВ-2. Каждая подсистема работает с собственной табличной базой данных, где указаны слова и выражения, а также частота их употребления. Ниже приведены схемы данных таблиц.

Таблица 2

Схема базы данных для подсистемы предиктивного ввода СПВ-1

Слово	Частота употребления
сәлам	19000

Таблица 3

Схема базы данных для подсистемы предиктивного ввода СПВ-2

Слово_1	Слово_2	Частота употребления
ничек	хәлләр	13000

Принцип работы предиктивной системы показан на рисунке 7.

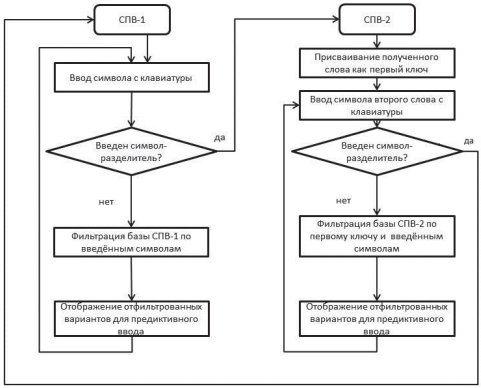


Рис. 7. Алгоритм работы подсистем предиктивного ввода

На четвертом этапе были созданы необходимые словари для предиктивной системы. Поиск и анализ был произведен с использованием ресурсов сети Интернет. В результате было получено два словаря объемом около миллиона слов и словосочетаний.

На пятом этапе в программный продукт были включены предварительные версии словарей, после чего разрабатывался код и тестировалась предиктивная система. Необходимо было оптимизировать код и словари исходя из трех критериев: объем приложения, быстродействие и объем словаря. В результате оптимизации был усовершенствован алгоритм работы предиктивной системы, а также на 50%. укорочены словари Уменьшение объема словарей не сказывается на качестве работы словаря, так как были удалены слова и словосочетания, частота употребления которых крайне мала. Однако количество таких элементов составляет большую часть базы, их удаление уменьшило объем занимаемой памяти, что дало преимущество разработчикам. Также на данном этапе была скомпилирована первая тестовая версия приложения. Процесс работы с приложением представлен на рисунках 8 и 9.

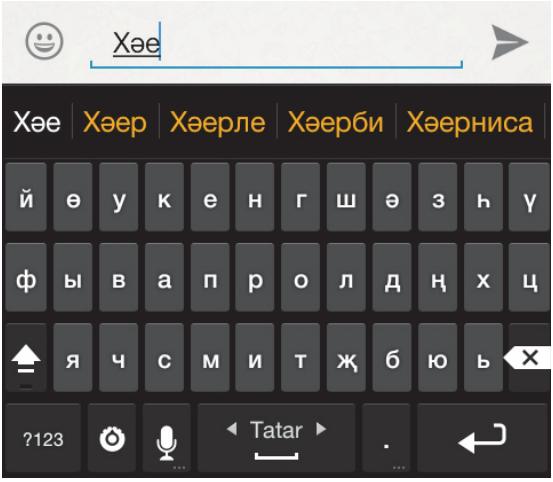


Рис. 8. Работа приложения и подсистемы предиктивного ввода СПВ-1. При наборе символов “*хәе*” подсистема предлагает популярные варианты для быстрого набора

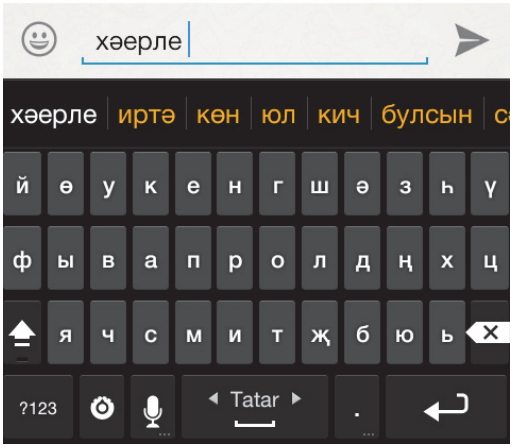


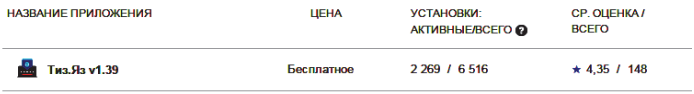
Рис. 9. Работа подсистемы предиктивного ввода СПВ-2.
При введении слова *хәерле* система уже предлагает
для ввода несколько вариантов словосочетаний

На шестом этапе приложение было тщательно протестировано, выявлены ошибки в приложении, которые впоследствии были исправлены.

Третья задача представляла публикацию приложения на площадке распространения программных продуктов для операционной системы

Android – Google Play Market [6]. После релиза данный программный продукт стал доступным для всех пользователей ОС Android.

На момент публикации статьи цель достигнута: создана виртуальная клавиатура для татарского языка на базе мобильной операционной системы Android. Релиз на площадке Google Play Market состоялся 8 декабря 2014 года. Приложение названо «Тиз.Яз» и находится в свободном доступе в сети Интернет. В период с 8 декабря 2014 года по 9 сентября 2015 года данное приложение загружено 6516 пользователями (рис .10)



НАЗВАНИЕ ПРИЛОЖЕНИЯ	ЦЕНА	УСТАНОВКИ: АКТИВНЫЕ/ВСЕГО	СР. ОЦЕНКА / ВСЕГО
Тиз.Яз v1.39	Бесплатное	2 269 / 6 516	★ 4,35 / 148

Рис. 10. Статистика использования программного продукта «Тиз.Яз»

Был разработан программный продукт, который способствует расширению сферы употребления татарского языка как языка инфокоммуникационных технологий.

Литература

1. Информационные материалы об окончательных итогах Всероссийской переписи населения 2010 года / Сайт Федеральной службы государственной статистики, 2010. // URL: http://www.gks.ru/free_doc/new_site/perepis2010/perepis_itogi1612.htm (дата обращения : 22.01.2015)

2. Клавиатура JBak/ 2012. // URL: http://jbak.ru/jbakkeyboard/ (дата обращения: 13.06.2014)

3. Клавиатура Multilang плагин поддержки татарского языка. / 2013 // URL: http://play.google.com/store/apps/details/id/Dklye.plugin.tt (дата обращения: 13.06.2014)

4. Клавиатура GingerBread Keyboard. / 2014 // URL: http://4pda.ru/forum/dl/post/4397320/1336908529_GingerBread_Keyboard.apk (дата обращения: 13.06.2014)

5. Татарская клавиатура TatarKey. / 2014 // URL: https://play.google.com/store/apps/details?id=com.tatarkey&hl=ru (дата обращения: 13.06.2014)

6. Клавиатура Тиз.Яз. /2014 // URL: https://play.google.com/store/apps/details?id=ru.antat.tatar_keyboard (дата обращения: 22.01.2015)

УДК 004 934

ЗАВИСИМОСТЬ ЭНЕРГИИ СЕГМЕНТОВ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ РЕЧЕВОГО СИГНАЛА ОТ ЗНАЧЕНИЯ МАСШТАБНОГО КОЭФФИЦИЕНТА

[1]П.В. Желтов, [2]В.П. Желтов, [3]В.И. Семенов, [4]А.К. Шурбин
ФГБОУ ВПО «Чувашский государственный университет им. И. Н. Ульянова», Чебоксары
[1]chnk@mail.ru, [2]zheltov42@mail.ru,
[3]syundyukovo@yandex.ru, [4]shurti@mail.ru

Картину расположения фонем в слове или предложении можно установить, исследуя зависимость энергии сегментов вейвлет-спектра от масштабного коэффициента. Для исследования используется МНАТ-вейвлет. Вейвлет-анализ речевого сигнала показывает, что гласные фонемы имеют максимальные энергии при средних значениях.

Ключевые слова: *вейвет-преобразование, масштабный коэффициент, сегмент, Фурье преобразование, речевой сигнал.*

Для вычисления вейвлет-спектра речевого сигнала используется формула непрерывного вейвлет-преобразования:

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} S(t)\ \psi\left(\frac{t-b}{a}\right)\ dt\ .$$

Для записи, отображения, воспроизведения, редактирования сэмплов и сохранения в текстовом файле используется звуковой редактор: частота дискретизации акустического сигнала – 8000 Гц, разрешение – 16 бит, режим записи – моно. Длительность речевого сигнала составляет четыре секунды. Сохраненные данные используются для преобразования в вейвлет-спектр $W(a,b)$ исследуемого сигнала $S(t)$. Для вычисления Фурье-спектра сегментов вейвлет-спектра используется преобразование Фурье:

$$F(\nu) = \int_{-\infty}^{\infty} f(t) e^{-i\,2\pi\nu t}\, dt\,.$$

Для вычисления энергии сегментов фонем используется формула Парсеваля:

$$\int_{-\infty}^{\infty} f^{\,2}(t)dt = \int_{-\infty}^{\infty} |F(\nu)|^2 d\nu.$$

Для исследования зависимости энергии сегментов вейвлет-спектра от масштабного коэффициента a вычисляется энергия сегментов функций $W(a,b)$. Полученные вейвлет-коэффициенты (функции) $W(a,b)$ разбиваются на сегменты фиксированной длительности (n = 128), что

соответствует 16 мс, где n - количество отсчетов в сегменте, Количество сегментов равно 256. Длительность сегмента не меньше длительности произношения фонем, но превышает максимально возможный период основного тона фонем. Энергия сегментов для каждого масштабного коэффициента a вычисляется по формуле:

$$E = \sum_{i=1}^{n} F(i).$$

Картину расположения фонем в слове или предложении можно установить, исследуя зависимость энергии сегментов вейвлет-спектра от масштабного коэффициента a. Для исследования используется *МНАТ*-вейвлет. Вейвлет-анализ речевого сигнала показывает, что гласные фонемы и фонемы н, м, л имеют максимальные энергии при средних значениях a. Энергия фонем н, м, л много меньше энергии гласных звуков речи, но значительно выше энергии шума. Фонемы к, т, п, д выделяются при больших значениях a. Перед фонемами к, т имеется пауза. Такая закономерность наблюдается при многократном повторении и не зависит от случайных факторов. Шипящие и свистящие фонемы при малых значениях масштабного коэффициента a имеют энергию $W(a,b)$, сравнимую с энергией гласных фонем. При средних значениях a они имеют энергию на уровне шума. Многомасштабный анализ, основанный на ВП, позволяет объединять слова в разные группы. В результате уменьшается время распознавания и увеличивается точность распознавания, так как базу данных слов можно разбить на подгруппы и представить в виде дерева поиска [1]; [2]; [3]; [4]; [5].

Зависимость энергии сегментов ВП $W(a,b)$ от значения масштабного коэффициента a для фонем п и а, представлена на рис. 1.

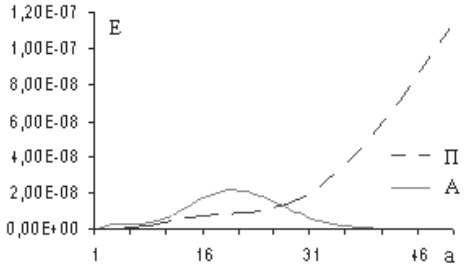


Рис. 1. Зависимость энергии сегментов ВП $W(a,b)$
от значения масштабного коэффициента a фонем п и а

Видно, что фонема п имеет во много раз большую энергию сегмента при большом масштабном коэффициенте, чем фонема а, и меньшую

энергию при среднем значении *a*. Для фонем т и к зависимость энергии сегментов ВП $W(a,b)$ от значения масштабного коэффициента *a* такая же, как и для фонемы п. Зависимость энергии сегментов ВП $W(a,b)$ от значения масштабного коэффициента *a* для фонем н, м, л отличается только тем, что эти фонемы всегда имеют энергию, меньшую, чем гласные фонемы. На рис. 2 показано это отличие.

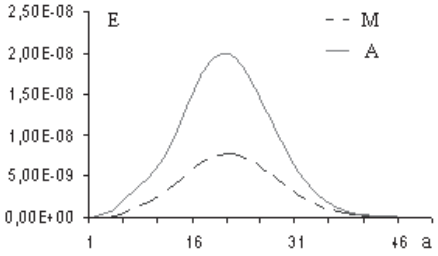


Рис. 2. Зависимость энергии сегментов ВП $W(a,b)$
от значения масштабного коэффициента *a* фонем м и а

На рис. 3 приведена зависимость энергии сегментов $W(a,b)$ от масштабного коэффициента *a* фонем а и ш.

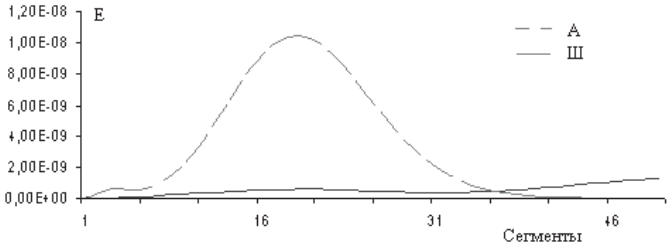


Рис. 3. Зависимость энергии сегментов ВП $W(a,b)$
от значения масштабного коэффициента *a* фонем ш и а

Свистящие фонемы, которые выделяются при малом масштабном коэффициенте наравне с гласными фонемами, имеют зависимость энергии сегментов ВП $W(a,b)$ от значения масштабного коэффициента *a* такую же, как фонема ш. Приведенные примеры показывают, что многомасштабное представление позволяет визуализировать динамику изменения речевого сигнала вдоль «оси масштабов». Эти изменения по «масштабной переменной» дают важную информацию о речевом сигнале.

Литература

1. Семенов В.И., Желтов П.В. Вейвлет-анализ акустического сигнала КГТУ им. А.Н. Туполева // Вестник КГТУ. 2008. Вып. 4. С. 210–212.
2. Семенов В.И., Желтов П.В. Распознавание речи на основе вейвлет-преобразования // Деп. в ВИНИТИ РАН 29.02.08, №174-В2008. Чебоксары: Чуваш. ун-т, 2008. 16 с.
3. Семенов В.И., Желтов П.В. Вейвлет-обработка речевых сигналов // Математические модели и их приложения: сб. науч. тр. Чебоксары: Изд-во Чуваш. ун-та, 2008. Вып. 10. С. 230–237.
4. Семенов В.И., Желтов П.В. Выделение границы между гласными и согласными фонемами при распознавании речи // Компьютерные технологии и моделирование: сб. науч. тр. Казань: КГТУ им. А.Н. Туполева, 2008. Вып. 1. С. 24–28.
5. Семенов В.И., Желтов П.В. Применение вейвлет-анализа сигнала в распознавании речи // Компьютерные технологии и моделирование: сб. науч. тр. Казань: КГТУ им. А.Н. Туполева, 2008. Вып. 2. С. 55–65.

УДК 004.855.5

КАРАКАЛПАКСКО-УЗБЕКСКИЙ ПЕРЕВОД ТЕКСТОВ НА ОСНОВЕ РЕКУРРЕНТНОЙ НЕЙРОННОЙ СЕТИ

А.А. Кадыров
Нукусский филиал Ташкентского университета информационных технологий
censor2005@mail.ru

В статье рассматривается вопрос применения нейронных сетей для разработки каракалпакско-узбекского переводчика текстов. Благодаря близкой родственности языков, данный подход может упростить обучение нейронной сети. Дается предварительная оценка сложности разработки.

Ключевые слова: *перевод, нейронная сеть, каракалпакский, узбекский.*

Текущий 2016й год по праву можно назвать годом нейросетевых алгоритмов. На фоне новостей о первых суперкомпьютерах на основе чипа TrueNorth, о победе AlphaGo над человеком, становится очевидным, что в будущем нейронные сети будут развиваться по закону Мура. И если раньше применение нейронных сетей для перевода текстов можно было считать неоправданно дорогим в плане затраченных процессорных и временных ресурсов, то скоро такой подход может стать повсеместным.

Нейронные сети (точнее, искусственные нейронные сети) – это одно из направлений исследований в области искусственного интеллекта, основанное на попытках воспроизвести нервную систему человека. А именно: способность нервной системы обучаться и исправлять ошибки, что должно позволить смоделировать, хотя и достаточно грубо, работу человеческого мозга [1].

К одному из сложных видов искусственных нейронных сетей относятся рекуррентные, в которых имеются обратные связи. В первых рекуррентных искусственных нейронных сетях главной идеей было обучение своему выходному сигналу на предыдущем шаге. Рекуррентные сети реализуют нелинейные модели, которые могут быть применены для оптимального управления процессами, изменяющимися во времени, то есть обратные связи позволяют обеспечить адаптивное запоминание прошлых временных событий. Обобщение рекуррентных нейронных сетей позволяет создать более гибкий инструмент для построения нелинейных моделей [2].

Рекуррентные нейронные сети могут быть обучены моделированию естественных языков. Данный факт порождает интересный вопрос: можно ли применить рекуррентную нейронную сеть для перевода текста? Как оказалось, да, это возможно [3].

Для автоматизации процесса работы с нейронными сетями корпорация Google опубликовала исходные коды библиотеки TensorFlow, предназначенной для решения задач с помощью нейронных сетей. В связи с этим возникла идея разработать переводчик текстов на основе нейронной сети.

Для обучения нейронной сети необходимо произвести ее обучение на параллельных корпусах. Так возникает необходимость собрать большую базу параллельных текстов. В качестве такой базы можно использовать базу данных законодательства Республики Каракалпакстан, которая как правило ведется на каракалпакском и узбекском языках, либо тексты народных эпосов каракалпаков.

В данной работе рассматривается план работ и дается общий анализ запланированных работ.

Процесс разработки планируется разбить на 3 этапа:

1. Сбор текстов на узбекском и каракалпакском языках.
2. Приведение текстов к единой форме, разбиение на фразы, выравнивание слов, подготовка данных к обучению.
3. Обучение рекуррентной нейронной сети.

Остановимся более подробно на каждом этапе. На первом этапе планируется собрать базу параллельных текстов, наиболее пригодных к обучению нейронной сети. Данную базу необходимо будет выровнять,

то есть соотнести каждое предложение исходного текста соответствующим предложениям конечного текста.

Второй этап является наиболее сложным и длительным, так как требует тщательной проверки текстов и выравнивания фраз. Так как искусственная нейронная сеть обучается на примере пар предложений, необходимо исходные тексты разбить правильным образом, чтобы размещение как минимум предложений в обоих языках было одинаковым, а в идеале – чтобы одинаковым было и размещение слов в этих предложениях на обоих языках. Хотя в принципе нейронная сеть довольно гибкая в плане размещения слов, тем не менее тот факт, что у нас небольшая обучающая база, повышает требования к ее качеству.

Третий этап является наиболее рутинным и уже вполне автоматизированным благодаря системе TensorFlow. На сайте tensorflow.org можно найти примеры обучения нейронной сети на базе англо-французского корпуса, результатом которого является достаточно качественный перевод текста [3].

Для обучения нейронной сети на тестовом примере с сайта tensorflow.org используется параллельный корпус состоящий примерно из 22 миллиона предложений. Используется подход последовательность-в-последовательность, обучение занимает довольно длительное время и требует не менее 20Гб свободного пространства на диске. При необходимости можно регулировать количество скрытых слоев и количество нейронов в каждом слое, а также периодичность сохранения данных, что может повлиять как на качество, так и на скорость обучения нейронной сети. Для ускорения обучения рекомендуется использовать мощности видеопроцессора, данная функция доступна в библиотеке TensorFlow. Один проход по базе требует примерно 340 тысяч шагов обучения [3]. Так что качественное обучение может потребовать дней и даже недель.

Для проверки качества обучения модели используется тестовая база. На сайте [3] приводится пример перевода фразы *«Who is the president of the United States?»*, результатом работы нейронной сети является фраза на французском *«Qui est le président des États-Unis?»*.

Данный опыт по разработке нейросетевой модели перевода с узбекского на каракалпакский может стать промежуточным этапом в задаче разработки каракалпакско-английского и каракалпакско-русского переводчика на основе нейронной сети, и любая информация, полученная в результате данного опыта окажется очень полезной в будущем.

Литература

1. Искусственный интеллект — Портал. URL: http://www.aiportal.ru/
2. Лила В.Б., Пучков Е.В. Методология обучения рекуррентной искусственной нейронной сети с динамической стековой памятью. Международный журнал "Программные продукты и системы", Тверь, №4, 2014 г.
3. TensorFlow -- an Open Source Software Library for Machine Intelligence. URL: https://www.tensorflow.org/

ЭЛЕКТРОННЫЙ КАТАЛОГ ВИРТУАЛЬНОГО МУЗЕЯ-БИБЛИОТЕКИ М.И. МАХМУТОВА: ПРЕДСТАВЛЕНИЕ ДОКУМЕНТОВ И ПОИСК

М.И. Курманбакиев, О.А. Невзорова,
Д.Ш. Сулейманов, Д.М. Шакирова
НИИ «Прикладная семиотика» АН РТ, Казань
write@marat.link, onevzoro@gmail.com,
dvdt.slt@gmail.com, shdilyara_m@mail.ru

В статье представлена модель организации электронного каталога Виртуального музея-библиотеки М.И. Махмутова, описана структура и представление научного наследия М.И. Махмутова, а также особенности организации поиска по электронному каталогу.

Ключевые слова: *виртуальный музей-библиотека, информационная система. электронный каталог, поиск информации, модель метаданных*

Введение

Многообразие научных школ, языков, национальных культур и конфессий является уникальным историко-культурным наследием Республики Татарстан и должно быть в полном объеме представлено в мировом информационном пространстве.

В области науки, образования и культуры национальные достижения складываются из персональных достижений талантливых личностей, живших и творивших в определенную историческую эпоху в конкретной социально-экономической и социально-культурной среде.

Проект научно-образовательного портала «Научные школы Академии наук Республики Татарстан» ставит целью создать Интернет-ресурс, ориентированный на представление объективных исторических, научных, культурологических характеристик общества через представление достижений и роли ученого в конкретный исторический период. Цели

проекта носят общекультурный, образовательный и просветительский характер и призваны привлечь внимание научной общественности и особенно молодежной аудитории к научному потенциалу АН РТ. Портал «Научные школы АН РТ будет содержать совокупность информационных ресурсов, представленных в виде информационных систем – виртуальных музеев-библиотек (ВМБ) ученых Академии наук РТ. Информационная система в виде виртуального музея-библиотеки будет включать архивные материалы - научные труды, статьи, переписки, творческие материалы, фото-видео материалы, воспоминания элиты нашего общества в разные исторические периоды - и современные информационные интерактивные ресурсы - форумы, видеоконференции, интерактивные модели, обучающие курсы.

Архитектура разрабатываемого портала представляет совокупность модулей, образующих единое информационное пространство, в основе которого лежит максимально полная база знаний о личности ученого и его времени. В состав портала входят следующие модули:

- модуль «Персоны», включающий биографическую информацию, информацию о научных достижениях, публикации, документальные материалы (воспоминания, официальные документы);
- модуль «Медиа-банк» (электронная коллекция фото и видео записей);
- модуль «Библиотека» (электронная коллекция публикаций).

Четкое разделение портала на модули дает возможность многоаспектной навигации по порталу в зависимости от целей пользователя. Модули портала взаимодействуют друг с другом. Взаимосвязи данных, представленных в различных модулях, отражены на рисунке 1. Данные о публикациях лиц, представленных в модуле «Персоны», публикациях с упоминаниями о них берутся из модуля «Библиотека», медиа — из модуля «Медиа-банк». Это становится возможным благодаря дополнительным единицам метаописаний, содержащим информацию об отношениях лиц к контенту, в соответствующих разделах. Такая схема взаимодействий позволяет автоматизировать процесс заполнения персональных кабинетов, а также наиболее удачно структурировать информацию в подразделах.

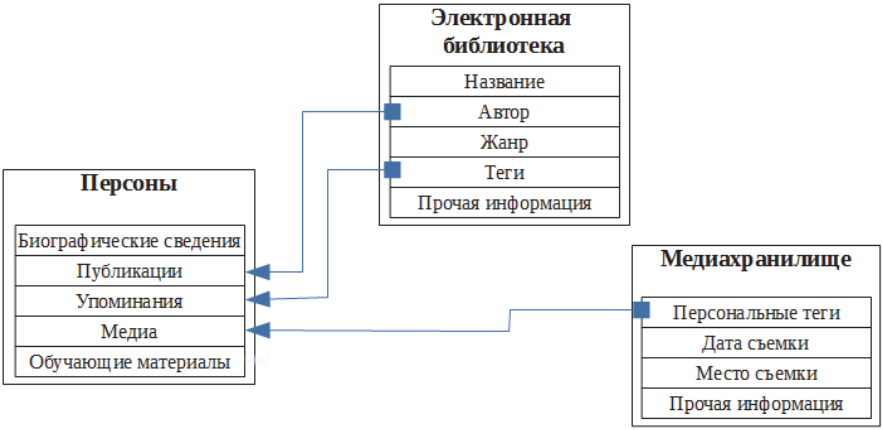


Рис. 1. Схема взаимодействия данных различных модулей

Можно отменить ряд близких по целям проектов в сети Интернет. Например, сайт посвященный академику РАН М.А. Лаврентьеву [1]. На этом сайте представлена биографическая информация, научная деятельность и достижения академика РАН М.А. Лаврентьева, собраны медиаматериалы, публикации и документальная литература. Сайт с библиографией академика РАН Н.Л. Добрецова [2] представляет из себя электронный каталог, в котором можно посмотреть библиографические записи трудов Н.Л. Добрецова отсортированные в хронологическом и в алфавитном порядке. Также имеется раздел со списком соавторов Добрецова. Доступ к тексту самих трудов с сайта не предоставляется. Сайт Мемориальной библиотеки В.А. Коптюга[3], несмотря на то что является сайтом физической библиотеки, имеет несколько публикаций представленных в электронном виде. Публикации представлены в виде html страниц.

Проект «Виртуальный музей-библиотека М.И. Махмутова» разрабатывается с 2011 года[4]. В настоящее время реализуется новая версия проекта, в которой существенно переработана программно-техническая часть. Проект обрел модульную архитектуру, в которой каждый из модулей выполняет определенную функцию, что позволяет реализовать более удобную навигацию по порталу и новый функционал, а также более функциональную систему администрирования проекта.

Электронный каталог ВМБ М.И. Махмутова

Одним из ключевых модулей проекта является модуль «Библиотека», включающий электронный каталог публикаций, в котором содержатся

развернутые метаописания документов, а также их исходные тексты. На рис. 2 представлена общая архитектура модуля «Библиотека».

Ядром электронного каталога является база данных, содержащая библиографические и семантические метаописания документов. Подсистема поиска выполняет поиск по данным метаописаний. Интерфейс навигации выводит список публикаций и их основных метаданных, на основе результатов поиска выполняется переход к карточке публикации, которая отображает как полный набор метаданных, так и идентификатор ресурса URI для доступа к документу.

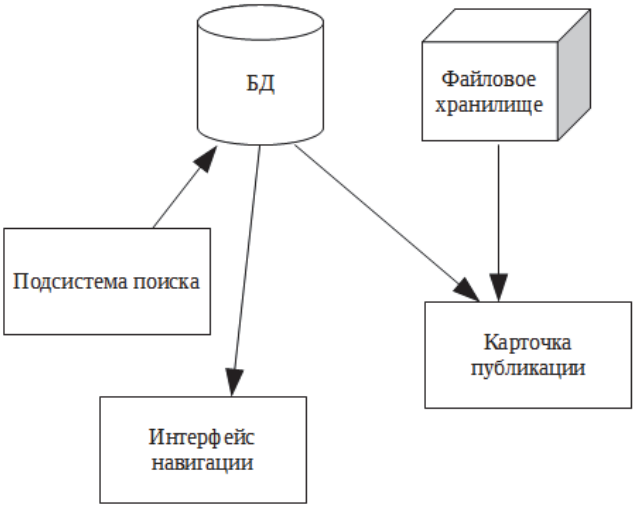


Рис. 2. Общая архитектура электронного каталога публикаций

Каждая единица контента в каталоге должна иметь необходимый набор метаописаний.

Основным критерием при выборе набора метаописаний является тип документа (тип публикации). Нами выделены следующие типы документов:

- Книга (учебник, монография, словарь и др.)
- Статья (подразделяется на статья в журнале, статья в сборнике)

В соответствии с типами документов определены базовые наборы метаданных.

Базовый набор метаданных для книг включает библиографические параметры (Название, Автор, Год издания, Издательство, Город издания, Тип издания, Номер редакции, Редактор, Язык документа, Переводчик на представленный язык), а также набор семантических

параметров (ссылки на Персоны, Общая тематика, Расширенная тематика, Ключевые слова, Список терминов).

Базовый набор метаданных для словарей включает библиографические параметры (Название, Автор, Язык (Языки для многоязычных словарей), Год издания, Издательство, Город издания, Номер редакции, Редактор).

Базовый набор метаданных для статей из сборников включает набор библиографических и семантических параметров, аналогично описанию книг.

Поскольку статьи являются частью сборников, сами сборники также должны быть описаны:

Сборники конференций:

- Название сборника
- Название конференции
- Дата проведения конференции
- Место проведения конференции

Научные журналы:

- Название журнала
- Номер журнала
- Год выхода журнала

Значения поля «список терминов» генерируются автоматически из содержимого документа на основе словарей профессиональных терминов, например для генерации списка терминов материалов педагогической направленности планируется использовать ряд словарей, например, словарь А.М. Новикова «Педагогика: словарь системы основных понятий», «Толковый словарь терминов понятийного аппарата информатизации образования» И.В.Роберта и Т.А.Лавиной, «Словарь основных терминов и понятий в области качества образования» и др.

Структура контента ВМБ М.И. Махмутова

Одной из самых сложных задач создания исторически достоверного, увлекательного, научно-обоснованного и полезного для людей разного возраста, типа, профессии виртуального музея является отбор информации. При построении концепции контента виртуального музея нами используются подходы, разработанные историками, библиографами, писателями для мемуарной литературы. Так, постсоветская мемуаристика, особенно ее «популярная» часть, впервые предложила, в отличие от обезличенной «правды» документа, личную историю, индивидуальный факт, биографию конкретных людей, живших в истории. Story на данном этапе главенствует над History, на смену универсальному опыту приходит опыт индивидуальный: история отдельной человеческой жизни сочетается

с историей науки и общества. По-видимому, происходит некая приватизация истории, при которой мемуары, дневники, автобиографии, подборки научных книг, статей, фотографий, отзывов и воспоминаний современников предлагают читателю более достоверную версию прошлого, чем исторический документ.

С этих позиций особый интерес представляют появляющиеся в последние годы виртуальные музеи университетов с отдельными сайтами ученых, которые интегрируют музейную и библиотечную составляющие.

Коллекция публикаций Мирзы Исмаиловича Махмутова, имеющаяся в распоряжении разработчиков проекта составляет 550 документов. Подготовка документа коллекции включает оцифровку, распознавание и редактирование, создание метаописания документа. Последний процесс является весьма трудоемким, поскольку формирование семантического блока параметров требует специальной библиографической работы. В ближайших планах проекта – автоматизация заполнения значений библиографических и семантических данных метаописаний документов.

Публикации разделены по следующим тематикам:

- Педагогика
- Национальное образование
- Словари
- Страноведение
- Языкознание
- Религия
- Социальные проблемы общества

Поисковые решения

Важным компонентом электронного каталога является подсистема поиска. Для поиска документов в стадии разработки находятся различные поисковые решения, различающиеся технической сложностью и использованием семантических данных.

Первый вариант поиска - «Быстрый поиск», который использует одно поле для ввода поискового запроса. Данный тип поиска ищет вхождения поискового запроса в следующие единицы метаданных: Автор; Название; Ключевые слова; Тематика; Ссылки на Персоны.

Расширенный поиск предполагает более конкретизированный поиск по описаниям метаданных. Расширенный поиск использует следующие поля:

Первый поисковый набор: Название; Автор; Год; Тип материала (список для выбора).

Как можно заметить, в первый поисковый набор входят параметры метаописаний общие для всех типов документов. Во второй набор поисковых полей входят все типы метаописаний, в зависимости от выбранного поля «Тип» в первом наборе.

Для использования расширенного поиска необходимо заполнение как минимум одного поля. Для исключения ошибок в поисковом запросе в расширенном поиске реализована система подсказок, которая выдает варианты, имеющиеся в базе данных по мере ввода пользователем поискового запроса. Подсказки формируются на основе существующих в базе данных метаописаний.

Заключение

Портал «Научные школы АН РТ» представляет собой системное многоуровневое объединение различных научных и образовательных ресурсов и сервисов. Цели проекта с одной стороны носят образовательный и просветительский характер, а с другой - имеют актуализирующую направленность, призванную привлечь внимание к научному потенциалу АН РТ.

В статье рассмотрен подход к начальной реализации электронного каталога печатного контента модуля «Библиотека» портала «Научные школы РТ». Электронный каталог позволяет осуществлять удобную навигацию и поиск по публикациям. В дальнейшем планируется улучшать поисковые возможности данного инструментария, в том числе по данным полученным в результате семантического анализа текста.

В перспективе планируется разработка инструментария позволяющего работать с содержимым публикаций непосредственно на портале. Для этого будет разработан программный инструмент для чтения, позволяющий организовать удобное чтение документов через браузер. Инструмент будет включать в себя такой функционал как: изменение кегля, размера и типа шрифта, изменение цветовой гаммы текста, создание примечаний к тексту и закладок.

Благодарности. Работа выполнена при частичной финансовой поддержке РФФИ (проект № 15-47-02472).

Литература

1. Академик Михаил Алексеевич Лаврентьев. [Электронный ресурс]. – Режим доступа: http://www.sbras.ru/win/sbras/acad/lavren.html – Заглавие с экрана. – (Дата обращения: 01.04.2016).

2. Николай Леонтьевич Добрецов - Библиография научных трудов. [Электронный ресурс]. – Режим доступа: http://www.sbras.ru/ppls/dbr/bibl/intro.htm – Заглавие с экрана. – (Дата обращения: 01.04.2016).

3. Мемориальная библиотека В.А.Коптюга. [Электронный ресурс]. – Режим доступа: http://www.prometeus.nsc.ru/koptyug/library/ – Заглавие с экрана. – (Дата обращения: 01.04.2016).

4. Виртуальный музей-библиотека М.И. Махмутова. [Электронный ресурс]. – Режим доступа: http://vml.antat.ru/ – Заглавие с экрана. – (Дата обращения: 01.04.2016).

5. Д.М. Шакирова, Д.Ш. Сулейманов, О.А. Невзорова, М.И. Курманбакиев. Личность и эпоха: виртуальное представление научного наследия // Научный сервис в сети Интернет: труды XVII Всероссийской научной конференции (21-26 сентября 2015 г., г. Новороссийск). -- М.: ИПМ им. М.В.Келдыша, 2015. – С. 321-328

6. Сулейманов Д. Ш., Шакирова Д. М., Гильмуллин Р. А. Виртуальный музей-библиотека “Научные школы АН РТ” как образовательная интернет-среда // Образовательные технологии и общество 2013.— № 3. - С. 655–663.

7. Халилбеков Т. Р. Каталогизация электронных ресурсов // Известия ЮФУ. Технические науки. 2003.— № 1. - С. 70–71.

8. Толстиков А.Н., Павленко Е.П., Толстиков Н. Г. Выбор метода представления библиографической информации // ВЕЖПТ ..2010.— № 2(46). - С. 69–71.

УДК: 004.8

КОРРЕКЦИЯ ПРАВОПИСАНИЯ С ПОДДЕРЖКОЙ ЧАНКИНГА В МОДЕЛИ ДЕРЕВЬЕВ ЗАВИСИМОСТЕЙ В РУССКОМ И АНГЛИЙСКОМ ЯЗЫКАХ

И.С. Анисимов
ООО «Яндекс», Москва, Россия
ivananisimov2010@gmail.com
Е.А. Макарова
ИЯз РАН, Москва, Россия
antaresselen@mail.ru
В.Н. Поляков
ИЯз РАН, НИТУ «МИСиС», Москва, Россия
pvn-65@mail.ru

Статья описывает способ коррекции правописания с использованием чанкинга в модели деревьев зависимостей. С помощью данного метода можно устранить возможные альтернативные генерации коррекций ошибочных слов в соответствие с морфологическим словарем. Метод основан на генерации коррекций с помощью метода Левештейна, построении графа всех возможных чанков и в последующем преобразования графа в множество деревьев. Затем выбирается дерево

чанков, состоящее из наибольшего количества слов. Метод может быть применен в пакетном режиме.

Ключевые слова: *коррекция правописания, чанкинг, синтаксис, модель зависимостей, русский, английский.*

Введение

Коррекция правописания является одной из самых известных задач обработки естественно-языкового текста. Для русского языка она, как правило, базируется на известном словаре А.А. Зализняка [19]. Со временем, в связи с ростом потребностей поисковых систем, в связи расширением сферы применения корректоров правописания в текстовых редакторах, в смс-сообщениях, в твиттере, блогах и на интернет-форумах, электронный словарь А.А. Зализняка был дополнен научно-техническими словарями, словарями имен собственных, словарем географических названий, словарем медико-биологических и химических терминов.

Но часть проблем коррекции правописания так и осталась нерешенной.

В первую очередь это касается:

- неграмматичностей, которые человеческий мозг выделяет довольно легко: неологизмы, имена собственные (фамилии, имена, отчества, названия компаний, продуктов), но которые представляют очевидную сложность для корректора – автомата;
- синтез нескольких вариантов коррекции, когда корректор правописания, программа, затрудняется принять решение в пакетном режиме и необходимо вмешательство человека;
- многоязычный режим, когда слово или фраза пишется на иностранном языке (при сохранении базового алфавита), и корректору правописания необходимы мета-знания для идентификации этой ситуации.

В настоящем исследовании делается попытка построения корректора правописания с поддержкой с помощью синтаксической модели. Это даст в потенциале возможность принимать решение в пакетном режиме, если программа-корректор сгенерировала несколько вариантов коррекции. Синтаксическая модель обеспечивает необходимый в таких ситуациях контекст, которым, кстати, чаще всего пользуется и человек-редактор.

О синтаксисе

Наиболее распространенными являются две модели синтаксиса, модель Хомского [5, 6] и модель Теньера [14]. В настоящей работе

синтаксическая модель базируется на модели Теньера, то есть, на модели зависимостей. Она получила рабочее название «чанкинг в модели зависимостей».[1]

В синтакисе чанкинга описание выполнено максимально скупыми средствами, в отличие, например от синтаксиса Мельчука [2, 13], где описание приведено максимально полно и подробно для каждой синтаксической ситуации. Краткость описания позволяет перейти сразу же к испытанию модели на практике. Синтаксис чанкинга, описанный в настоящей работе, базируется преимущественно на исследовании [3, 4], при этом учтен многолетний опыт его эксплуатации и в модель чанкинга введены новые эвристики, которые призваны улучшить ее показатели.

Описание алгоритма

Программа корректора правописания использует в своем составе библиотеку NLP@CLOUD и написана на языке Java в среде UIMA. Оговоримся сразу, что все указанные средства были выбраны не случайно. Несмотря на тот факт, что язык Java медленнее, чем язык C/C++, допустим; сама библиотека UIMA работает медленнее, чем [17, 18]; возможности переноса на любую платформу и возможности масштабирования (например, за счет применения кластерной архитектуры) в UIMA и Java гораздо шире, а значит, они являются более перспективными в будущих приложениях.

На вход алгоритма подается текст на русском языке. Обработка входа осуществляется в несколько этапов, в том числе, предусмотрены этапы:

- предобработка
- токенизация
- сегментация по предложениям
- морфонализ
- синтез потенциальных коррекций
- построение расширенного грамматического вектора слов
- поиск главных членов предложения
- сегментация по клаузам (не реализовано в текущей версии)
- формирование множества потенциальных чанков
- построение множества деревьев чанков
- выбор наилучшего дерева чанков
- вывод результатов

На выходе алгоритма получаем исправленный текст и набор аннотаций разметки в формате XML.

Алгоритм работает на базе библиотеки Apache UIMA (https://uima.apache.org/). Формат аннотаций описан в: http://uima.

[1] Слово «чанк» образовано от английского chunk, что значит *ломоть, глыба, кусок, фрагмент*.

apache.org/downloads/releaseDocs/2.2.2-incubating/docs/html/references/references.html#ugr.ref.xmi.

Чанкинг для английского находится в стадии проектирования, поэтому представлен эскизно.

В большинстве представленных на рынке примерах и в публикациях [1, 16] чанкинг для английского строится на следующих принципах:

- он базируется на грамматиках Хомского [5];
- в качестве чанков выбраны непосредственно составляющие в грамматиках Хомского.

В предлагаемой модели чанкинга для английского применены следующие

новшества:

- чанкинг базируется на грамматике деревьев зависимостей (модель Теньера [14]);
- в качестве чанков используются фрагменты дерева зависимостей, максимально приближенные к логическим предикатом;
- кроме того, применены эвристики, позволяющие построить двухчастные чанки.

Заключение

Библиотека NLP@CLOUD носит универсальный. Часть допущенных ошибок в русской коллекции можно устранить, повысив тем самым показатели Re и F1. В реальной выборке случаев поддержки синтаксическим контекстом правописания оказывается больше, чем это было представлено в данной выборке. Это форумы и блоги в интернете, твиты, смс – сообщения со смартфонов. Таким образом будет повышен показатель Pr.

Кроме того, библиотека NLP@CLOUD обладает рядом выгодных преимуществ, это высокая переносимость и масштабируемость библиотеки за счет языка Java, а следовательно, она имеет свою нишу.

Наша цель выхода на соревнование – получение максимального количества критических замечаний и их последовательное устранение.

Благодарности. Исследование поддержано грантом РНФ № 15-11-10019

Литература

1. Abney Steven (1991), Parsing By Chunks, Kluwer Academic Publishers, pp. 257–278. Available at: http://www.vinartus.net/spa/90e.pdf (Accessed 14 March, 2016).

2. Boguslavsky I., Iomdin L., Tsinman L., Sizov V., Petrochenkov V. (2011). Rule-Based Dependency Parser Refined by Empirical and Corpus Statistics // Proceedings of the International Conference on Dependency Linguistics (Depling'2011). Barcelona, September 5-7, 2011. P. 318-327. ISBN 978-84-615-1834-0.

3. Bushtedt V. (2011). Model of Decision Making based on Syntactic Analysis in Tasks of Patent Information Procession [Model prinyatiya resheniya na osnove sintaksicheskogo analiza v zadachah obrabotki patentnoi informacii]. PhD paper. Specialty 05.13.01. "System analysis, information management and processing (in production)", defended in 2011.

4. Bushtedt V., Polyakov V. (2009). Heuristics for Improvement of Partial Syntactic Analyzer Work [Evristiki dlya uluchsheniya raboty chastichnogo sintaksicheskogo analizatora]. Scientific Notes of Kazan State University [Uchenie zapiski Kazanskogo Gosudarstvennogo Universiteta]. V. 151, book 3, pp. 214-228.

5. Chomsky N. (1956). Three Models for Description of Language. // IRE Trans. Informat. Theory, 1956, v. IT-2, p. 113-124.

6. Chomsky N. (1957). Syntactic Structures. — The Hague: Mouton, 1957. (Reprint: Chomsky N. Syntactic Structures. — De Gruyter Mouton, 2002. — ISBN 3-11-017279-8).

7. Gladky A. & I. Melchuk. Elements of mathematical linguistics [Elementy matematicheskoy lingvistiki]. Moscow: Nauka Publ., 192 p., 1969.

8. Gladky, A. Sintactic structures of natural language in computer-based communication systems [Sintaksicheskie struktury estestvennogo yazyka v avtomatizirovannyh sistemah obscheniya]. Moscow: Nauka Publ., 144 p., 1985.

9. Melchuk I. & N. Pertsov. Surface syntax of English: A formal model within the Meaning-Text framework. Amsterdam; Philadelphia: Benjamins, 1987. ISBN 90-272-1515-4

10. Melchuk I. (1987). Dependency syntax : theory and practice. Albany: State University Press of New York. ISBN 978-0-88706-450-0

11. Melchuk, I. (2003). Levels of dependency in linguistic description: Concepts and problems. In Ágel et al., 170–187.

12. Noy N, Crubezy M, Fergerson RW, Knublauch H, Tu S, Vendetti J, Musen M. Protege-2000: an open-source ontology-development and knowledge-acquisition environment. AMIA Annu. Symp. Proc. 2003;2003:953.

13. Sannikov V., Boguslavky I., Iomdin L., Apresyan Yu. (2010). Theoretic Problems of Russian Syntax [Teoreticheskie problemy russkogo sintaksisa]. Languages of Slavic Cultures [Yazyki Slavyanskih Kultur]. 408 pp.

14. Tesniere, L. (1959). Elements of Structural Syntax (Éléments de syntaxe structural), Klincksieck, Paris. Préface by Jean Fourquet, professeur à la Sorbonne. Second edition, reviewed and corrected. ISBN 2-252-02620-0. Re-edition of: Tesniere, L. (1959). Éléments de syntaxe structurale, Klincksieck, Paris. ISBN 2-252-01861-5

15. Tesniere, L. (1988). Dependency Syntax : Theory and Practice, Albany, N.Y.: SUNY Press, 1988. 428 pp.

16. Punyakanok V. and D. Roth, The Use of Classifiers in Sequential InferenceNIPS (2001) pp.995—1001. Available at:
http://cogcomp.cs.illinois.edu/papers/nips01.pdf (Accessed 14 March, 2016).

17. Vydrin D. Programs for morphological analysis. http://macrocosm.narod.ru/madown.html (Accessed on February, 29, 2016).

18. Vydrin D., Polyakov V. (2002). Realization of Electronic Dictionary Based on N-Grams [Realizaciya electronnogo slovarya na osnove n-gramm]. // Proceedings of III Inter-

national Scientific-Practical Conference “Artificial Intelligence – 2002” [Iskustvenny Intellect]. Publishing house “Institute of Problems of Artificial Intelligence” [Institut problem iskustvennogo intellecta]. Kacevelli, vol. 2, pp. 79-84.

19. Zaliznyak A.A (1977). Grammar Dictionary of Russian. Inflection. [Grammatichesky slovar russkogo yazyka. Slovoismenenie]. – Moscow, Russky Yazyk.

УДК 510.5, 519.768.2

РАЗРАБОТКА СЕМАНТИЧЕСКОЙ МОДЕЛИ СИСТЕМЫ МАШИННОГО ПЕРЕВОДА ДЛЯ РУССКО-КАЗАХСКОЙ ЯЗЫКОВОЙ ПАРЫ

Д.Р. Рахимова
Казахский Национальный Университет имени аль Фараби, Алматы, Казахстан
diana.rakhimova@kaznu.kz

В статье описывается семантическая модель системы машинного перевода для русско-казахской языковой пары на основе предложенной расширенной атрибутной грамматики. Данный метод был разработан с учетом специфики двух различных языков и их сопоставления на примере простых предложений.

Ключевые слова: *семантика, машинный перевод, русский и казахский язык.*

Процесс машинного перевода можно описать как: декодирование смысла исходного текста и перекодирование этого смысла на целевом языке. За этой кажущейся простой процедурой скрывается сложная когнитивная деятельность. Для декодирования смысла исходного текста в целом, переводчик должен интерпретировать и проанализировать все особенности текста - процесс, который требует глубокого знания грамматики, семантики, синтаксиса, идиом и т.д. исходного языка, а также культуры его носителей. Переводчику необходимы такие же углубленные знания в целевом языке для перекодировки смысла [1, с.10-11].

Анализ различных подходов в области показывает, что наиболее качественный машинный перевод получается с использованием подхода, основанного на правилах с расширением семантических свойств языка. Ниже представлена архитектура модели машинного перевода с семантическим анализом.

Концептуальная модель машинного перевода простых предложений для русско-казахской языковой пары представлена из следующих модулей [2, с.8-16]:

Модуль лексикографического анализа (ЛГА) текста. В данном модуле введенный текст разбивается на предложения и лексемы (слова) предложения. Учитывается наличия и расположения знаков препинания в предложений.

Модуль морфосемантического анализа слов. В модуле будут произведены морфологический и семантический анализ слов. Определение основы и окончаний слова и их свойств.

Модуль семантико-синтаксического анализа (ССА) текста. На основе результатов предыдущих модулей производится синтаксический разбор простых предложений с учетом его семантики и синтетической природы ЕЯ. Выясняется, какую роль играет каждое слово в предложении, строится семантико-синтаксическое дерево разбора входного текста. Для построения семантико-синтаксического дерево разбора необходимо определить смысловые словосочетания (фразы), которые можно определить за счет семантических атрибутов и семантических правил на уровне фраз.

Модуль построения онтологии**.** После анализа предложения входного языка на метаязыке строится схемы структуры предложения на основе онтологической модели предложения с семантическими атрибутами и отношениями. Далее необходимо произвести оценку и уничтожение несовместимых и\или дублирующих отношений из множества подграфов семантических словосочетании. Подбирается соответствующая ей схема предложения на целевой язык. Обработка многозначных слов производится с учетом контекста их употребления.

Модули семантической, синтаксической и лексической генерации преобразовывает входной текст с учетом свойств и правил каждого модуля на выходной язык.

Рис. 1. Схема работы семантического анализа
и генератора системы машинного перевода

В связи с вышеизложенным в данной работе для разработки семантической модели и системы машинного перевода русско-казахской пары было решено выбрать подход основанного на правилах с расширением семантических свойств языка. Для семантического анализа и синтеза структуры предложения будет использован предложенный метод Расширенной атрибутной грамматики (РАГ), основанный на атрибутной грамматике Кнута [3, с. 127-143]. Расширенная атрибутная грамматика для описания предложений естественных языков представляется в следующем виде [4]:

$$AAG = <G, A, R^W, R^F, R^S>, \quad (1)$$

где G является контекстно-свободная грамматика для естественного языкаа;

A- конечное множество семантических атрибутов для нашей грамматики;

R^W–множество семантических правил на уровне слов;

R^F –множество семантических правил на уровне фраз (словосочетании);

R^S –множество семантических правил на уровне предложения.

$$G = < V, N, P, S >, \quad (2)$$

Итак, контекстно-свободная грамматика G определяется следующими характеристиками:

V — набор (алфавит) терминальных символов;

N — набор (алфавит) нетерминальных символов;

P — набор правил вида: "левая часть" "правая часть", где:

"левая часть" — непустая последовательность терминалов и нетерминалов, содержащая хотя бы один нетерминал;

"правая часть" — любая последовательность терминалов и нетерминалов;

S — стартовый (начальный) символ из набора нетерминалов.

Большой интерес к атрибутной грамматике вызван не тем, что она представляет собой идеальное определение двоичной системы записи, а тем, что она демонстрирует взаимодействие унаследованных и синтезированных атрибутов. Семантические правила подобные правилам не приводят к зацикленности определения атрибутов. Важность унаследованных атрибутов состоит в том, что они естественно возникают в практике и в очевидном смысле "двойственны" синтезированным атрибутам. Хотя для определения смысла двоичной записи достаточно только синтезированных атрибутов, существует ряд языков, для которых такое ограничение приводит к неуклюжему и неестественному определению семантики [5].

Описание атрибутной грамматики состоит из раздела описания атрибутов и раздела правил. Раздел описания атрибутов определяет состав

атрибутов для каждого символа грамматики и тип каждого атрибута. Правила состоят из синтаксической и семантической части. В синтаксической части используется расширенная БНФ. Семантическая часть правила состоит из семантических атрибутов и семантических отношений [6].

Семантические правила дополняют КС грамматику G следующим образом. С каждым символом $X \in V$ связывается конечное множество атрибутов A(X). A(X) разбивается на два непересекающихся множества: множество синтезированных атрибутов $A_0(X)$ и множество унаследованных атрибутов $A_1(X)$. Множество $A_1(S)$ должно быть пустым (то есть начальный символ S не должен иметь унаследованных атрибутов); аналогично, множество $A_0(X)$ пусто, если X - терминальный символ. Каждый атрибут a из множества A(X) имеет множество значений H. Для каждого вхождения X в дерево вывода семантические правила позволяют определить одно значение из множества H для соответствующего атрибута [6].

Пусть P состоит из m правил, и пусть p-е правило имеет вид

$$X_{p0} \rightarrow X_{p1}X_{p2}...X_{pnp}, \qquad (3)$$

где $n_p > 0$, $X_{p0} \in N$ и $X_{pj} \in V$ для $1 \leq j \leq n_p$. Семантическими правилами называются функции $f_{pj} \in R$, определенные для всех $1 \leq p \leq m$, $0 \leq j \leq n_p$ и некоторых $\alpha \in A_0(X_{pj})$, если $j = 0$, или $\alpha \in A_1(X_{pj})$, если $j > 0$. Где R состоит из $\{R^w, R^f, R^s\}$ множеств семантических правил на различных уровнях семантического анализа. Каждая такая функция представляет собой отображение из $V\alpha_1 \times V\alpha_2 \times ... \times V\alpha_t$ в H для некоторого $t = t(p, j, \alpha) > 0$, где все $\alpha_i = \alpha_i(p, j, \alpha)$ являются атрибутами некоторых X_{pki}, при $0 \leq k_i = k_i(p, j, \alpha) \leq n_p$, $1 \leq i \leq t$. Другими словами, каждое семантическое правило отображает значения некоторых атрибутов символов X_{p0}, X_{p1}, ..., X_{pnp} и значение некоторого атрибута символа X_{pj} [7].

На практике была реализована система машинного перевода с словарем 13000 слов единиц.

Для оценки качества машинного перевода использована методика BLEU. Проведен сравнительный анализ качества машинного первода известных он-лайн переводчиков (как Sanasoft, Pragma6, Audaru (Soylem)) и предложенной программой. Для эксперимента использовались словосочетания и простые предложения. Ниже в таблице показаны практические результаты.

Таблица 1

Результаты оценки качества машинного перевода по методике BLEU

Наименование переводчика	Униграмма %	Биграмма %	Триграмма %	Итого %
Pragma	38,22727	15,11905	6,060606	19,802308
Audaru (Soylem)	25,45455	4,090909	0	9,848486
Sanasoft	53,68182	26,78572	11,74242	30,736653
Разработанная программа	51,13636	24,36147	14,54545	30,014426

По результатам оценки качества переводчики показали различные результаты. Можно отметить , что наиболее лучший результат показал перевод компаний Sanasoft (30,736653 %) и результаты реализованной программы на основе разработанных моделей и алгоритмов имеет хорошие результаты (30,014426%). Можно предположить, если пополнив словарную базу и правил анализа и синтеза текста можно добиться более высоких результатов при оценке качества машинного перевода. Необходимо отметить, что задачей данного анализа не является выявление ошибок или недостатков того или иной системы МП. Автор хотел показать эффективность предлагаемого метода в реализации машинного перевода, которые имеет сравнительно не плохие практические результаты при переводе для русско-казахской языковой пары.

Заключение

К настоящему времени известно, что в существующих методах и технологиях систем машинного перевода семантический анализ обычно следует после синтаксического. А в предложенном методе РАГ семантический анализ будет производиться поэтапно на каждом из уровней анализа предложения МП, благодаря которому можно добиться лучшего качества перевода. Выше был описан метод семантического анализа предложения на основе РАГ для КС грамматике G, с помощью которой были разрешены следующие задачи:

– Определение множества атрибутов предложения входного языка;

– Определение и классификация семантических атрибутов из множества атрибутов входного предложения;

– Вычисление семантических правил на основе семантических атрибутов на различных этапах анализа текста (на уровне слов, на уровне фраз, на уровне предложения);

– Систематизация множества семантических правил;
– Приведение в единое смысловое выходное значение предложения;
– Построение системы знаний (онтологии) предложения.

Разработанный метод РАГ представляет собой упрощенную модель семантического анализа предложений русского и казахского языка. Разработанная упрощенная модель семантического анализа казахского и русского языков позволяет разработать эффективные алгоритмы преобразования грамматических и семантических характеристик входного языка в грамматические и семантические характеристики целевого языка на различных этапах анализа (на уровне слов, фраз и предложения). Разработанные семантические атрибуты и семантические отношения в методе РАГ является достаточными для определения смыслового значения предложения и удобны в реализации алгоритмов семантического анализа и синтеза предложения при переводе с одного языка на другой язык [8].

Литература

1. Разработка эффективных технологии компьютерного перевода казахского языка на английский и русский языки (и обратно) на основе методов формальных грамматик и статистических методов: отчет о НИР (промежуточный) / ДГПНИИ ММ при КазНУ им аль-Фараби: рук. Тукеев У.А. Алматы, 2012.- 84 с.- № ГР 0112РК01467
2. Разработка эффективных технологии компьютерного перевода казахского языка на английский и русский языки (и обратно) на основе методов формальных грамматик и статистических методов: отчет о НИР (промежуточный) / ДГПНИИ ММ при КазНУ им аль-Фараби: рук. Тукеев У.А. Алматы, 2013.- 78 с.- № ГР 0112РК01467.
3. Knuth D.E. Semantics of Context-free Languages // Mathematical Systems Theory. - 1968.- Vol.2(2).- P.127-145.
4. Tukeyev U., Rakhimova D.R. Augmented attribute grammar in meaning of natural languages sentences // SCIS-ISIS 2012 The 6th International Conference on Soft Computing and Intelligent Systems. The 13th International Symposium on Advanced Intelligent Systems. - Japan: Kobe, 2012. - P.1080-1084.
5. Neven, F. Attribute grammars for unranked trees as a query language for structured documents// Journal of Computer and System Sciences.-2005.-№70. -P.221-257.
6. Серебряков В.А., Галочкин М.П. Атрибутные грамматики [Электрон.ресурс].- http://citforum.ck.ua/programming/theory /serebryakov /5.
7. Методы распознавания образов при идентификации объектов бинарного класса в автоматизированных телекоммуникационных комплексах систем управления. [Электрон.ресурс].- URL: http://www.bibliofond.ru/view.aspx? id=587496
8. Рахимова Д.Р. Построение семантических отношении в машинном переводе // Вестник КазНУ им. аль- Фараби. Серия "Математика , механика и информатика". - Алматы, 2014.- №1. - С. 90-101.

УДК 81`33

О НЕКОТОРЫХ ПОДХОДАХ К РЕШЕНИЮ ЗАДАЧИ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Б.П. Тажев
Институт информатики
и проблем регионального управления КБНЦ РАН
boristazhevart@mail.ru,

И.А. Гуртуева
Институт информатики
и проблем регионального управления КБНЦ РАН
gurtueva-i@yandex.ru

В данной работе приводится краткий обзор существующих подходов к решению актуальных проблем автоматического распознавания речи. Рассмотрены акустико-фонетический подход, распознавание паттернов и подходы с использованием искусственного интеллекта и искусственных нейронных сетей.

Ключевые слова: *распознавание речи, скрытые марковские модели.*

Наиболее эффективным и удобным способом коммуникации людей является речь. Взаимодействие человека с компьютером посредством речи было бы более предпочтительно, чем использование специальных устройств (интерфейсов, клавиатур и указательных устройств). Это было бы возможно с разработкой системы автоматического распознавания речи, целью которой является конвертация речевого высказывания в письменный текст. Математически задача распознавания речи может быть описана как поиск функции, определяющей отображение акустической формы в слово или последовательность слов [9]. Эта задача крайне сложна, поскольку речевое высказывание содержит в себе огромный объем лингвистической информации [8]. Кроме того, речь несет массу информации о говорящем (пол, возраст, социальное и географическое происхождение, эмоциональное состояние и состояние здоровья) [12]. Технологии распознавания речи находят широкое применение в самых разных сферах. Так, разработка системы автоматического распознавания речи даст импульс развитию устройств для естественно-языкового управления мультиагентными роботами. Речевые технологии являются ядром приложений, разрабатываемых для истребительной авиации, а именно, приложения для установки радиочастот, управления автопи-

лотом, установки рулевой координаты точки и параметры выпуска оружия и управления отображением полета). Разработки систем распознавания и понимания речи необходимы для создания голосового интерфейса для управления системами «умный дом», голосовых ключей, голосовых навигаторов для управления программным и аппаратным обеспечением, программ для диктовки – ввод текста и цифровых данных. Разработка и практическая реализация систем распознавания речи расширит возможности исследования актуальных проблем лингвистики.

В течение последних шести десятилетий сформировались три основных подхода к решению задачи распознавания речи: акустический, распознавание паттернов, а также подход с использованием искусственных нейронных сетей. В настоящее время ведутся активные разработки по созданию систем распознавания речи на базе искусственного интеллекта [1, 2]. Наиболее ранним подходом является акустический подход [2, 5]. Он исходит из предположения о том, что в разговорном языке существует конечный набор характерных фонетических единиц (фонем), которые характеризуются набором акустических свойств. И, несмотря на то, что, акустические свойства фонетических единиц, будучи сильно зависимыми от говорящего и соседних звуков (так называемый эффект коартикуляции), высоко вариативны, данный подход предполагает, что правила, определяющие вариативность, однозначны и хорошо формализуемы. Первым шагом в данном подходе являются спектральный анализ речи и детектирование, целью которых является конвертация спектральных измерений в набор характеристик, описывающих акустические свойства фонем в широких пределах. Следующий шаг – сегментирование высказывания на стабильные акустические области и фонетическая разметка полученных фреймов. В результате получается фонетическая сетка, характеризующая речевое высказывание. На заключительном этапе предпринимается попытка определить значимое слово (или строку). В процессе идентификации слова учитываются лингвистические пределы (т.е., словарный запас, синтаксис и другие семантические правила).

Сущность подхода, основанного на распознавании паттернов, заключается в использовании точно сформулированной математической структуры и создании устойчивых речевых паттернов для надежного сравнения с набором размеченных шаблонов на основе формального обучающего алгоритма [5, 11]. Речевые паттерны могут быть представлены в форме речевых шаблонов или статистических моделей и применены к звуку, слову или фразе. Распознавание паттернов осуществляется в два этапа – обучение и сравнение. На стадии сравнения с шаблоном производится прямое сравнение между распознаваемой речью со всеми возможными паттернами, которым обучилась система на этапе обучения. В рамках данного подхода развивались два метода, а именно, шаблонный и стохастический.

В рамках шаблонного подхода к решению проблем распознавания речи было разработано целое семейство техник. Идея, лежащая в их основе, проста. Коллекция прототипических речевых паттернов хранится как архив эталонов, представляющий словарный запас диктора. Распознавание осуществляется путем сравнения неизвестного речевого высказывания с каждым из архивных шаблонов и последующего выбора категории по принципу лучшего совпадения с ним. Как правило, шаблоны строятся на основе целых слов. Это позволяет избежать ошибок, возникающих при сегментировании и классификации меньших акустических, но более вариативных единиц, таких, как фонемы. Ключевая идея шаблонного подхода в том, чтобы произвести типичные последовательности речевых фреймов для шаблона с помощью процедуры усреднения и сравнить паттерны на основе измерений локального спектрального расстояния. Для согласования паттернов по времени применяется одна из форм динамического программирования, что помогает учесть скорость речи, изменяющуюся от диктора к диктору, а также повторы одного и того же слова. Стохастический метод использует вероятностные модели для решения проблем, связанных с неопределенностью и неполнотой информации. В распознавании речи неопределенность и неполнота возникают как следствие ошибочных звуков, вариативности говорящего, контекстуальных эффектов и гомоморфных слов. Наиболее популярный стохастический подход – это скрытая модель Маркова [9, 10]. По сравнению с шаблонным подходом скрытое марковское моделирование – более общая модель и имеет строгое математическое обоснование. Скрытая модель Маркова легко интегрирует источники знаний в архитектуру. Недостатком марковского моделирования является то, что оно не обеспечивает понимание процесса распознавания. Поэтому при попытках улучшить представление очень трудным оказывается проанализировать ошибки скрытой модели Маркова.

Модели, основанные на применении искусственных нейронных сетей, принципиально полагаются на стратегии обучения, так же как и стохастические техники. Но подход на базе нейронных сетей стремится оптимизировать или организовать сеть обрабатывающих элементов. Но не делает никаких предположений о вероятностных распределениях. Многослойные нейронные сети способны генерировать сложные нелинейные классификаторы или функции отображения [3, 4].

Подход с использованием искусственного интеллекта предпринимает попытку автоматизировать процесс распознавания подобно тому, как человек применяет интеллект при визуализации, анализе и характеристике речи [1, 2]. Среди техник, используемых в данном классе методов, применяются экспертные системы, которые интегрируют фонетическую, лексическую, синтаксическую, семантическую и даже прагматическую информацию для сегментирования и разметки, а также такие инструменты, как

искусственные нейронные сети для определения отношений между звуковыми событиями. В данном подходе знания или ограничения не кодируются в индивидуальных единицах, правилах или процедурах, а распределяется между простейшими вычислительными единицами. Нечеткость моделируется не как подобие или функция плотности вероятности, а как паттерн активности большого числа единиц.

Литература

1. Бова В.В., Дуккарт А.Н., Нагоев З.В., Токмакова Д.Г. Метод формального представления семантики естественного языка на основе мультиагентной рекурсивной когнитивной архитектуры. // Известия КБНЦ РАН №6 (62) 2014. 46-51

2. Нагоев З. В. «Интеллектика, или мышление в живых и искусственных системах», Изд-во КБНЦ РАН, Нальчик, 2013.

3. Овчинников П. Е. «Методы обнаружения и детектирования сигналов с априорно неизвестными статистическими свойствами с применением искусственных нейронных сетей». Дисс.канд.физ.-мат. наук, Н. Новгород, 2009.

4. Оганезов А. Л. «Применение нейронных сетей в задачах распознавания образов». Дисс.канд.физ.-мат.наук, Тбилиси, 2006.

5. Bhargaval S., Purohit G. N. An Experimental Survey on Parsing with Neural and Finite Automata, International Journal of Computer Science Issues, Vol. 8, Issue 4, No 2, July 2011.

6. Cole R., Mariani J. et all A Survey of the State of the Art in Human Language Technology, web edition Cambridge University Press and Giardini, 1997.

7. De Mulder et all A survey on the Application of Recurrent Neural Networks to Statistical Language Modeling.//Computer Speech and Language 30(2015) 61-98

8. Ginzburg R. S. A Course in Modern English Lexicology, M.: Vysshaya Skola, 1979.

9. Jurafsky D., Martin J. H. Speech and Language Processing., 2004.

10. Mote R Natural Language Processing. A Survey., 2002.

11. Reddy R. Speech Recognition by Machine: A Review// Proceedings of the IEEE, vol. 64, No. 4, April 1976.

12. Saini P., Kaur P. Automatic Speech Recognition: A Review//International Journal of Engineering Trends and Technology Volume 4 Issue2- 2013.

УДК 004.825

АВТОМАТИЧЕСКИЙ МОРФЕМНЫЙ РАЗБОР ГЛАГОЛОВ АГГЛЮТИНАТИВНОГО ЯЗЫКА

Б.П. Тажев, И.П. Тажев, А.М. Ксалов
Институт информатики и проблем регионального управления Кабардино-Балкарского научного центра РАН, Нальчик
Кабардино-Балкарский государственный аграрный университет им. В.М. Кокова, Нальчик
boristazhevart@mail.ru, itazhev@yandex.ru, arsenksal@gmail.com

В статье рассматриваются возможности автоматического морфемного разбора глаголов агглютинативного языка на примере кабардино-черкесского языка и его использование в различных системах.

Ключевые слова: *агглютинативный язык, морфемный разбор, контекстно-свободные грамматики*

Агглютинативные языки (от лат. agglutinatio - приклеивание) - языки, имеющие строй, при котором доминирующим типом словоизменения является агглютинация («приклеивание») различных формантов (суффиксов или префиксов), причём каждый из них несёт только одно значение [1].

Агглютинативными называются языки, которые характеризуется наличием как словообразовательных, так и словоизменительных аффиксов. Агглютинативный строй противопоставлен флективному, в котором каждый формант несёт сразу несколько неразделимых значений (например, падеж, род, число и т. п.). Связь между корнем и аффиксом имеет специфические характеристики, отличающие агглютинативные языки от флективных. Они проявляются в следующем.

Во-первых, все морфы представлены вариантами, обусловленными фонетическими закономерностями. Далее в морфах отсутствуют чередования, которые выступают в качестве грамматических, то есть отсутствуют чередования, называемые внутренней флексией.

Во-вторых, аффиксы агглютинативных языков выражают только одно значение, что приводит к нанизыванию аффиксов при выражении разных грамматических значений.

В-третьих, в агглютинативных языках границы между морфемами характеризуются четкостью; на стыке морфем обычно не возникает значительных звуковых изменений, что, напротив, характерно для

флективных языков. Такое свободное соединение морфем называется агглютинацией, т.е. склеиванием.

Отметим другую особенность аффиксов. Они могут примыкать к основе с разных сторон.

Еще одним качеством агглютинативных аффиксов является то, что для выражения грамматических значений используется один определенный аффикс (в его фонетических вариантах), поэтому в агглютинативных языках обычно отсутствуют разнообразные типы склонения и спряжения.

В-четвертых, в агглютинативных языках основа слово без аффиксов может употребляться как самостоятельное слово, поскольку именно исходная форма слова зачастую совпадает с основой, к которой присоединяются аффиксы [2].

Словоизменительный принцип в агглютинативных языках, подразумевающий регулярный порядок присоединения аффиксов к основе, низкий процент грамматической омонимии, отсутствие инфиксов и прочих подобных случаев, затрудняющих реализацию автоматического распознавания грамматической формы слова, облегчают практическую реализацию программного средства для морфологического разбора слова [3]

Для автоматического морфемного анализа слов мы применяем КС-грамматики. Этот подход выбран из-за того, что мы используем в качестве объекта исследования агглютинативный язык, организация которого коррелирует с организацией правил в КС-грамматиках. В связи с этим применение КС-грамматик является наиболее эффективным для решения данной задачи.

Контекстно-свободная грамматика — частный случай формальной грамматики, у которой левые части всех продукций являются одиночными нетерминалами [4]. Смысл термина «контекстно-свободная» заключается в том, что возможность применить продукцию к нетерминалу, в отличие от общего случая неограниченной грамматики Хомского, не зависит от контекста этого нетерминала.

Пусть $G=(N, \Sigma, P, S)$ – КС-грамматика, правила которой занумерованы 1, 2, …, p. Пусть $\alpha \in (N \cup \Sigma)^*$. Тогда

1) левым выводом цепочки α называется вывод, на каждом шаге которого очередное правило применяется к самому левому нетерминалу;

Левым разбором цепочки α называется последовательность правил, примененных при левом выводе цепочки α из S;

2) правым выводом цепочки α называется вывод, на каждом шаге которого очередное правило применяется к самому правому нетерминалу;

Правым разбором цепочки α называется обращение последовательности правил, примененных при правом выводе цепочки α из S [5].

Эти разборы можно представить в виде последовательности номеров из множества {1,2, …, p}.

Пусть π=i_1… i_n – левый разбор цепочки w∈L(G), где G – КС-грамматика. Зная π, можно построить дерево разбора цепочки w следующим “нисходящим” образом. Начнем с корня, помеченного S. тогда i_t дает правило, которое надо применить к S. Допустим, что i_1 – номер правила S→ X_1… X_k. Присоединим k потомков к вершине, помеченной S, и пометим их X_1,X_2,…,X_k. Если X_i – первый слева нетерминал в цепочке X_1… X_k, то первыми i-1 символами цепочки w должны быть X_1… X_(i-1). Правило с номером i_2 должно тогда иметь вид X_j → Y_1… Y_l, и можно предложить построение дерева разбора цепочки w, применяя ту же процедуру к вершине, помеченной X_i. Продолжая в том же духе, можно построить все дерево разбора цепочки w, соответствующее левому разбору π.

Существует естественный класс грамматик – они называются LL(k)-грамматиками – для которых левый разбор можно сделать детерминированным с помощью простого приема, состоящего в том ,что анализатор заглядывает на входе на k символов вперед и делает очередной шаг на основе того, что он при этом видит. LL(k)-грамматики – это те, которые “естественным образом” анализируют детерминированным левым анализатором.

А правый разбор цепочки w в грамматике G=(N, Σ, P, S) – это последовательность правил, с помощью которых можно свернуть цепочку w к начальному символу S.

Пусть α=xβ – такая левовыводимая цепочка в грамматике G=(N, Σ, P, S), что x ϵ Σ*, а β либо начинается нетерминалом, либо пустая цепочка. Будем называть x законченной частью цепочки α, а β – незаконченной частью. Границу между x и β будем называть рубежом.

Идею, лежащую в основе понятия LL(k)-грамматики, интуитивно можно объяснить так: если мы строим левый вывод S⟹_l^* w и уже построили S⇒_l α_1 ⟹_l α_2 ⟹_l… ⟹_l α_i так что α_i ⟹_l^* w, то можно построить α_(i+1), т.е. сделать очередной шаг вывода, видя только законченную часть цепочки α_i и “еще немножко”, а именно следующие k входных символов цепочки w.

Метод рекурсивного спуска — алгоритм синтаксического анализа, реализуемый путём взаимного вызова парсящих процедур, соответствующих правилам контекстно-свободной грамматики или БНФ. Применения правил последовательно, слева направо поглощают токены, полученные от лексического анализатора. Это один из самых простых алгоритмов парсинга, подходящий для полностью ручной реализации

Используя метод рекурсивного спуска, были построены грамматики для кабардино-черкесского языка (рис 1.), который является агглютинативным [6]. Созданный алгоритм, используя грамматики и правила, разбирает слова и выдает словарную карточку с объяснением каждой морфемы слова (рис. 2). Данный алгоритм, созданный для кабардино-черкесского языка, может быть применен и к другим агглютинативным языкам при написании соответствующих грамматик.

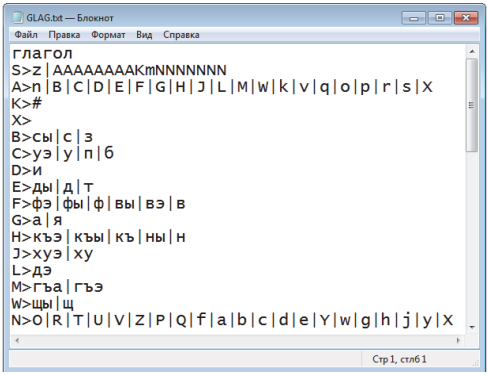


Рис. 1. Часть формальной грамматики глаголов агглютинативного языка

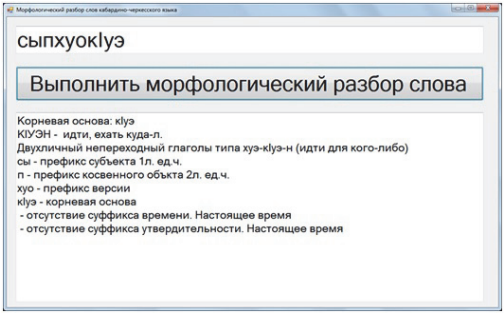


Рис. 2. Результат работы парсера слов кабардино-черкесского языка

С использованием данного метода был создан Эксплейн-сервис: адыгский язык, который представлен на сайте http://www.adygtion.org/. Эта система осуществляет морфологический разбор введенного слова, восстанавливает инфинитную (словарную) форму слова из корневой основы, выдает словарную карточку из лексикографической базы данных кабардино-черкесского языка с озвучкой данного слова, выдает объяснения каждой морфеме слова с указанием времени, лица и числа слова.

Вывод

Разработанные в рамках исследования алгоритмы с использованием КС-грамматик, позволяют выполнить морфологический (морфемный) анализ слов агглютинативного языка. На основе таких морфологических анализаторов могут быть созданы системы естественно-языкового интерфейса, которые могут быть применены в различных областях, таких как робототехника, речевой ввод, системы обучения языкам и др.

Литература

1. Н.Т. Гишев. Избранные труды по языкознанию. – Майкоп: ООО «Качество», 2008. - 538 с.
2. Г.С. Зенков, И.А. Сапожникова. Введение в языкознание. Учебное пособие для студентов дистанционного обучения КГНУ. Бишкек: ИИМОПКГНУ, 1998. - 218 с.
3. Б.В. Орехов, Е.А. Слободян. Проблемы автоматической морфологии агглютинативных языков и парсер башкирского языка // Информационные технологии и письменное наследие: материалы международной научной конференции (Уфа, 2010 г.) / отв. ред. В. А. Баранов. — Уфа; Ижевск: Вагант, 2010. — С. 167–171.
4. С. Гинзбург. Математическая теория контекстно-свободных языков. - М.: Мир, 1970.
5. А. Ахо, Дж. Ульман. Теория синтаксического анализа, перевода и компиляции. - Т. 1,2. - М.: Мир, 1979.
6. А.М. Ксалов, В.А. Денисенко, Ф.М. Гошокова. Алгоритм вывода неопределенной формы глагола агглютинативного языка на основе грамматик. - Известия КБНЦ РАН № 2 (46), Нальчик, 2012

УДК 519.766

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ КАЗАХСКОГО ЯЗЫКА НА ОСНОВЕ ПОЛНОЙ СИСТЕМЫ ОКОНЧАНИЙ

У.А.Тукеев, А.Тургынова
Казахский Национальный Университет им. Аль-Фараби, Алматы, Казахстан
ualsher.tukeyev@gmail.com, turgynovaa@gmail.com

В статье описывается подход к морфологическому анализу казахского языка на основе полной системы окончаний языка, что гарантирует правильный анализ любого слова языка. Предлагается построить алгоритм морфологического анализа в два этапа: выделения основы и окончания слова, а затем по окончанию слова нахождение грамматических характеристик слова. Второй этап предлагается строить

в виде трансдюсера (автомата Мили) с одним состоянием. Эксперименты показали высокую эффективность предлагаемого подхода морфологического анализа.

Ключевые слова: морфологический анализ, казахский язык, трансдюсер

1. Введение

Морфологический анализ предложений языка является начальным, базовым этапом во многих задачах обработки естественных языков, таких как, машинный перевод, анализ текстов и т.д.

В работе [1] разработана полная система окончаний казахского языка, что является базой для предлагаемого алгоритма морфологического анализа казахского языка. Основная идея предлагаемой работы заключается в том, что на основе полной системы окончаний строится решающая таблица, входом которой является окончание анализируемого слова, а выходом является набор грамматических характеристик данного окончания слова, которая полностью характеризирует состояние данного слова. Но, для полной реализации этой идеи необходимо выполнить также следующие действия для исходного предложения языка: - разделить предложение на слова; - для каждого слова выполнить операцию «стемминга» - разделение слова на основу и окончание; - по окончанию слова найти его (слова) грамматические характеристики. В последующем найденные грамматические характеристики могут быть использованы для последующих задач в зависимости от основной задачи. Например, если основная задача – машинный перевод, то необходимо на следующем этапе найти перевод данного слова на целевой язык. Здесь возможно, что необходимо будет решать проблему многозначных слов.

Предлагаемый подход к реализации морфологического анализа отличается от существующих тем, что: 1) базируется на полной системе окончаний языка, тем самым гарантируется анализ любого слова исходного языка, в данном случае казахского языка;

2) базируется на двухэтапном процессе морфологического анализа: выделении окончания слова, а затем по решающей таблице трансдюсера Мили с одним состоянием нахождении соответствующих грамматических характеристик слова, в отличии от подхода на основе FST(finite state transducers).

2. Краткий обзор по морфологическому анализу

Основным подходом к морфологическому анализу естественных языков является подход двухуровневой морфологии, предложенной [2], реализованной через использование конечных преобразователей - трансдюсеров (FST).

Имеется достаточно много публикаций в направлении использования двухуровневой морфологии и аппарата FST для морфологического анализа различных языков, в том числе, и для казахского языка [3-6].

3. Полное множество окончаний казахского языка.

В казахском языке окончания подразделяются на два больших класса:

- окончания на именной основе, к которым относятся окончания существительных, прилагательных и числительных;
- окончания на глагольной основе, к которым относятся окончания глаголов, причастий, деепричастий, наклонений и залогов.

В работе [7] впервые было рассмотрено построение множества окончаний казахского языка и возможность его использования для эффективного обучению казахскому языку.

Базовые аффиксы казахского языка, из которых формируются окончания казахского языка могут быть следующих четырех типов: - аффиксы множественного числа (обозначены через K), - аффиксы притяжательные (обозначены через T), - аффиксы падежные (обозначены через C), - аффиксы личные (обозначены через J).

В работе [1] рассмотрены всевозможные размещения варианты размещений типов аффиксов в окончаниях слов казахского языка: из одного типа, из двух типов, из трех типов и из четырех типов. Число размещений определяется формулой: $A_n^k = n!/(n-k)!$. Тогда, количество размещений будет определяться следующим образом:

$A_4^1=4!/(4-1)!=4$, $A_4^2= 4!/(4-2)!=12$, $A_4^3 = 4!/(4-3)!=24$, $A_4^4 = 4!/(4-4)!=24$. Всего возможных размещений 64.

В работе [1] путем рассмотрения семантически допустимых размещений определено полное множество семантически допустимых типов окончаний казахского языка.

Так, для слов с именными основами семантически допустимое количество типов окончаний казахского языка равно 15. Семантически допустимое количество типов окончаний казахского языка для причастий равно 11, для глаголов равно 25, для деепричастий равно 1, для наклонений равно - 6 и для залогов – 8. Итак, общее количество типов окончаний слов с глагольными основами будет 51.

Итого, общее количество окончаний с именными основами плюс общее количество типов окончаний слов с глагольными основами будет равно 66. По данным типам окончаний построены конечные множества окончаний для всех основных частей речи казахского языка. Так, для частей речи с именными основами количество окончаний равно 1213(учтены все варианты множественного числа), а количество окончаний частей речи с глагольными основами составляет: глаголы – 432, причастия- 1582, деепричастия- 48, наклонения – 240, залоги- 80. Итого, 3565 всего окончаний.

4. Алгоритмы морфологического анализа на основе полного множества окончаний

Общая схема морфологического анализа на основе полного множества окончаний включает два этапа:

- выделение основы и окончания анализируемого слова,
- вывод грамматических характеристик по окончанию слова.

4.1. Алгоритм стемминга на основе полного множества окончаний

Принцип выделения окончания (стемминг) слова, основанного на полной системе окончаний заключается в следующем. В системе окончаний казахского языка все окончания разбиваются на классы по длине окончаний. В слове сначала ищется окончание максимальной длины для данного слова: оно будет на два символа меньше длины слова (предполагается, что основа не может длины меньше, чем 2). Предполагаемое окончание длины L ищется в соответствующем классе окончаний длины L. Если окончание не находится в данном классе, то длина предполагаемого окончания уменьшается на единицу и ищется в соответствующем классе окончаний и т.д. до тех пор, пока не найдется окончание или слово будет без окончания.

Если окончание находится в соответствующем классе окончаний длины L, то выделяется основа слова путем отнимания окончания от слова, как правой ее части. Остальная часть слова будет являться основой, которая для достоверности еще проверяется поиском ее по словарю.

Обозначения:

$L(e)^{max}$ - есть максимальная длина окончаний в системе окончаний языка;

w - анализируемое слово:

e(w) – окончание анализируемого слова;

L(w) - длина анализируемого слова L(w);

L[e(w)] - предполагаемая длина окончания данного слова.

$L[e(w)]^{max}$ - максимальная длина окончания данного слова.

Шаги алгоритма:

Входом является слово w.

1. определяется длина анализируемого слова L(w).

2. определяется максимальная длина окончания анализируемого слова:

$L[e(w)]^{max} = L(w) - 2$.

где 2 – есть минимальная длина основы слова.

3. **Если** $L(w) \leq L(e)^{max}$ (если длина слова w меньше или равно максимальной длины окончаний в системе окончаний языка),

то предполагаемой длине окончания данного слова L[e(w)] присваиваем значение максимальной длины окончания анализируемого слова:

$L[e(w)] = L[e(w)]^{max}$. Далее перейти на 4.

4. **Иначе**: предполагаемой длине окончания данного слова L[e(w)] присваиваем $L(e)^{max}$:

$L[e(w)] = L(e)^{max}$.

5. Сделать выборку окончания e(w) длины L[e(w)] из данного слова w.

6. Проверка e(w) на совпадение с окончанием из списка окончаний длины L[e(w)].

Если совпадает,

то определяется основа данного слова:

{st(w) = w – e(w), т.е. из данного слова выделяется основа.

Выделенная основа для достоверности проверяется по словарю. В случае отсутствия основы в словаре, необходимо добавить основу в словарь.

Вариант основы и окончания для слова w сохраняется. Переход на п.7.}

Иначе

7. Уменьшаем предполагаемую длину окончания данного слова на единицу:

$L[e(w)] = L[e(w)] - 1$.

Если $L[e(w)] < 1$,

то выдаются варианты разбора или слово w без окончания. Переход на п.8.

Иначе – переход на п.5

8. Выдаются варианты разбора слова. Конец

4.2. Вывод грамматических характеристик по окончанию слова

Вывод грамматических характеристик по окончанию слова производится с использованием метода конечного преобразователя с многозначным отображением. Конечный преобразователь с многозначным отображением (Finite State Transducers with multivalued mappings) предложен в работе [8]. Так как данный FST используется только с одним состоянием, то это тривиальный конечный преобразователь Мили, а именно, FST Мили с одним состоянием: $y(t) = f_y(x(t))$, где x (t) – вход преобразователя, y (t) - выход преобразователя, t – текущее время, f_y – выходная функция преобразователя. Тогда, это будет Single State Transducer (SST) с многозначным отображением, т.е. это будет недетерминированный SST. В работе [8] показано как решать недетерминированность SST.

Для морфологического анализа слов казахского языка входом преобразователя SST будет множество окончаний казахского языка, а выходом – будут грамматические характеристики окончаний. Для каждого анализируемого слова сначала производится выделение основы и окончания, затем окончание подается на вход SST, в результате на выходе SST будет получен набор грамматических характеристик данного окончания.

5. Экспериментальные результаты

Для эксперимента взят текст биографии Абая Кунанбаева на казахском языке порядка 500 слов. Процент правильно проанализированных слов составляет 98 % для слов, основы которых присутствуют в словаре. 2% составляют слова, которые изменяют свою основу в соответсвии с правилом сингармонизма казахского языка. Неизвестные слова анализируются по разным вариантам длин окончании и представляются пользователю для включения в словарь. Программа также выдает варианты разбора слов в случае множественного грамматического разбора. Вопрос выбора варианта разбора слова решается на другом этапе, с учетом контекста данного слова.

6. Заключение

В работе представлен подход к разработке морфологического анализа казахского языка на основе полной системы окончаний казахского языка. Данный подход позволяет гарантировать морфологический анализ любого слова исходного языка, основа которого присутствует в словаре. В

предлагаемом подходе возникают вопросы множественного разбора слова, что должно решаться с учетом контекста данного слова. Предлагаемый подход применим и для других агглютинативных языков. В будущем планируются работы по расширению предлагаемого подхода для других языков.

Литература

1. Tukeyev, U., Automaton models of the morphology analysis and the completeness of the endings of the kazakh language. Proceedings of the international conference "Turkic languages processing" TURKLANG-2015 September 17–19, Kazan, Tatarstan, Russia, 2015. 91-100 pp

2. Koskenniemi K. 1983. *Two-level morphology: A general computational model of word-form recognition and production*. Tech. rep. Publication No. 11. Department of General Linguistics. University of Helsinki.

3. Oflazer K. 1994. *Two-level description of Turkish morphology*, Literary and Linguistic Computing Volume9, Issue2. 137-148.

4. Washington J. N., Salimzyanov I., Tyers F.M. 2014. *Finite-state morphological transducers for three Kypchak languages*. Proceedings of the 9th Conference on Language Resources and Evaluation.

5. Kairakbay B.M., Zaurbekov D. L. 2013. Finite *State Approach to the Kazakh Nominal Paradigm.*. Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing. St Andrews–Sctotland. 108–112.

6. Kessikbayeva G., Cicekli I. 2014. Rule *Based Morphological Analyzer of Kazakh Language*. Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, Baltimore, Maryland USA . 46–54.

7. Бектаев К., Ахабаев А., Керимбаев Е., Молдабеков К. Краткий казахско-русский словарь. Приложение 1- Список окончаний казахского языка. - Алма-Ата: Главная редакция Казахской советской энциклопедии, 1991, -256 стр.

8. Tukeyev, U., Miłosz, M., Zhumanov, Zh. Finite-State Transducers with Multivalued Mappings for Processing of Rich Inflectional Languages. In New trends in intelligent information and database systems (Vol. 598, pp. 271-280). Springer. 2015.

УДК 004.934

СРАВНЕНИЕ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ СИСТЕМЫ РАСПОЗНАВАНИЯ ТАТАРСКОЙ РЕЧИ

А.Ф. Хусаинов
«Институт прикладной семиотики «Академии наук Республики Татарстан», Казань
Казанский федеральный университет, Казань
Khusainov.aidar@gmail.com

В статье приводятся результаты экспериментов по созданию различных языковых моделей для татарского языка. Данные модели

могут быть использованы в различных приложениях: системах машинного перевода, проверки орфографии и т.д. В данном исследовании предполагалось дальнейшее использование моделей в составе системы автоматического распознавания татарской речи.

Ключевые слова: *языковая модель, татарский язык, распознавание речи*

1. Введение

Задача создания языковых моделей возникает при решении множества задач, от проверки орфографии и до систем машинного перевода. Во всех случаях языковая модель призвана описывать существующие в языке закономерности и на их основе уметь оценивать вероятности произнесения конкретных последовательностей слов.

Для определённого класса задач в качестве языковой модели может выступать набор грамматических правил, которые бы описывали структуру возможных в контексте данной предметной области фраз. Например, в задаче распознавания телефонного номера абонента, правила языковой модели могут допускать только повторения цифр нужное количество раз (в зависимости от формата телефонных номеров). Для записи грамматических правил часто используются логические операторы (например, оператор «ИЛИ») и именованные группы слов. Для унификации записи международным консорциумом W3C была создана спецификация Speech Recognition Grammar Specification (SRGS) [1], регламентирующая запись грамматических правил для систем распознавания речи.

Однако для более общих задач распознавания невозможно описать все возможные варианты фраз. В таких случаях в качестве языковой модели используют основанную на статистике модель n-грамм [2]. Модель n-грамм исходит из предположения, что вероятность произнесения слова можно рассчитать на основе последовательности предшествующих слов, а умея вычислять вероятность появления каждого слова во фразе, можно рассчитать и вероятность произнесения фразы целиком.

Вероятности произнесения каждого слова в различных контекстах определяются на этапе подготовки языковой модели на основе больших текстовых корпусов. В качестве оценки для условных вероятностей берётся отношение количества наблюдений последовательностей слов:

$$P(w_i | w_1,..., w_{i-1}) = \frac{N(w_1,..., w_i)}{N(w_1,..., w_{i-1})},$$

где $N(w_i,..., w_j)$ - количество наблюдений последовательности слов $w_i,..., w_j$ в корпусе.

С теоретической точки зрения, чем больше информации мы знаем об уже произнесенных словах, тем более точной является оценка вероятности текущего слова. Однако на практике приходится ограничивать анализируемый контекст, используя для оценки вероятности 1 или 2 предыдущих слова, то есть 2- или 3-граммы, соответственно. Ограничение вызвано вычислительной сложностью создаваемых моделей: количество параметров 3-граммной модели из 100 000 слов может составлять до 10^{20} [3]. Другой существенной проблемой является недостаток текстовых данных для полного обучения моделей: существует множество последовательностей слов, которые оказываются либо вообще не представленными в корпусе, либо встречаются там малое количество раз, недостаточное для точной оценки вероятностей. Наличие в произнесенной фразе хотя бы одной подпоследовательности слов, не встретившейся на этапе обучения (и, соответственно, имеющей нулевую вероятность), приведёт к обнулению вероятности всей фразы. Для преодоления данной ситуации разработаны методы сглаживания вероятностей, по имени создателей называющиеся методами Katz (1987), Kneser-Ney (1995), Good-Turing (1953), Jelinek-Mercer (1980), Witten-Bell (1990), Church-Gale (1991) [4].

Существуют разновидности описанной выше статистической модели n-грамм. Например, модели, основанные на классах слов, позволяют повышать размерность n-грамм на базе имеющихся текстовых корпусов; триггерные модели моделируют взаимоотношение пар слов в более длинном контексте [5].

Отдельно стоит отметить вариант n-граммных моделей, которые основываются на элементах, по размеру меньших слова (particle-based models). В этом случае слова представляются в виде морфем, и происходит анализ статистических закономерностей между ними, а не целыми словами. Данная особенность ценна для случаев распознавания языков с богатой морфологией, например, для флективных и агглютинативных языков: для финского, турецкого, эстонского, венгерского, русского.

2. Создание языковых моделей для татарского языка

Татарский язык относится к группе агглютинативных языков и имеет богатую морфологию. При построении стандартных статистических языковых моделей для таких языков возникает проблема с большим числом словоформ, которые необходимо включать в словарь. Большое количество различных аффиксальных цепочек, которые могут следовать за основой слова, делает невозможным построение словаря адекватных размеров с небольшим уровнем OOV (out of vocabulary) слов. Решением этих проблем, как было отмечено в п. 1, является уменьшение базовой

моделируемой единицы до элемента, меньшего чем слово. В качестве базовых подходов в текущем исследовании были выбраны следующие:

- морфемы;
- основы плюс аффиксальная цепочки;
- статистически выделенные морфы;
- слоги;
- буквы.

2.1. Инструментальные средства создания языковых моделей

В качестве основного инструмента для построения языковой модели татарского языка был выбран SRILM (Speech Technology and Research (STAR) Laboratory) [6]. Он включает в себя функционал по построению n-граммных моделей языка, алгоритмы интерполяции различных моделей, оценки качества построенных моделей. Работа с этим инструментом состоит из трёх этапов:

1. Вызов функции ngram-count для расчёта количества n-грамм;
2. Вызов функции ngram-count для построения языковой модели на основе результатов работы первого пункта с указанием выбранной функции сглаживания модели;
3. Оценка качества построенной модели на тестовом корпусе с помощью функции ngram и параметра -ppl.

Кроме того, были созданы программные средства по обработке текстового корпуса татарского языка и автоматизации процессов. Они включают в себя следующие основные модули:

1. Предобработка корпуса (фильтрация, разделение на тестовую и обучающую части);
2. Разбиение слов текстового корпуса на необходимые для моделирования элементы;
3. Средства автоматизации построения языковых моделей всех видов, проведения тестов, а также построения отчётов по итогам тестирования.

Остановимся на инструментарии, использованном при разделении слов текстового корпуса на необходимые для моделирования элементы. Так, разделение на отдельные морфемы или основы с аффиксальными цепочками осуществлялось с помощью морфоанализатора MorphAn [7]. Выделение «морфов» – определённых статистически частей слова – с помощью инструмента Morfessor [8]. Разбиение татарских слов на слоги происходило на основе знаний о 6 основных типов слогов в языке (Г, СГ, ГС, СГС, ГСС, СГСС) без учёта специфики разбиения заимствованных слов.

2.2. Текстовый корпус

Исходной информацией для обучения языковой модели выступил текстовый корпус татарского языка [9]. Полученный для работы фрагмент после процедуры фильтрации (удаления повторов, фрагментов русского и английского текстов, удаления специальных символов и т.д.) и разделения на обучающую и тестовую части имеет следующие характеристики, Табл. 1.

Таблица 1

Характеристики текстового корпуса

Корпус	Обучающая часть	Тестовая часть	Всего
Количество файлов	200 000	17 294	217 294
Количество слов	64 629 794	5 180 239	69 810 033
Количество слогов	172 193 048	13 821 430	186 014 478 (2,66/слово)
Количество морфем	102 131 309	8 149 139	110 280 448 (1,58/слово)
Количество морфов	86 507 729	6 950 813	93 458 542 (1,34/слово)
Количество основ и афф. цепочек	90 253 214	7 208 004	97 461 218 (1,4/слово)
Количество букв	402 356 569	32 279 979	434 636 548 (6,23/слово)
Размер	834 МБ	67 МБ	901 МБ

3. Результаты экспериментов

С учётом отсутствия до настоящего момента публикаций на тему построения и сравнения различных видов статистических языковых моделей для татарского языка, схема эксперимента была составлена таким образом, чтобы собрать максимально полную оценку влияния факторов на качество итоговой языковой модели. Так, были построены и проанализированы отдельные статистические модели для всех комбинаций из следующий категорий:

1. Тип элемента – 6 типов: слово, слог, морфема, морф, основа и аффиксальная цепочка, буква;

2. Размерность n-грамм: биграммы, триграммы, 4-граммы (5-граммы для модели на основе букв);
3. Алгоритм сглаживания модели – 5 типов: абсолютное сглаживание, Good-Turing, Kneser-Ney, Witten-Bell, модифицированный алгоритм Kneser-Ney.

Качество построенной модели оценивалось по таким показателям, как логарифм вероятности для тестового подкорпуса, perplexity (степень уверенности модели при анализе тестовых данных), OOV (количество элементов тестовой выборки, не вошедших в словарь) и размеру модели (по числу используемых n-грамм).

По результатам построения моделей был сделан вывод о том, что с точки зрения алгоритма сглаживания наилучшие результаты показали основной и модифицированный алгоритмы Kneser-Ney. Данные по значению параметра perplexity на примере морфемной модели представлены в Табл. 2.

Таблица 2

Значение perplexity для морфемной модели татарской языка при разных алгоритмах сглаживания

Сглаживание	2-грамм	3-грамм	4-грамм
Absolute	72,6082	37,2884	29,9665
Good-Turing	81,0384	42,8639	33,1613
Kneser-Ney	**72,0003**	**36,2964**	28,6693
Witten-Bell	73,193	37,2586	29,3679
Mod. Kneser-Ney	**72,0003**	**36,2964**	**27,9677**

Среди 95 построенных моделей наилучшее качество показала пословная модель, далее следуют модели на основе морфем и основ с аффиксальной цепочкой, морфов, слогов и букв, Табл 3.

Таблица 3

Сравнение языковых моделей

Базовый элемент	n-грамм	Размер словаря	Log вероятности, тыс.	OOV, %	Количество использованных n-грамм
Слово	4	1 029 311	**-12 209**	1%	30,5 млн.
Морфема	4	748 349	-12 638,7	0,5%	25,2 млн.
Морф	4	95 691	-12 772,4	**0%**	27,7 млн.
Слог	4	147 957.	-14 282	0,1%	17,5 млн.

Основа+цепочка	4	758 752	-12 386,7	0,5%	27,5 млн.
Буква	5	51	-20 741,5	**0%**	3,3 млн.

Как отмечалось выше, одной из основных проблем при статистическом моделировании языков с богатой морфологией является большой размер словаря, необходимый для покрытия лексикона, что приводит либо к снижению скорости работы систем с большим словарем, либо к возрастанию числа внесловарных слов при снижении размера словаря. С этой точки зрения, модели, построенные на основе элементов, меньших чем слово, показали значительное снижение числа OOV слов. Для эксперимента были построены словари для разных типов базовых единиц, состоящие из 20, 50 и 200 тысяч элементов. Результаты оценки построенных моделей представлены в Табл. 4. Наименьшее число элементов в словаре для полного покрытия тестового подкорпуса необходимо для моделей на основе слогов и статистически выделенных морфов.

Таблица 4

Результаты сравнения моделей со словарями
в 20, 50 и 200 тысяч элементов

Базовый элемент	Размер словаря	OOV	Размер словаря	OOV	Размер словаря	OOV
Слово, 3-грамм	20 тыс.	17%	50 тыс.	10%	200 тыс.	5%
Морфема, 3-грамм	20 тыс.	7%	50 тыс.	5%	200 тыс.	3%
Морф, 3-грамм	20 тыс.	3%	50 тыс.	0%	200 тыс.	-
Слог, 3-грамм	20 тыс.	0%	50 тыс.	0%	200 тыс.	-
Основа+цепочка, 3-грамм	20 тыс.	5%	50 тыс.	2%	200 тыс.	1%

В заключительном эксперименте была построена биграммная модель на основе классов слов для словаря в 20 тысяч элементов. Для выделения классов слов также использовался инструмент SRILM, реализующий для этих целей алгоритм Брауна [10]. Построенная модель отличается нулевым значением внесловарных слов, небольшим размером, однако уступает стандартным пословным моделям в качестве описания тестового подкорпуса. Интерес представляет результат разбиения словаря из 20 тысяч слов на классы: автоматически выделенные классы объединяют слова со схожими значениями. Например, в отдельные классы выделены названия населённых пунктов, чисел, годов, фамилий, названий стран, профессий и т.д.

Заключение

Необходимость построения языковых моделей возникает при решении большого спектра задач: распознавания речи, машинного переводчика, предиктивного набора. В контексте анализа агглютинативных языков стандартные подходы на основе построения пословных n-граммных моделей имеют серьёзные ограничения, вызванные богатой морфологией данных языков. Для решения данной проблемы при моделировании используются составные элементы слова. В данной работе для татарского языка были впервые произведено построение и сравнение моделей на основе слов, морфем, основ и аффиксальных цепочек, морфов, слогов и букв. Результаты эксперименты показали, что наилучшее качество моделирования особенностей татарского языка достигается при использовании пословных моделей, однако возможно существенное снижение необходимого размера словаря при использовании моделей слогов и морфов при относительно небольшом снижении качества моделирования.

Полученные результаты и модели планируется в дальнейшем внедрить в систему распознавания слитной речи на татарском языке.

Литература

1. Speech Recognition Grammar Specification Version 1.0 [Electronic resource]. URL: http://www.w3.org/TR/speech-grammar/ [Дата обращения: 01.04.2013].

2. Manning, C.D. Foundations of Statistical Natural Language Processing / C.D. Manning, H. Schutze. Cambridge, Massachusetts: MIT - Press, 1999. 704 p.

3. Карпов, А.А. Модели и программная реализация распознавания русской речи на основе морфемного анализа: дис. канд. техн. наук: 05.13.11 / Алексей Анатольевич Карпов. СПб., 2007. 132 с.

4. Chen, F. Goodman, J. An empirical study of smoothing techniques for language modeling. Computer speech and language (1999). 13. P.359–394. [Electronic resource]. URL: http://www.ideallibrary.com [Дата обращения: 02.07.2012].

5. Кипяткова, И.С., Карпов, А.А. Разработка и исследование статистической модели русского языка / И.С. Кипяткова, А.А. Карпов. // Труды СПИИРАН. – 2010. Вып. 1(12). С.35–49.

6. SRILM - The SRI Language Modeling Toolkit [Electronic resource]. URL: http://www.speech.sri.com/projects/srilm/ [Дата обращения: 01.02.2016].

7. Suleymanov, D. Sh. Tatar phonological rules as a base of two-level morphological analyzer, in Proceedings of LP'2000, ed. B.Palek and O.Fujimura/ D. Sh. Suleymanov, R. A. Guilmoulline, A. A. Guilmoulline – Prague : The Karolinum Press. – P. 495–504.

8. Mathias Creutz and Krista Lagus. Unsupervised discovery of morphemes. In Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02. (Philadelphia, Pennsylvania, 11 July, 2002). P. 21-30.

9. Dz. Suleymanov, O.A. Nevzorova, and B. Khakimov. National Corpus of the Tatar Language “Tugan Tel”: Structure and Features of Grammatical Annotation. Proc. International Conference Georgian Language and modern Technology. (Tbilisi, 2013). P. 107-108.

10. P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai and R. L. Mercer. Class-Based n-gram Models of Natural Language, Computational Linguistics 18(4), 1992. P. 467-479.

УДК 004.82; 004.912; 81.322.2

ОНТОЛОГИЧЕСКОЕ МОДЕЛИРОВАНИЕ МОРФОЛОГИЧЕСКИХ ПРАВИЛ ПРИЛАГАТЕЛЬНЫХ КАЗАХСКОГО И ТУРЕЦКОГО ЯЗЫКОВ

А. Шарипбай, Г. Бекманова, Л. Жеткенбай
Евразийский национальный университет им. Л.Н.Гумилева, Астана, Казахстан
sharalt@mail.ru, gulmira-r@yandex.ru, jetlen_7@mail.ru

Данная работа посвящается онтологическому моделированию морфологических правил прилагательных казахского и турецкого языков. Она позволит сравнивать сходства и различия между вышеуказанными языками. Результаты будут применяться для создания систем семантического перевода с казахского языка на турецкий язык, и наоборот, и для электронного обучения указанным языкам через компьютеры или через Интернет.

Ключевые слова: *Обработка естественного языка, имя прилагательное, морфологические правила, онтологическое моделирование, машинный перевод.*

1. Введение

Исследование «Обработка естественного языка (ОЕЯ) – Natural Language Proccesing (NLP)», которое отразилось в данной работе входит в область информатики «Искусственный интеллект». Мы рассмотрим проблемы компьютерной обработки казахского и турецкого языков, входящую в тюркскую группу естественного языка. В него входит интеллектуальные системы такие, как машинный перевод с одного языка на другой, смысловой ответ к заданному вопросу, автоматическое принятие решений, проверка правильности высказывания.

Казахский язык входит в тюркским, входящий в кыпчакскую группу, а турецкий в огузскую. Однако, они как и другие тюркскиеязыки имеют агглютинативное свойство. Это свойство характеризуется возможностью

к каждому коренному слову или основе посредством добавления аффиксов (суффиксы и окончания) формировать новые слова или словообразования. Здесь суффиксы изменяют семантику слов (значения), которая входит в семантическую категорию, формирующую новые слова, а окончание – в структурную категорию, меняющую только состав.Это свойство тюркоязычия дает возможность легко формализовать морфологические правила.

На сегодняшний день известно несколько методов формализации морфологических правил тюркских языков. В основном они не рассматривают семантику слов, а предназначенылишь для формального описания структуры, то есть формализованы только правила сочетания оканчаниис кореными словами. Среди таких работ можно выделить связанных с казахским языком работы [3-6]. Результатыданных исследованииииспользуются для разработки систем в части проверки написания казахских слов или обучения посредством компьютера, но они не дают возможности разработать машинный переводчик основанный на формальных правилах вследствие отсутствия семантического описаия слов. Поэтому мы разработаем такую систему, которая позволит формализовать правила словообразонвании.Для проведения исследования будем использовать метод онтологического моделирования дающая возможность семантически описать грамматические правила естественных языков.

2. Связанная работа

Сейчас есть разнообразные системы и методы машинного перевода языков [7-9]. Использование данных методов обусловлено сложностью формализации языка или наличием корпуса национальных языков.

К работам по машинному переводу в основе которых лежатграмматические правила с казахского на другие языки можно отнести [10-12]. В данных работах на основе системы Apertiumдля пар казахско-русский и казахско-английский были разработаны двуязычная база данных и структурированная база грамматических правил (морфологический, лексический, синтаксический); алгоритм и модель лексического анализа; технология автоматического формирования грамматических правил машинного перевода.

По проблематике машинного перевода с турецкого языка на другие посвящены работы [13-17]. В исследованиях по машинному переводу описывается компьютерный анализ морфологии турецкого и уйгурского языков. Создана архитектура системы по машинному переводу тюркских языков. Широко используемый метод двухуровневого анализа осуществлен посредством конечных автоматов. Результат морфологичекого

анализа серьезный шаг для задач обработки языков тюркской группы, поскольку никакая система обработки естественных языков не проектируется без морфологического анализа. Кроме того, были рассмотрены задачи по улучшению качества двухязычных корпусов статистического машинного перевода.

Использование достижении вышеуказанных работ значительно облегчают создание систем машинного перевода с казахского на турецкий язык, и наоборот.

3. Онтологическое моделирование морфологических правил прилагательных казахского и турецкого языков

Онтология является мощным и широко используемым инструментом для моделирования отношений между объектами, принадлежащими к различным предметным областям. Можно классифицировать онтологии на основе степени зависимости от задачи или прикладной области, модель представления онтологических знаний и выразительности, а также других критериев. Основная часть формально представленных знаний основана на концептуализации: объекты, концепции и другие объекты, которые считаются существуют в некоторой области интересов и отношения, которые держат среди них (Генесерет&Nilsson, 1987). Концептуализации является абстрактным, упрощенным взглядом на мир, который мы хотим представить для какой-то цели. Онтология явная спецификация концептуализации. Когда знания предметной области представлены в декларативном формализме, множество объектов, которые могут быть представлены называется универсум дискурса. Этот набор объектов, и описываемые отношения между ними, отражаются в изобразительной лексике, с помощью которого программа, основанная на знаниях представляет знания. Таким образом, в контексте искусственного интеллекта, мы можем описать онтологию программы путем определения набора репрезентативных терминов. В такой онтологии определения связывают имена сущностей во вселенной дискурса (например, классы, отношения, функции или другие объекты) с текстом читабельной, описывающие то, что означают имена и формальные аксиомы, которые ограничивают интерпретацию и хорошо сформированные использование этих терминов [18,19].

1. Какие области охватывает онтология? Ответ: Имя прилагательное.
2. Для чего нам нужна онтология? Ответ: Нужен для разработки сравнительной онтологической модели по прилагательным казахского и турецкого языков.

3. На какие виды вопросов должна отвечать информация в онтологии? Ответ: Нужен для спряжения по составу и значению прилагательных и определения видов имен прилагательных.

4. Кто будет использовать и поддерживать онтологию? Ответ: Лингвисты и программисты.

Согласно вышеуказанным вопросам сравнительная онтологическая модель имени прилагательного будет выглядеть O(X, R,I), здесь X - наименования входящие в структуру прилагательных (объекты и понятия), R – связи между наименованиями, а I – множество наименовании этих структур и связей

Сравнительная онтологическая модель имен прилагательных разработана в среде Protege (http://protege.stanford.edu). Язык Protege OWL дает возможность описать не только понятия, но и конкретные объекты. Онтологическая модель прилагательных казахского языка разработанная в среде Protege отображен на 1-рисунке, а онтологическая модель прилагательных турецкого языка на 2-рисунке.

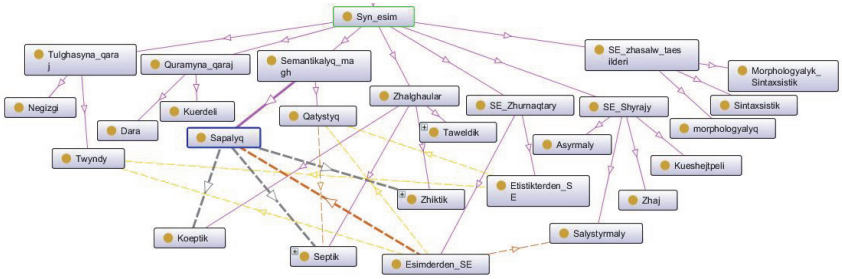


Рис. 1. Онтологическая модель имен прилагательных казахского языка

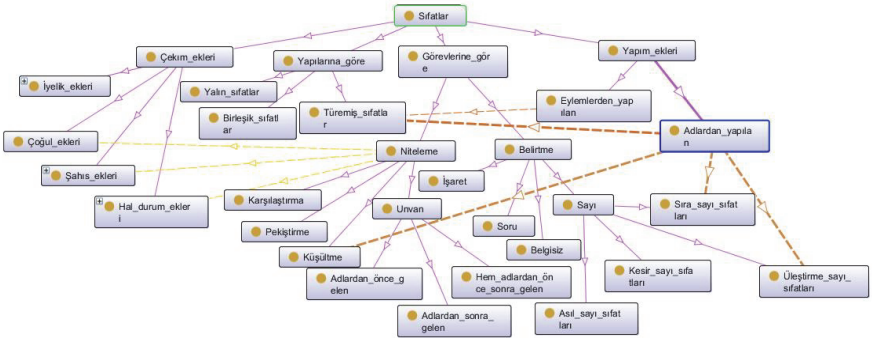


Рис. 2. Онтологическая модель имен прилагательных турецкого языка

Таким образом, сравнительная онтологическая модель прилага-тельных предназначенная для систем машинного перевода охватывает все компоненты множества морфологического анализа, в частности: семантические группы имен прилагательных казахского языка подразделяются на относительное и качественное, морфологический состав индивидуальный и сложный, по форме основной и производный. Имена прилагательные по способу образования делятся на морфоло-гические, синтаксические, морфолого-синтаксические методы. Имена прилагательные преобразуются по степени. Для того, чтобы связать прилагательные друг с другом использована функция Zhalghaw_ zhalghanady. Качественное прилагательное возрастает (Koeptik), спрягается (Zhiktik), склоняется (Septik), а относительное прилагательное только склоняется, таким образом имя прилагательное субстантивируется. С помощью функции Zhurnaq_ arqyly_zhasalady показано то, что сравнительная степень прилагательных формируется посредством добавления суффиксов к слову (Имени), а относительное прилагательное формируетсяне только от глаголов, но и от слов (Имени) с помощью суффиксов. А также прилагательные турецкого языка подразделяются на качественные и относительные (belirtme sıfatlar) прилагательные. По составу бывают основными, производными, совокупными (Birleşik). Качественные (Niteleme) прилагательные спрягаются на 4 вида: сравнительные (Karşılaştırma), усилительные (Pekiştirme), уменьшительный (Küçültme) и по званию (Unvan). В турецком языке качественное (Niteleme) прилагательное возрастает (Koeptik), спрягается (Zhiktik), склоняется (Septik), таким образом имя прилагательное субстантивируется, а относительные (Belirtmesıfatlar) прилагательные не возрастает, не спрягается, не склоняется. Путем добавления суффиксов к прилага-тельным образовывается производное прилагательное, уменьшительное прилагательное, порядковое прилагательное, распределительное числительное (Üleştirme Sayı Sıfatlar), детальное число (Kesir Sayı Sıfatlar).

Теперь остановимся на идентичности и различиях прилагательных казахского, турецкого языков. Спряжение субстантивного (овеществленного) положения прилагательных казахского и турецкого языков указан на 1-таблице.

Таблица 1

Спряжение субстантивного (овеществленного) положения прилагательных казахского и турецкого языков

Болымды		Отрисательный		Вопросительный	
Казахский	*Турецкий*	*Казахский*	*Турецкий*	*Казахский*	*Турецкий*
Ақылды-мын	Akıllı-yım	Ақылды емес-пін	Akıllı değil-im	Ақылды-мынба?	Akıllımı-yım?
Ақылды-сың	Akıllı-sın	Ақылды емес- сің	Akıllı değil-sin	Ақылды-сыңба?	Akıllımı-sın?
Ақылды-сыз		Ақылды емес-сіз		Ақылды-сызба?	
Ақылды	Akıllı(dır)	Ақылды емес	Akıllı değil(dir)	Ақылдыма?	Akıllı(dır)mı?
Ақылды-мыз	Akıllı-yız	Ақылды емес-піз	Akıllı değiliz	Ақылды-мызба?	Akıllımı-yız??
Ақылды-сыңдар	Akıllı-sınız	Ақылды емес- сіңдер	Akıllı değil-siniz	Ақылды-сыңдар ма?	Akıllımı-sınız?
Ақылды-сыздар		Ақылды емес-сіздер		Ақылды-сыздар ма?	
Ақылды	Akıllı-lar(dır)	Ақылды емес	Akıllı değil-ler(dir)	Ақылдыма?	Akıllı-lar(dır) mı?
Әдемі-мін	Güzel-im	Әдемі емес-пін	Güzel değil-im	Әдемі-мін бе?	Güzel mi-yim?
Әдемі-сің	Güzel-sin	Әдемі емес-сің	Güzel değil-sin	Әдемі-сің бе?	Güzel mi-sin?
Әдемі-сіз		Әдемі емес-сіз		Әдемі-сіз бе?	
Әдемі	Güzel(dir)	Әдемі емес	Güzel değil(dir)	Әдеміме?	Güzel(dir) mi?
Әдемі-міз	Güzel-iz	Әдемі емес-піз	Güzel değil-iz	Әдемі-міз бе?	Güzel miy-iz?
Әдемі-сіңдер	Güzel-siniz	Әдемі емес-сіңдер	Güzel değil-siniz	Әдемі-сіңдер ме?	Güzel mi-siniz?
Әдемі-сіздер		Әдемі емес-сіздер		Әдемі-сіздер ме?	
Әдемі	Güzel(dir)ler	Әдемі емес	Güzel değil-ler(dir)	Әдеміме?	Güzel(dir)ler mi?

В казахском языке усилительная степень образуется путем добавления согласного п к переднему слогу прилагательного, а в турецком к переднему слогу прилагательного добавляются согласные*м, n, p, c* (*m, p,*

r, s). Например, в казахском **ап**-ащы, **жап**-жаңа, а в турецком **yem**yeşil, **ap**acı, **yep**yeni, **ter**temiz, **mas**mavi и др.

Теперь создадим правила преобразования усилительной степени казахского, турецкого языкови рассмотрим формализацию скобочным методом.

Согласование: Сначала систематизируем буквы казахского и турецкого языков, обозначим их следующим образом.

Қазақша	Түрікше
АОҰЫ!01	AOUI!01
ӘӨҮЕІ!02	EİÖÜ!02
МНҢ!03	MN!03
РУЙ!04	RUY!04
Л!05	L!05
БГҒД!06	BGĞD!06
ЖЗ!07	CZ!07
П!08	P!08
К!09	K!09
Қ!10	H!10
СТШ!11	STŞ!11

1-правило. При добавлениив случае начинающихся с гласной буквы прилагательныхк переднему слогу согласного п (p) в казахском, турецком языке формируется усилительная степень. В казахском языке усилительная степень пишется через дефис, а в турецком слитно. По согласованию:

01X!((01p)-X)-**ап**-ащы	01X!((01p)X)-**ap**acı
01X!((01p)-X)-**ап**-аз	01X!((01p)X)-**ap**az
01X!((01p)-X)-**ұп**-ұзын	01X!((01p)X)-**up**uzun
01X! ((01p)-X)-**ып**-ыстық	01X!((01p)X)-**ıp**ıslak
02X! ((02p)-X)-**еп**-ескі	02X! ((02p)X)-**ep**eski
02X! ((02p)-X) - **үп**-үлкен	02X! ((02p)X)-**ip**ince

Правило 2. При добавлениив казахском языке в случае начинающихся с согласной буквы прилагательныхк переднему слогу п формируется усилительная степень, а в турецком к переднему слогу прилагательного добавляется согласные *м, п, р, с* (*m, p, r, s*). По согласованию:

0301X!((0301p)-X)-**моп**-момын	0301X!((0301s)X)-**mas**mavi
0602X!((0602p)-X)-**біп**-биік	0602X!((0602m)X)-**bem**beyaz
0601X!((0601p)-X)-**дап**-дайын	0601X!((0601p)X)-**dap**dar
0602X!((0602p)-X)-**дәп**-дәл	0602X!((0602m)X)-**düm**düz
0701X!((0701p)-X)-**жап**-жаңа	0702X! ((0702p)X)–**yep**yeni
0702X!((0702p)-X)-**жіп**-жіңішке	0702X!((0702m)X)-**yem**yeşil
0702X!((0702p)-X)-**жеп**-жеңіл	
0902X!((0902p)-X)-**көп**-көне	0901X!((0901p)X)-**kap**karanlık

1101X!((1101p)-X)-**тап**-таза 1101X!((1101p)X)- **tap**taze
1101X!((1101p)-X)-**сап**-сары 1101X!((1101p)X)-**sap**sarı

Сравнительная степень оценивает вид одной вещи к другой. Эти сравнительные знаки означают преимущества или неполноценность друг друга. Сравнительная степень в казахском языке преобразуется с помощью следующих суффиксов.

Казахский язык: Прилагательное+[рАқ, ЫрАқ, лАу, дАу, тАу, Қыл, Қылт, тым, шЫл, қай, аң]. Турецкий язык: daha (en)+имя прилагательное.

На казахском	***На турецком***
ақылды-*рақ*	*daha* akıllı
жылдам-*ырақ*	*daha* hızlı
арзан-*дау*	*daha* ucuz
кәрі-*леу*	*daha* yaşlı

4. Заключение

Разработанные онтологические модели для компьютерной обработки казахского и турецкого языков являются важным шагом при сравнительном исследовании двух тюрских языков. Поэтому исследование структуры и значений схожих прилагательных казахского и турецкого языков, и результаты их сравнения безусловно дает большую возможность для систем машинного перевода и разработки системы обработки естественных языков.

5. Будущая работа

В будущем будет разработана онтологическая модель морфологических правил имен числительных, глаголов и других частей речи казахского и турецкого языков.

Литература

1. Қазақграмматикасы. Фонетика, сөзжасам, морфология, синтаксис, Астана, 2002. – 784 бет.

2. Lewis, Geoffrey (2001). Turkish Grammar. Oxford University Press. ISBN 0-19-870036-9.

3. Шәріпбаев А.Ә. Қазақ тіліниңматематикалықтеориясыныңнегіздері // ҚазақстанРеспубликасы Ұлттық Ғылым Академиясының Информатика және басқару институтының еңбектер жинағы. Ғылым, – Алматы, 1996. б. 189-199.

4. Шарипбаев А.А., Бекманова Г.Т. Формализация морфологических правил казахского языка с помощью семантической нейронной сети // Доклады Национальной академии наук Республики Казахстан. – Алматы: 2009. – №4. – с. 11-16.

5. Шарипбаев А. А., Бекманова Г. Т. Построение логической семантики слов казахского языка // Материалы Всероссийской конференции с межународным участием «Знания-Онтологии-Теории». – Новосибирск: 2009. – Том 2. – с. 246-249.

6. Sharipbayev A.A., Bekmanova G.T. Construction the logical semantics of the Kazakh words // Proceedings of the All-Russian conference with international participation "Knowledge-Ontologies-Theories" - Novosibirsk, 2009. - Volume 2. - P. 246-249.

7. Koehn, F. J. Och, Marcu D. 2003. Statistical phrase-based translation. In Proc. of NAACL-HLT, Edmonton, Canada, P. 48–54.

8. Koehn H., Hoang A., Birch C., Callison-Burch M., Federico N., Bertoldi B., Cowan W., Shen C.Moran R., Zens C., Dyer O., Bojar A., Constantin E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the ACL 2007 Demo and Poster Sessions, Prague. AssociationforComputationalLinguistics. P. 177–180.

9. Lagarda A.L., Alabau V., Silva F. R., D´ıaz-de-Lianono E. Statistical Post-Editing of a Rule-Based Machine Translation System// Proceedings of NAACL HLT 2009: Short Papers, Boulder, Colorado, June 2009. AssociationforComputationalLinguistics. P 217–220.

10. Abduali B., AkhmadievaZh., Zholdybekova S., Tukeyev U., Rakhimova D. 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 "Turkic Languages Processing: TurkLang-2015". - Kazan: Academy of Sciences of the Republic of Tatarstan Press, 2015. P. 5-9.

11. Abduali B., Sundetova A., Zhanbussunov N., MusabekovaZh., Study of the problem of creating structural transfer rules for the Kazakh-English and Kazakh-Russian machine translation systems on Apertium platform. // VestnikKazNU. Al-Farabi, том 20, Proceedings of the International Conference "Computational and Informational Technologies in Science, Engineering and Education", - Almaty: Al-FarabiKazNU Press, 2015. P. 77-82.

12. Tukeyev U., ZhumanovZh.,Rakhimova D., Kartbayev A. Combinational Circuits Model of Kazakh and Russian Languages Morphology. Abstracts of International Conference "Computational and Informational Technologies in Science, Engineering and Education" (CiTech-2015, 24-27 September, 2015). - Almaty: Al-FarabiKazNU Press, 2015. P. 241-242.

13. Tantuğ A. C., Adalı E., Oflazer K., "Computer Analysis of the Turkmen Language Morphology," in FinTAL, Lecture Notes in Computer Science. vol. 4139: Springer, 2006, P. 186-193.

14. Orhun M., Tantuğ A. C., Adalı. E. Rule Based Analysis of the Uyghur Nouns.Proceedings of the International Conference on Asian Language Processing (IALP) .Chiang Mai, 2008. Thailand.

15. Orhun M., Tantuğ A. C., Adalı E. "Morphological Disambiguation Rules For Uyghur Language", IEEE International Conference on Software Engineering and Service Sciences (ICSESS), Beijing, China, 2010.

16. Tantuğ A.C. "Document Categorization with Modified Statistical Language Models for Agglutinative Languages", International Journal on Computational Intelligence Systems, vol. 5(3), 2010.

17. Ilgen B., Adali E., Tantuğ A.C. A Comparative Study to Determine the Effective Window Size of Turkish Word Sense Disambiguation Systems, 28th International Symposium on Computer and Information Sciences, Paris, France, 2013.

18. Gruber T.R. A Translation Approach to Portable Ontology Specifications / Gruber T.R.// Knowledge Acquisition,1993, P.199-220.

19. Gruber T.R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing / Gruber T.R. // International Journal Human-Computer Studies. – 1995, - Vol. 43, P.907-928.

УДК 004.82; 004.912; 81.322.2

МОРФОЛОГИЧЕСКИЙ АНАЛИЗАТОР КАЗАХСКОГО ЯЗЫКА НА ОСНОВЕ ОНТОЛОГИЧЕСКОГО МОДЕЛИРОВАНИЯ МОРФОЛОГИЧЕСКИХ ПРАВИЛ

А. Шарипбай, Б. Разахова, А. Зулхажав
Евразийский национальный университет им. Л.Н.Гумилева, Астана, Казахстан
sharalt@mail.ru, utalina@mail.ru, altinbekpin@gmail.com

В настоящей работе описывается универсальный морфологический анализатор казахского языка, построенный на основе онтологического моделирования морфологических правил. Данный анализатор может использоваться в машинном переводе, в информационном поиске, в извлечении знании из неструктурированного текста.

Ключевые слова: *обработка естественного языка, морфологический анализатор, онтология, морфология.*

1. Введение

Казахский язык входит в тюркскую группу и обладает свойством агглютинативностью, которая характеризуется большим числом словообразованием и словоформ для каждого корневого слова, образованных путем добавления к его концу аффиксов (суффиксов и окончаний). В этих языках определен строгий порядок добавления аффиксов: вначале к корню слова прибавляются суффиксы затем окончания множественности, притяжательные окончания, падежные окончания, окончания спряжения. Также определен строгий порядок слов в предложении: сначала размещается подлежащее, затем один из следующих членов предложения (дополнение, определение, обстоятельство), а в конце сказуемое [1].

Эти свойства тюркских языков допускают легкую формализации морфологических и синтаксических правил, поскольку с точки зрения морфологии в них строго определен порядок присоединения аффиксов, а с точки зрения синтаксиса строго определен порядок слов в предложении.

Для разработки морфологического анализатор любого языка, требуется формализация грамматических правил (фонетических, морфологических) и база слов или так называемые текстовые корпуса языка в неком структурированном виде, например как wordnet или OpenCorpora.

2. База слов или тезаурусов

При решении многих лингвистических задач, в том числе при морфологическом анализе часто используются так называемые текстовые корпуса как OpenCorpora. Наиболее информативными являются размеченные корпуса, то есть такие, в которых частям текста приписана лингвистическая информация. Например, каждое слово в корпусе снабжено исчерпывающими грамматическими характеристиками: к какой части речи оно принадлежит, в какой форме оно находится, какова его синтаксическая роль.[2], [3] корпус доступен бесплатно и в полном объёме (под лицензией CC-BY-SA). Разработчики корпуса создают хранилище текстов, специально предназначенное для текстов с лингвистической разметкой, удобный интерфейс редактирования разметки и исправления ошибок, инструменты для контроля качества и стандарт разметки для русского языка. На сегодняшний день OpenCorpora используется в морфолгическом анализаторе pymorphy2.

Также в морфологических анализаторах русского языка используется "Грамматический словарь русского языка" А.А.Зализняка (110 тыс.слов). В их числа входят On-line морфологический парсер от Yandex, морфлогический парсер Mystem. Тезаурусы (или семантические сети) — другой тип широко востребованных входных данных. Пожалуй, самый известный тезаурус — это WordNet, представляющий собой ресурс, в котором слова связаны с помощью так называемых семантических отношений: синонимии, гиперонимии (частное — обобщение), гипо- нимии (обобщение — частное), меронимии (часть — целое) и др. WordNet полезен в задачах машинного перевода, генерации текстов, классификации текстов, в морфлогическом анализе. Wordnet в основном используют при лемматизации, то есть при приведении слов к нормализованному виду. Если кратко описать процесс нормализации, то он заключается в следующем:

1. Для каждой части речи загружаются из WordNet по 2 файла — индексный словарь (имеет название index и расширение согласно части речи, например index.adv для наречий) и файл исключений (имеет расширение exc и название согласно части речи, например adv.exc для наречий).

2. При нормализации сперва проверяется массив исключений, если слово там есть, возвращается его нормализованная форма. Если слово не является исключением, то начинается привидение слова по грамматическим правилам, то есть отсекается окончание, приклеивается новое окончание, затем слово ищется в индексном массиве, и если оно там есть, то слово считается нормализованным. Иначе применяется следующее правило и тд, пока правила не закончатся или слово не будет нормализовано раньше[4].

3. Алгоритм морфологического анализатора казахского языка

Существует три основных вида проведении морфологического анализа (далее МА): декларативный, процедурный и комбинированный. При декларативном методе в словаре хранятся все возможные словоформы каждого слова с приписанной им морфологической информацией (далее МИ). В этом случае задача МА состоит просто в поиске словоформы в словаре и переписывании из словаря МИ, поэтому можно считать, что в этом методе отсутствует как таковой МА, а хранится только его результат. Алгоритм плох тем что требует больших затрат памяти, так как количество различных словоформ у каждого слова довольно велико. Процедурный МА выполняет следующие функции: выделяет в текущей словоформе основу, идентифицирует ее и приписывает данной словоформе соответствующий комплекс МИ. Процедурный метод предполагает предварительную систематизацию морфологических знаний о ЕЯ и разработку алгоритмов присвоения МИ отдельной словоформе. Недостатком такого подхода является высокая трудоемкость составления словарей совместимости. Для нашего алгоритма мы выбрали комбинированный тип, где сперва прогоняем процедурный МА, после проводим коррекцию морфологического анализа используя декларативный МА, если разбор слова вернула несколько вариантов анализа. Опишем одну из них. В казахском языке имеются окончания и суффиксы которые пишутся и произносятся одинково. Например словоизмениетельные морфемы – ды/-ді, - ты/-ті иногда являются:

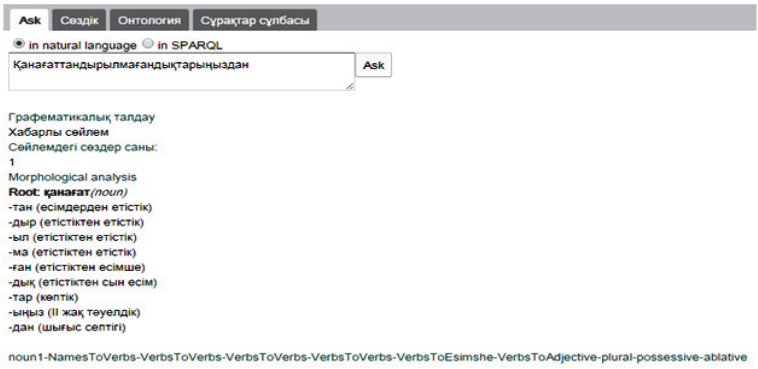
1. суффиксами образующие относительные прилагательные, например атты әскер (конница, кавалерия), ақылды (умный) и т.д.;
2. суффиксами прошедшего времени глагола, например атты (выстрелил), келді (пришел), и т.д.;
3. окончаниями винительного падежа: атты (коня), кемені (корабля), и т.д.

Как видим из примеров приведенных выше слово “ат” является омонимом (имя существительним конь или глаголом стреляй) и разбор слова “атты” может вернуть следующие вариации: коня (имя сущ.+ окон. вин. падежа), конница (имя. сущ + суффикс образ. отн. прил.) и выстрелил (глагол + оконч.прош.вр.глагола). В таких случаях мы сперва опираемся на порядок слова в предложении (если разбираем предложение), так как обычно глагол в прошедшей форме встречается в конце предложений, а после относительного прилагательного должен стоять существительное. Если и после этих разборов встречаются вариации, переходим на декларативный разбор с использованием семантического словаря.

Алгоритм анализа:

1. Начало. Слово ищется из базы словоформ. Если слово встречается, то определяем часть речи, в противном случае переходим к второму шагу. Если имеется словоизменительная часть, то преходим к третьему шагу;
2. Из слова циклический убираем последнюю букву и переходим к первому шагу. А убранные буквы сохраняем в памяти как строку суффиксов. Если достигли первую букву и слова не нашлось в базе, то вернем сообщение о неизвестной слове.
3. Циклический относительно части речи строка суффиксов ищется среди словопреобразующих суффиксов. Если суффикс встречается то меняем часть речи относительно к суффиксу и если строка окончании не пустая то ее значение присваиваем строке суффиксов, если суффикс не встречается то переходим к четвертому шагу.
4. Из строки суффиксов убираем последнюю букву и переходим к третьему шагу. А убранные буквы сохраняем в памяти как строка окончаний. Если достигли первую букву и слова не нашлось в базе, то переходим к пятому шагу.
5. Циклический строка окончаний ищется среди окончаний. Если окончание встречается то пишется морфологическая информация о ней и если остаток от строки окончаний не пустой то его значение присваиваем к строке окончаний, если окончание не встречается то переходим к шестому шагу.
6. Циклический из строки окончаний убираем последнюю букву и переходим к пятому шагу. А убранные буквы сохраняем в памяти как остаток от строки окончаний. Если достигли первую букву и остаток строки от окончаний не пустое, то запускаем коррекцию МА. Конец.

Скриншот результата работы алгоритма:



4. Заключение

Разработка морфологического анализатора казахского языка с комбинированным методом нам предоставляет хороший возможность объединить преимущества процедурных и декларативных морфологических разборов и избежать от ошибок. А использование семантического словаря даст возможность улучшить качество машинного перевода казахского языка и большие преимущества при извлечений знаний из неструктурированного текста.

Литература

1. Sharipbay, G. Bekmanova, A.Buribayeva, B. Yergesh, A. Mukanova, A.Kaliyev. Semantic neural network model of morphological rules of the agglutinative languages// The 6th International Conference on Soft Computing and Intelligent Systems / The 13th International Symposium on Advanced Intelligent Systems. – Kobe, Japan, 2012, P.1094-1099;

2. http://opencorpora.org/;

3. Дмитрий Ильвовский, Екатерина Черняк, Системы автоматической обработки текстов, www.osmag.ru, 01/2014, Открытые системы;

4. Аня Компанец, Частотный анализатор английских слов, написаный на python 3, умеющий нормализовывать слова с помощью WordNet и переводить с помощью StarDict, https://habrahabr.ru/post/161073/

РАЗДЕЛ IV | ЛЕКСИКОГРАФИЧЕСКИЕ БАЗЫ ДАННЫХ

УДК 81'374.81

К ВОПРОСУ О СОСТАВЛЕНИИ И РАЗРАБОТКЕ ЭВЕНКИЙСКО-АНГЛИЙСКОГО СЛОВАРЯ

А.Б. Анисимов
Северо-Восточный федеральный университет, Якутск
anis_and@mail.ru

Данная статья посвящена рассмотрению вопроса о составлении и разработке эвенкийско-английского словаря. В современной отечественной лексикографии подобный словарь еще не составлялся. В статье рассматриваются два основных этапа работы над эвенкийско-английским словарем.

Ключевые слова: *эвенки, эвенкийский язык, словарь*

Актуальность составления эвенкийско-английского словаря обусловлена тем, что в нынешней ситуации для возрождения и развития эвенкийского языка назрела необходимость использования максимально эффективных методов его сохранения и популяризации. Одним из этих методов является составление эвенкийско-английского словаря. Тем более, что подобный словарь еще не разрабатывался в отечественной лексикографии.

Некоторые лингвисты, делая акцент исключительно на коммуникативной функции языка, исходят из той позиции, что человек – существо социальное, и поэтому не может не общаться, не обмениваться информацией. На наш взгляд, такой односторонний подход не совсем оправдан. Мы считаем, что в отношении миноритарных языков в первую очередь ориентация на кумулятивную функцию языка представляется наиболее перспективной. В ситуации, когда естественная передача, в частности, эвенкийского языка имеет тенденцию к исчезновению, для лингвиста важно заметить жизненно важную функцию языка – функцию накопления и сохранения информации в словарном составе языка. Именно

такая позиция может позволить сохранить язык и разработать соответствующую стратегию его ревитализации.

В эпоху глобализации мир становится все более связанным и взаимозависимым. В глобализирующемся мире благодаря интенсивным контактам и взаимодействиям между различными культурами и языками стираются политические и идеологические границы. Глобализация, с одной стороны, ведет к нивелировке этнических различий, а с другой стороны, способствует синтезу локальных и глобальных элементов, что в свою очередь приводит к возрождению языка и культуры этнических меньшинств. К тому же в настоящее время социально-экономическая маргинализация малочисленных народов Севера стоит очень остро. Данная проблема может быть уменьшена благодаря повышению статуса языков этих народов. Таким образом, создание эвенкийско-английского словаря может быть одной из форм сохранения эвенкийского языка и способствовать повышению самосознания эвенкийского народа и стимуляции его социально-экономической активности.

Кроме того, сопоставление и сравнение языковых структур двух неродственных языков может позволить в перспективе выявить универсальные свойства человеческого языка.

В первую очередь перед составлением данного словаря были проанализированы русско-эвенкийские и эвенкийско-русские словари, составленные нашими отечественными исследователями Болдыревым Б.В. [3], Василевич Г.М. [5], В. Д. Колесниковой, О.А. Константиновой [7] и А.Н. Мыреевой [10]. Мы также опирались на работу М.И. Матусевич «Очерки системы фонем ербогоченского говора эвенкийского языка на основе экспериментальных данных» [9], которая сохраняет свое значение для изучения эвенкийской фонетики и в наши дни. Была изучена работа Т.Е. Андреевой [1], представляющая собой попытку поиска акустических коррелятов ударения в эвенкийском языке. Нами рассмотрен общий свод правил эвенкийской графики и орфографии, изданный в 1958 г. отечественными исследователями В.А. Горцевской, О.А. Константиновой и В.Д. Колесниковой [6]. Этот свод правил сохраняет статус нормативного документа и по сей день, поскольку эвенкийская графика со времени его выхода в свет не подвергалась пересмотру. Значительную помощь в составлении эвенкийско-английского словаря оказали исследование О.А. Константиновой «Эвенкийский язык. Фонетика и морфология» [8] и фундаментальная научная монография Б.В. Болдырева «Морфология эвенкийского языка» [4]. При изучении основ теоретической лексикографии исключительно важную роль сыграла работа академика Л.В. Щербы «Опыт общей теории лексикографии» [11].

При создании данного эвенкийско-английского словаря мы пользовались следующими методами исследования: методом компонентного анализа значений языковых единиц и методом семантического толкования, который необходим при выявлении значений слов.

В недалеком прошлом малочисленные народы Севера и Арктики воспринимались как носители архаической отживающей культуры. В результате такого отношения эти народы превратились в районах своего расселения в национальное меньшинство, а их самобытное развитие оказалось под серьезной угрозой. В настоящее время рождается новая концепция развития арктических и субарктических территорий, предусматривающая и новое отношение к самобытной культуре северных народов. С этой целью реализуются различные программы и проекты в поддержку нематериального наследия малочисленных народов Севера России.

Работа над данным словарем осуществляется в два основных этапа.

Первый этап работы над созданием эвенкийско-английского словаря носит аналитико-теоретический характер. Хотя следует отметить, что само разделение на этапы условно, ведь в действительности труд лексикографа имеет творческий характер.

На этом этапе работы были проанализированы русско-эвенкийские и эвенкийско-русские словари, труды отечественных лингвистов по исследованию фонетики, орфографии и морфологии эвенкийского языка. Следует отметить, что нормы литературного языка до сих пор окончательно не сформированы, сохраняется значительная диалектная раздробленность. Но, несмотря на значительную разобщённость говоров, диалектные различия не столь велики и не препятствуют общению эвенков между собой [2].

Нами также проведено исследование потребностей адресата – конечного пользователя словаря. В нашем случае конечным пользователем разработанного словаря является, в первую очередь, либо зарубежный специалист-филолог, занимающийся изучением эвенкийского языка, либо отечественный исследователь, владеющий английским языком. Словарь ориентирован на специалистов-филологов, тунгусо-маньчжуроведов и окажет неоценимую помощь в исследовании эвенкийского языка. Процесс интеграции локальных традиций и современного мира предполагает, что зарубежные исследователи тунгусо-маньчжурских народов будут стремиться овладеть эвенкийским языком, в частности. Этнический изоляционизм – это проблема, часто встречающаяся именно у малочисленных народов. Ошибочно предполагается, что этнический язык должен изучаться только членами данной этнической группы. Мы считаем, что языки малочисленных народов должны быть доступными всем, кто хочет их изучать.

Второй этап данного проекта представляет собой практический этап лексикографической работы, то есть работа непосредственно со словником. Что касается словника, его состав определяется потребностями пользователей и включает наиболее распространенные эвенкийские слова, включая и диалектные формы.

Составленный нами эвенкийско-английский словарь начинается с инструкции, как им пользоваться. В инструкции даются сокращения, используемые в словаре: adj –adjective; adv – adverb; fig – figurative; n – noun; num – number; pron – pronoun; v – verb и т.д.

Необходимость кириллического написания эвенкийских лексем объясняется тем, что современный эвенкийский алфавит основан на кириллице. Поскольку мы разработали эвенкийско-английский словарь, нам представляется целесообразным сохранение кириллической графики при написании эвенкийских лексем. Тем более в начале словаря прилагаются эвенкийский алфавит и руководство по произношению эвенкийских гласных, согласных и дифтонгов.

Эвенкийские слова в словаре имеют английскую транскрипцию:

абдукāн [abdu‘ka: n] (*acc* абдукāнмэ [abdu‘ka:nme]) *n* toy

илтэнмэк [ilten‘mek] *prep* past, *adv* incidentally

камая! [ka‘maja] *excl* hurry! be quick!

ливгэми [livge‘mi:] *v* snow

надан [na‘dan] *num* (*acc* наданма [na‘danma]) seven

Если в алфавитном порядке следуют друг за другом эвенкийские слова, которые имеют различную грамматическую форму, но на русский язык переводятся одинаково, то они помещаются рядом в строку, например: абул [a‘bul], абулкачин [a‘bulkatʃin] *n* lack [of something], shortage.

Если слово употребляется только в словосочетаниях или вне словосочетаний не поддаётся переводу, оно даётся на своём алфавитном месте, после него ставится двоеточие и приводятся сочетания, в которых это слово употребляется, например: угдывун [ug‘divun]: угдывун хувунин [ug ‘divun ‘huvunin] *n* fret saw (for ivory work).

Практически во всех языках мира обнаруживается явление лакунарности. При изучении словарного состава эвенкийского языка неизбежно возникает проблема передачи языковых лакун и безэквивалентной лексики. Мы исходим из лингвострановедческой позиции, ориентированной на понимание специфики эвенкийской культуры, отраженной в языке. Так, в данном словаре отражены слова-понятия, отражающие специфику языковой картины мира эвенкийского народа, например, дэтулэ [detu’le] – high moss cover in the woods, мушу [mu’ʃu] – a puddle in

the meadow, хукэлки [hu'kelki] – low spot on the road. Этот пласт лексики передается в словаре посредством описательного перевода.

К словарю также прилагается список наиболее часто используемых эвенкийских словообразовательных суффиксов. Это должно облегчить восприятие словаря зарубежными специалистами.

В разработанной нами модели эвенкийско-английский словарь состоит из следующих блоков:

1. Справочный материал (включая общую информацию о словаре, условные обозначения, алфавит эвенкийского языка, руководство по произношению эвенкийских гласных, согласных и дифтонгов).
2. Корпус словаря.
3. Список наиболее употребительных эвенкийских словообразовательных суффиксов.

Таким образом, мы получили следующую содержательную схему словарной статьи.

Основные зоны (типы) информации:

1. Входное слово (эвенкийская лексема).

1.1. Заглавная форма слова.

1.2. Фонетическая информация (транскрипция).

2. Зона переводных эквивалентов (английские лексемы).

2.1. Переводные эквиваленты на английском языке (отдельно для каждого узуального значения входного слова) – заглавная форма слова.

2.2. Грамматическая информация (принадлежность к грамматическому классу (часть речи)).

При составлении данного словаря мы стремились с возможной полнотой отразить лексическое богатство современного эвенкийского языка. Диалектная лексика даётся в тех случаях, когда она дополняет и обогащает общий лексический фонд эвенкийского языка.

В качестве основных критериев при отборе лексических единиц следует отметить важность или семантическую ценность лексемы, ее употребительность (частотность). При отборе слов, вошедших в словарь, представлялось разумным руководствоваться критерием целесообразности, то есть отбирать только те эквиваленты, которые с большой вероятностью понадобятся пользователю или те, которые незаменимы.

В заключение следует отметить, что данный эвенкийско-английский словарь представляет собой словарь общей лексики. Составители стремились отразить многовековое богатство языка эвенкийского народа и полнее представить лексику эвенкийского языка.

Настоящий словарь содержит около 2500 слов. Слова расположены в алфавитном порядке. По целевой установке данный словарь не имеет аналогов, как по своему построению, так и назначению и может послужить образцом при создании подобных словарей для других языков

малочисленных народов Севера. Главным преимуществом данного словаря является его двуязычность: полное и четкое толкование значений слов дается на эвенкийском и английском языках, что поможет зарубежным исследователям точнее и легче воспринимать и понимать семантическую структуру достаточно большого пласта лексики эвенкийского языка.

Несомненно, настоящей проверкой будет непосредственное использование данного словаря реальными пользователями, которые смогут вносить свои предложения по устранению недостатков и усовершенствованию словаря. Так как словарь электронный, то его обновление и внесение исправлений не составит большого труда.

Благодарности. Настоящая работа по составлению эвенкийско-английского словаря выполнена при финансовой поддержке Российского гуманитарного научного фонда (проект №14-14-14005).

Литература

1. Андреева Т.Е. Словесное ударение в дисиллабах эвенкийского языка по экспериментальным данным (на материале говоров эвенков Якутии) // Языки Сибири и Монголии. Новосибирск, 1987. С. 152–164.
2. Афанасьева Е. Ф. Эвенки: язык, фольклор, литература, этнография. Улан-Удэ : Изд-во БГУ, 1998.
3. Болдырев Б. В. Эвенкийско-русский словарь (Эвэды-лучадытурэрук). Новосибирск: Изд-во СО РАН, филиал «Гео», 2000.
4. Болдырев Б.В. Морфология эвенкийского языка. Новосибирск : Изд-во «Наука», 2007.
5. Василевич Г. М. Русско-эвенкийский словарь. В 2 ч. Ч. 2. – 2-е изд., перераб. СПб: филиал изд-ва «Просвещение», 2005.
6. Горцевская В.А., Константинова О.А, Колесникова В.Д. Основные правила произношения и правописания эвенкийского языка. Л., 1958.
7. Колесникова В. Д., Константинова О. А. Русско-эвенкийский словарь. Л., 1960.
8. Константинова О.А. Эвенкийский язык. Фонетика и морфология. М.-Л.: Наука, 1964.
9. Матусевич М.И. Очерки системы фонем ербогоченского говора эвенкийского языка на основе экспериментальных данных // Ученые записки ЛГУ (вопросы фонетики). 1960. № 237. В. 40. С. 83–102.
10. Мыреева А.Н. Эвенкийско-русский словарь. Около 30 000 слов. Новосибирск: Наука, 2004.
11. Щерба Л.В. Опыт общей теории лексикографии // Языковая система и речевая деятельность / Л. В. Щерба. Л.: Наука, 1974.

УДК 811.51

ФОРМИРОВАНИЕ ТЕРМИНОЛОГИЧЕСКОЙ СИСТЕМЫ ТАТАРСКОГО И ЧУВАШСКОГО ЯЗЫКОВ В ОБЛАСТИ ИНФОКОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ И ЕГО РЕАЛИЗАЦИЯ НА ПРИМЕРЕ ЧЕТЫРЕХЪЯЗЫЧНОГО СЛОВАРЯ

Д.Ш. Сулейманов, А.Ф. Галимянов
Академия наук Республики Татарстан, НИИ «Прикладная семиотика»
dvdt.slt@gmail.com, anis_59@mail.ru

В статье рассматриваются проблемы терминологии по информатике и информационным технологиям и образование новых терминов для татарского и чувашского языков

Ключевые слова: *информатика и информационные технологии, образование новых терминов*

1. Сравнительный анализ терминопостроения в тюркских языках в области информатики и информационных технологий. Термино-построение в казахском языке

Проведем краткий сравнительный анализ терминопостроения в тюркских языках в области информатики и информационных технологий. Так как в постсоветских республиках наибольшая работа в этой области проведена в Казахстане, остановимся кратко в работах казахских ученых в данной области. Как отмечает **казахский** ученый **Амандык Амирхамзин,** многоязычие страны не может не учитываться при развитии терминологии. В 1991 году в связи с обретением независимости в Казахстане стала быстро развиваться национальная терминология. Вместе с тем, имеются и другие примеры, где образование новых терминов и понятий развивается гораздо активнее. Так, в странах Балтики официально утверждено около 200 000 слов. А в Казахстане по данным сотрудника Комитета по языкам Министерства культуры и информации Республики Кузекбая Ерлана, начиная с 1992 года, было утверждено 2500 терминов, а в 2000-2008 годах было утверждено 8751 терминов, которые вызывали затруднения при переводе. В общей сложности, в течение 1992-2008 годов утверждено около 189 000 слов и словосочетании, из них 155 000 отраслевых терминов (был выпущен 31 том казахско-русского и русско-казахского словаря). Активизировала свою работу Терминологическая комиссия при Правительстве Республики Казахстан, которая заседает ежеквартально. Выпускаются журналы и

газеты по проблемам государственного языка, в том числе по вопросам терминологии (газета «Ана тілі», журналы «Терминологиялық хабаршы», «Тіл және қоғам», «Тіл» и т.д.). Член Гостерминкома, кандидат филологических наук Т. Туякбай внес предложение, чтобы термины, которые согласовывает Гостерминкомиссия, утверждались законом, а статус Гостерминкома нужно повысить. Министр культуры и информации Республики Казахстан Мухтар Кул-Мухаммед в своем интервью в газете «Егемен Қазақстан» по вопросам казахской терминологии выразил свою точку зрения:«Во-первых, я считаю целесообразным оставить слова латинского, греческого происхождения и термины, широко используемые в европейских языках в том виде как они есть, не изменяя их. Не следует воспринимать это, как слепое подражание Западу, или подпадание под влияние «великого русского языка», а принимать как меру, которая продиктована стремлением сохранить первоначальное значение и историческую природу терминов. Во-вторых, можно полностью согласиться с утверждением, что к словам чисто русского, славянского происхождения обязательно можно подобрать казахские аналоги. В-третьих, при утверждении любого термина было бы предпочтительнее, взяв его аналог из русского, турецкого, европейских языков, провести сравнительный анализ и принимать только те из них, которые соответствуют языковым нормам казахского языка. В-четвертых, в случае, когда встречаются трудности при терминообразовании на казахском языке, можно прибегать к терминам родственных языков, например, к турецкому. В-пятых, не ограничиваться первым попавшимся толкованием данного термина, а охватывать весь спектр сходных по значению слов».

Анализ особенностей формирования технических терминов в казахском языке показывает следующее.

В книге «Пән сөздері» (терминологиялық сөздік - терминологический словарь), который вышел в Кызылорде 1927 году, встречаются такие слова, как «автомобиль – аптамабіл», «барометр – барометір», «воронка – бәренке», «гелий – ілі», «трапеция – қостабан», «щелочь – сілті». В этот период термины, особенно технические термины иностранного происхождения переводятся, но с учетом особенностей казахского языка, то есть за основу берется только само слово. В «Терминологиялық хабаршы» («Терминологический вестник»), вышедшем в свет в 1998 году в Астане, есть устоявшиеся и очень удачные переводы терминов и названий, утвержденные Гостерминкомом в 1972-1981 годах в области техники, такие как «вертушка - зырылдауық», «вибрация - діріл», «вязкость - тұтқырлық», «выстой - кідіріс», «горелка - жанарғы», «заклепка - тойтарма», «зацепление - ілініс», «зубоокругляющий станок – тіс

жұмырлағыш станок», «зубострогальный станок – тіс сүрлеуші станок», «колесо - доңғалақ», «маслораспылитель – май бүркуіш». Начиная с девяностых годов Терминологическая комиссия утвердила термины по разным отраслям, например, в 1995 году по земельным наукам и металлургии, по общей технике и инженерным наукам: «акватория - айдын», «отвал - үйінді», «движение винтовое – бұрандалы қозғалыс», «желоб - науаша», «привод - жетек», «обод - құрсау». На заседании Терминкома от 24 июня 1998 года были утверждены термины, которые часто применяются в современном казахском языке, такие как «дефект - ақау», «сотовый телефон – ұялы телефон». Также, можно отметить очень удачные переводы терминов по геологии и горно-добывающему делу, например, «изумруд – зүбәржат», «кремень - шақпақтас», «шлих - түпшайма», «шлиф - тілімтас». В 2001 году издательством «Рауан» были выпущены русско-казахские и казахско-русские терминологические словари в 31 томах, утвержденные Терминологической комиссией, охватывающего 155 000 слов, куда внесли свои вклады ведущие специалисты разных отраслей. Каждый из этих 31 тома казахско-русского и русско-казахского терминологического словаря состоит из 5000 тысяч слов. Также запущен сайт http://www.techkz.kz/ и который охватывает такие сферы как: геологическая разведка, нефтегазовая сфера, информационные технологии, машиностроение, строительство и архитектура, металлургия. Вместе с тем, на сайте http://www.til.gov.kz/ организована коллективная работа специалистов электронно-терминологического центра, целью которой является формирование сборника терминов казахского языка, а в комплексе словарей создан единый фонд лингвистических и специализированных словарей с толкованием значений слов с помощью языкового пособия и фонда для описания специальных терминов научных отраслей и отраслей общественных служб в многоотраслевом терминологическом словаре. Также часто пользователи обращаются к русско-казахскому словарю сайта http://www.sozdik.kz/. Но он больше рассчитан на массового пользователя. На данный момент общее количество технических терминов в казахском языке составляет 360 тысяч слов.

Необходимо констатировать, что в последние годы терминологическая работа для казахского языка практически является образцом активности и масштабности для других тюркских языков.

2. Работы по терминологии в области информатики и информационных технологий для чувашского языка

Исследования проблем компьютеризации чувашского языка, проводимые в Чувашском государственном институте гуманитарных наук

и Чувашский государственном университете им. И.Н. Ульянова (Г.А. Дегтярев, И.В. Алексеев, В.П. Желтов и др.) показывают следующее.

Чувашский язык является одним из государственных языков Чувашской Республики — это закреплено в Конституции ЧР и республиканском Законе о языках. Вместе с тем, специалисты констатируют, что по-настоящему государственным языком он так и не стал. Отсутствует чётко продуманная и материально обеспеченная национально-языковая политика, из-за чего многие положения закона не претворяются в жизнь. В республике нет полномочного органа, занимающегося определением языковой политики и координацией деятельности в области языкового строительства. В нынешних условиях витальность (жизнеспособность) языка во многом зависит от того, насколько полно он представлен в информационно-коммуникационных технологиях и обеспечен возможностями компьютерной поддержки. Чтобы чувашский язык соответствовал нынешним реалиям, национально-языковая политика должна быть нацелена на обеспечение ему статуса рабочего языка компьютеров и Интернета, создание в глобальной сети соответствующей культурно-образовательной среды, информационных ресурсов.

Признается, что интеграция чувашского языка в сферу компьютерных информационных систем, полноправное представление его в Интернете (возможность читать на нём новости, художественную и специальную литературу, а также общаться посредством электронной почты, мессенджеров и т. д.) должны стать приоритетным направлением в области национально-языкового строительства. Совершенно справедливо, что ученые ставят перед собой и соответствующими официальными органами следующие актуальные задачи, которые практически уже решены для ряда тюркских языков, таких как татарский, казахский и турецкий:

1. Принятие стандарта кодировки дополнительных (диакритизированных) букв чувашского алфавита.
2. Разработка методики обучения компьютерной грамоте на чувашском языке и подготовка соответствующих учебных пособий.
3. Создание компьютерных игр, способствующих изучению чувашского языка, повышающих уровень владения речью.
4. Подготовка электронных аналогов традиционных чувашских словарей (Н. И. Ашмарина, Н. В. Никольского, В. Г. Егорова, М. И. Скворцова и др.), снабдив их новыми сервисами и функциями: разными способами навигации, полнотекстовым поиском.
5. Составление и эффективное ведение онлайновых словарей чувашского языка.

6. Реализация систем машинного перевода с чувашского языка на другие языки, а также с этих языков на чувашский.

7. Разработка утилиты проверки правописания чувашских текстов, создание грамматического и стилистического корректора.

8. Формирование чувашской компьютерной терминологии и создание англо-русско-чувашского глоссария по информатике.

9. Локализация популярных компьютерных программ на чувашский язык.

10. Создание Национального корпуса чувашского языка — информационно-справочной системы, основанной на собрании текстов различных жанров (отражающих письменную и устную речь), в электронной форме.

По некоторым направлениям работа уже начата. Силами энтузиастов ведётся локализация свободных компьютерных программ и веб-сайтов на чувашский язык. Появился чувашский раздел многоязычной свободной энциклопедии «Википедия» (в настоящее время он содержит более 10 тыс. статей), переведены интерфейсы программных продуктов Mozilla Firefox, CMS Drupal.

Важным событием в развитии чувашского сегмента Интернета явилось создание чувашского народного сайта Chuvash.org, в его рамках были открыты форумы для обсуждения на чувашском и русском языках животрепещущих национальных проблем. Позднее появился сайт Samah.chv.su, предоставляющий возможность поиска слов по электронным версиям чувашских словарей. Составлен специальный словарь для инструмента проверки орфографии Hunspell, применяемого в офисном пакете OpenOffice.org и браузере Mozilla Firefox.

Новые информационные технологии входят и в практику преподавания чувашского языка и литературы. Для учителей-новаторов стали привычными мультимедийное сопровождение уроков, привлечение ресурсов Интернета для исследовательских работ, использование компьютерного тестирования, участие в интернет-конференциях, создание компьютерных игр обучающего характера. В.Ю. Андреевым подготовлены первые электронные учебники чувашского языка. Н.А. Плотников основал электронную библиотеку Lib.chuvash.org, насчитывающую более 2 тыс. произведений чувашских авторов.

Адаптация программного обеспечения к региональным особенностям предполагает знание и учёт многих нюансов, так или иначе влияющих на её успешность:

1) при формировании интерфейсных текстов разработчики исходят из особенностей фонетического и грамматического строя чувашского языка, к которым относятся вариативность аффиксальных морфем (использование аффиксов с динамическими параметрами), постпозиция служебных слов и т. п.;

2) превышение объёма текста при переводе пользовательского интерфейса ведёт к изменению размеров интерфейсных элементов;

3) в чувашеязычной версии регистр букв отличается от английской — прописные буквы допустимы лишь в начале предложения, акронимах (пункты меню считаются отдельными предложениями);

4) в предложениях с динамически представляемыми параметрами, независимо от логики построения фразы, порядок следования аргументов не меняется и т. д.

Преобразование программного продукта для удобства работы с ним носителей целевого языка имеет и культурологический аспект, который предусматривает учёт социо-культурной специфики потенциальных пользователей, особенностей национальной ментальности. К примеру, форма вежливого обращения на «Вы» не соответствует традиционным канонам чувашского речевого этикета. В локализованных версиях компьютерных программ глаголы второго лица повелительного наклонения (императивные обращения) используются в единственном числе. Основной проблемой локализации является подбор чувашских компьютерных терминов, эквивалентных англоязычным. От того, насколько они удачны и органичны, во многом зависит успех локализации программных продуктов. В первую очередь необходимо создать равноценную терминологию интерфейса наиболее употребительных прикладных программ и веб-дизайна.

Как показывает анализ, проблемы чувашской локализации компьютерных систем практически идентичны с теми проблемами, с которыми мы сталкивались при татарской локализации, которая осуществляется с середины 1990-х годов. И в связи с этим, исполнение данного проекта и создание «Англо-русско-татарско-чувашского словаря терминов по информатике и информационным технологиям» является и актуальной и своевременной и подготовленной практикой национальной локализации компьютерных систем, активно проводимой исполнителями в течение последних 10-15 лет.

Формирование национальной компьютерной терминологии стало велением времени, требованием современной жизни. Массовая компьютеризация и проникновение сетевых технологий в повседневную жизнь человека превратили современных людей в «юзеров» (пользователей), применяющих компьютер не только в профессиональной деятельности, но и в быту. Выросло «компьютерное» поколение (в том числе чувашей и татар), активно потребляющее информацию в цифровом формате, вовлеченное в новые формы коммуникативной активности. В последние десятилетия компьютерная терминология перестала быть узкоспециальной, коммуникативно замкнутой сферой функционирования лексики.

Базовые термины из подъязыка ИТ-специалистов перешли в обыденную речь, из разряда технических терминов превратились в коммуникативно важные слова. В тюркских языках (скажем, в татарском и чувашском в том числе) ещё не сформировалась своя система обозначения компьютерных реалий, не сложилась околотерми-нологическая среда (субстандартный слой), поэтому в речи чуваше-язычных и татароязычных пользователей доминируют русские обороты с чувашской и татарской «инкрустацией»: *компьютера включать ту, компьютерны включать ит; домашни страницӑна уç, домашняя страницаны ач; личнӑй даннӑйне удалить ту; личный данныйларны удалят ит* и т.п. В этих условиях разработка многоязычного словаря терминов с татарским и чувашским языками является существенным вкладом в повышение языковой культуры и компьютерной грамотности народов, весомой поддержкой развитию этих языков. Компьютерная культура практически становится неотъемлемой частью современной национальной культуры.

Как справедливо отмечают многие авторы публикаций, в обществе и сознании пользователей до сих пор господствует технократический подход к компьютеризации. Наделение программного обеспечения «национальной оболочкой» лежит в русле нового подхода — гуманизации процесса компьютеризации, в которой немаловажное значение придаётся оптимизации коммуникации между человеком и машиной, раскрытию творческого потенциала пользователей. Общение с компьютером на родном языке раскрепощает человека, вызывает положительные ассоциативные связи с прежним эмоциональным опытом, текст в интерфейсе воспринимается лучше — потребитель программного продукта тратит минимум времени на понимание того, что написано в служебной строке, он может сосредоточиться на быстром и качественном выполнении своей задачи. О повышении эргономических характеристик локализованного интерфейса свидетельствует и то, что у начинающих пользователей сокращается время овладения навыками работы, снижается количество ошибок, повышается продуктивность познавательной, учебной и трудовой деятельности, осуществляемой на компьютере.

При создании терминов и выработки рекомендаций по их применению необходимо постоянно следить за языковой стихией, практикой употребления компьютерной лексики. Наблюдается большое расхождение в переводе обозначений соответствующих понятий, некорректность и непоследовательность использования терминов, злоупотребление заимствованиями. Такая ситуация вызвана тем, что вплоть до последнего времени национальная компьютерная терминология и вопросы её формирования и упорядочения оставались вне поля зрения специалистов. Следует также учесть и влияние экстралингвистических факторов. Быстрое развитие компьютерных технологий, постоянное

появление новых понятий требует динамичной реакции языка, оперативного ввода в терминосистему дополнительных обозначений. В этих условиях добиться строгой регламентации употребления терминов очень трудно.

В чувашскую часть трёхъязычного глоссария компьютерных терминов включены и некоторые вариантные обозначения, дублетные наименования. На этапе становления терминосистемы формальную избыточность средств реализации понятий — в научно обоснованных пределах — следует признать нормальной и оправданной. Нормативная оценка, осознанное упорядочивание терминологических единиц должны опираться не на субъективное восприятие, а на результаты естественного отбора.

3. Терминотворчество на татарском языке

3.1. Принципы и правила создания терминов и понятий

Практически основополагающие постулаты татарской терминологии отражены в [1]. Тем ни менее, анализируя более чем 20 летний опыт практической работы по татарской локализации компьютерных систем, составления словников и толкового словаря татарских терминов по информатике, исследованиями в этой области в других тюркских языках, можно указать следующие основные способы терминообразования по информатике и информационным технологиям на татарском языке, а также внести предложения по обобщению и уточнению имеющегося свода правил татарской терминологии. Они отражены в приведенной таблице. За основу рассмотрения нами взяты только основополагающие принципы, в предположении, что в рамках представленных принципов в дальнейшем могут быть детально описаны всевозможные правила, включая также и возможные исключения.

Принципы образования понятий и терминов можно разделить на два типа: принципы отвечающие на вопрос – “Каким образом порождаются татарские понятия и термины?” и принципы второго типа, отвечающие на вопрос – “Какими должны быть татарские понятия и термины?”.

Таким образом, принципы создания терминов на татарском и чувашском языках можно объединить в единую таблицу:

Принципы и правила создания терминов и понятий по информатике
и информационным технологиям в татарском и чувашском языках

Принципы язык	Поиск готовых понятий и терминов в самом языке	Порождение новых понятий, используя правила языка	Заимствование терминов из родственных (тюркских) языков	Заимствование понятий из других (нетюркских) языков	Заимствование понятий из других языков путем их перевода
на татарском языке	1.из малоупотребляемых слов 2.Из диалектов 3.Из старых слов	1.<корень>+<аффиксы> 2.<корень>+<корень> 3.определение одного значения с помощью словосочетания 4.порождение нового значения с помощью парных слов 5.блендинг 6.формальное гнездо	Практически отсутствует	Заимствование без перевода путем ассимиляции	1.прямой перевод 2.перевод по смыслу
на чувашском языке	1.специализация значения 2.актуализация 3.транстерминологизация	1.морфологический способ 2.синтаксический способ	Практически отсутствует	1.прямое заимствование 2.транскрипционное или транслитерированное заимствование	1.скрытое заимствование 2.повторное заимствование 3.смешанное заимствование

3.2. Вторая группа принципов образования понятий и терминов в татарском и чувашском языках

Вторая группа принципов определяет то, ***какими*** должны быть новые термины и понятия.

Первый принцип: “краткость термина” - чем короче слово, определяющее понятие или термин – тем лучше. Наиболее пердпочтительной формой слова является корневая.

Второй принцип: наиболее выигрышным является обозначение понятия или термина одним словом (особенно, с точки зрения технологий).

Третий принцип: “избегать” омонимии, многозначности. Понятия, термины должны иметь только одно значение. Например, слово *печать – бастыру* предпочтительнее слова *язу,* потому что писать *(язу)* можно и на экране, однако печатать (*бастыру)* можно только на принтере.

Четвертый принцип: “понятность”, “ясность” понятий, их распространенность, привычность; активное и широкое использование терминов.

Пятый принцип: “благозвучие” термина. Фонетичекая ассимиляция с татарским “произношением” (*cash:* руское произношение – *кэш*, татарское: *кәш*).

Шестой принцип: использование редких вариантов синонимов (как правило, диалектных вариантов) (*окно – тәрәз/тәрәзә: тәрәз*).

Седьмой принцип: “прямой перевод” - перевод иностранных понятий и терминов на татарский язык с языка оригинала, исправление некорректных переводов (кальки), полученных через русский язык (*арифметическое действие – арифметик гамәл – арифметика гамәле)*.

Восьмой принцип: “уход от неологизмов”. Не порождать новых слов, которые создают трудности в понимании и использовании термина.

Девятый принцип: иностранные слова должны заимствоваться строго в качестве корневых слов. Любые грамматические, фонетические и другие изменения и проявления данного слова в языке должны подчиняться правилам языка.

Литература

1. Правила образования, совершенствования и применения татарских терминов. Комитет при Кабинете Министров РТ по релизации Закона РТ “О языках народов РТ” / Под ред. зам. председателя Комитета проф. М.З. Закиева, председ. терминол. комиссии Комитета, доц. И.М. Низамова. – Казан, 1995. – 13 с. (на татарском языке).

2. Татарский язык и новые информационные технологии / Серия: Интеллект. Язык. Компьютер. – Вып. 2. – Казань: Изд-во Казан. ун-та. – 1995. – 123 с.

3. Сулейманов Д.Ш., Гатиатуллин А.Р. Структурно-функциональная компьютерная модель. – Казань, Изд-во “Фэн”, 2003. -220 с.

4. Сулейманов Д.Ш., Галимянов А.Ф., Валиев М.Х. Термины информатики и информационных технологий. Англо-татарско-русский толковый словарь (толкования на тат. яз.). – Казань: Магариф, 2006. – 383 с.

5. Шакирзянов Р.А. Новые тенденции терминообразования в татарском языке // В научно-информационном журнале «Наука и язык» (на тат. яз.). № 1 (46) #2011. – С. 23 - 25.

6. Ахметьянов Р.Г. Исторические источники татарской терминологии // В научно-информационном журнале «Наука и язык» (на тат. яз.). № 1-2 #1999; № 1-4 #2000; №1-4#2001; №1-2#2002.

УДК 81’373=811.512

ЯЗЫКОВЫЕ СРЕДСТВА СОЗДАНИЯ ОБРАЗНОСТИ В ПРОИЗВЕДЕНИЯХ А.ЕНИКИ И В ИХ ПЕРЕВОДАХ НА РУССКИЙ ЯЗЫК
Ә.ЕНИКИ ӘСӘРЛӘРЕ ҺӘМ АЛАРНЫҢ РУС ТЕЛЕНӘ ТӘРҖЕМӘЛӘРЕНДӘ ОБРАЗ ТУДЫРУДА ТЕЛ ЧАРАЛАРЫ

Э.Н.Денмухаметова, Р.Г.Гараева
Казанский федеральный университет, Казань
denmukhametova@gmail.com, rezeda-92g@mail.ru

В статье рассматриваются языковые способы создания образности А.Еники и описываются их отражение в переводных текстах на русский язык; анализируются находки переводчиков в использовании переводческих трансформаций; описываются неточности в переводных текстах, которые привели к искажению образа, созданного автором.

Ключевые слова: *татарский язык, перевод, языковые образы, трансформации, литературный текст*

Хәзерге вакытта телләрне чагыштырып өйрәнү зур темплар белән алга таба бара. “Тәрҗемә эшенең аеруча киң колач җәйгән урыны – күпмилләтле Россия Федерациясе. Тәрҗемә – илебездә халыклар дуслыгын, аларның туганнарча хезмәттәшлеген ныгытучы иң әһәмиятле чараларның берсе” – дип яза Р. Юсупов [9]. Тәрҗемә – ул бары тик телләрне генә чагыштырып өйрәнү түгел, ә шул телләрдә сөйләшүче халыкның мәдәниятен чагыштыру да. Тәрҗемә материалы буларак әдәби әсәрләрне өйрәнү дә бүгенге көндә актуальлеген югалтмый, чөнки күренекле әдипләрнең матур әдәдият әсәрләре аша халык менталитетын, йола-гадәтләрен һәм яшәү рәвешен белеп, бу халык белән танышып була .

Татар әдәбиятында иң күренекле һәм төрле буын укучысы тарафыннан да укыла торган язучыларының берсе булып, һичшиксез, Әмирхан Еники тора. Аның әсәрләрендә тормыш вакыйгалары, гади халык кичергән михнәт ачысы, яшәүгә булган тирән ышаныч сурәтләнә. Әдип, барыннан да элек, сугыш чынбарлыгының кеше рухында, табигате-холкында чагылышын анализлый һәм көндәлек тормыш вакыйгалары, детальләре аша үз геройларының эчке дөньясын, характер үзенчәлекләрен психологик төгәллек белән, тәэсирле итеп ачуга ирешә. Геройларының күңел серләренә тирән үтеп керә белү осталыгы, фикернең фәлсәфи үткенлеге һәм драматик киеренкелеге, форманың пөхтә һәм җентекле эшләнүе, тел-сурәтләү чараларының төрле стилистик буяуларга, образлы фикерләүдәге милли үзенчәлекләргә бай булуы – болар Ә. Еники талантының төп сыйфатларын тәшкил итәләр. Әдип образлык тудыруда, татар теленең нечкәлекләрен бик үткен тоемлап, тел материалы буларак тотрыклы гыйбарәләр, метафора, чагыштыру һәм сынландыруларны бик оста файдалана һәм уңышлы нәтиҗәләргә ирешә.

Ә. Еникинең иҗат юлы һәм үзенчәлекле стиле төрле галимнәр тарафыннан тикшерелгән. Ә. Еники иҗатына багышлап күп кенә диссертацияләр, фәнни мәкаләләр һәм дәреслекләр басылып чыккан. Мәсәлән әдипнең әсәрләрендә тел үзенчәлекләрен тикшерүгә багышланган диссертация һәм монографияләр: А.З. Кәримова “Психологизм в творчестве А. Еники” (2005), Г.Х. Зиннатуллина “Поэтическая ономастика прозы А. Еники” (2005) Шулар арасында: Ф.М. Хатипов, Р.Х. Свергин, М.Ш. Җәләлиева, Д.Ф. Зайидуллина һ.б. Бу хезмәтләр, ниггездә, әдипнең әдәби эшчәнлеге, әсәрләрдәге тел-сурәтләү чараларының үзенчәлекләрен тикшерүгә багышланганнар. Әмма Ә. Еникинең рус теленә тәрҗемә ителгән әсәрләренең телен, тәрҗемә алымнарын тикшерүгә багышланган хезмәтләр әлегә бик аз [4 -8]. Шунысы да мәгълүм: әдипнең әсәрләре рус теленә һәм башка милләтләр телләренә тәрҗемә ителгән. Бу эш белән Х. Хусаинова, А.Бадюгина, И. Вагапов, М. Рафиков, М. Зарецкий Л. Лебедева, Р.Кутуй һәм башкалар шөгыльләнгән һәм Ә. Еники тудырган образларны башка телләрдәдә гәүдәләндерергә тырышканнар.

Әлеге хезмәтнең максаты – Ә.Еники образларының рус теленә тәрҗемәдә бирелешен ачыклау. Моңа ирешү өчен, Ә. Еникинең “Вөждан”, “Гөләндәм туташ хатирәсе” әсәрләре һәм аларның Р. Кутуй тарафыннан рус теленә тәрҗемәләре өйрәнү объекты буларак алынды. Әлеге тәрҗемәләрне өйрәнгәннән соң түбәндәгеләрне күзәтү мөмкинчелеге туды.

Халыкның төп үзенчәлекләрен чагылдырган лексик сурәтләү чараларын – тропларны тәрҗемә итү аеручы кыенлык тудыра. Әмма мондый төр сурәтләү алымы – Ә.Еникинең иң яратып файдаланган алымнарыннан берсе. Шунлыктан аның әсәрләрен рус теленә тәрҗемә иткәндә тәрҗемәчеләр 4 төрле ысул кулланып туган ситуациядән чыгалар.

1) Эквивалент тәрҗемә – тропның икенче телдәге тулы эквивалентын табып тәрҗемә итү. Мәсәлән, *"**Тормышның авыр, каты мәктәбен үткән**"* (1: 112) – *"**на своем горбу испытал нелегкую участь**"* (2: 178). *"белгечләр арасында **кайнаган**"* (1: 114) – *"**вращавшийся** в избранном обществе"* (2: 181). Сүзгә-сүз тәрҗемә иткәндә *кайнаган – кипел* сүзе белән тәрҗемә ителергә тиеш иде, ләкин бу очракта тәрҗемәче рус телендәге "вращатся в обществе" фразеологизмын бик уңышлы файдаланып, ситуациядән чыга.
2) Калькалаштыру – бер телдәге тропны икенче телгә сүзгә-сүз тәрҗемә итү. Әмма әдәби тәрҗемә вакытында бу төр текст табигый яңгырашын югалтмаска тиеш. *Мәсәлән,* татар халкында чарасыз, котылгысыз хәлдә калган кешенең тормышка яраклашуын белдергән ***"башыңа төшкәч, башмакчы буласың"***мәкален Ә.Еники үз әсәрендә бик уңышлы куллана. Рус халкында "Поклонишься и кошке в ножки" дигән эквивалент мәкаль бар. Әмма Р. Кутуй тәрҗемәсендә без калькалаштыру очрагын күрәбез: ***"ежели подопрет – и сапожником станешь, поднажмет – и соловьем засвистишь!"*** (2: 178). Шулай ук *"**ни аллага ни муллага**"* (1.: 238) дигәнне дә Р.Кутуй *"**ни мулле, ни аллаху**"* (2.: 22) дип тәрҗемә итүне кулай күрә һәм шул рәвешле татар теленең җорлыгын русча белдерә алмый.
3) Төп текстның бер компонентын алмаштырып тәрҗемә итү ысулы исә, икенче телдә лексик сурәтләү чаралары аңлаешлы булсын өчен, аларны тулыландырып, яисә артык элементларын төшереп калдырып тәрҗемә итүне аңлата. Без караган әсәрләрдә ул түбәндәгечә яңгырый: *"Шактый биек тауга **менеп җиткәндәй авыр** сулап калды"* (1.: 106) чагыштыруы тәрҗемәдә *"**как будто одолел** высокую гору"* (2: 170) дип бирелә. Икенче мисалда *"Туташны **Әлдермеш** егетләре урлап киткәннәр"* (1: 301) тәрҗемәче *"Туташ украли **неизвестные** джигиты"* (2: 85) дип бирә һәм Әлдермеш авылы турында мәгълүмат югала. *"Мин – бүре, ә сез – **бәрән**"* (1: 249) чагыштыруы "*я – волк, вы - **зайчишка**"* (2: 34) һ.б. буларак тәрҗемә ителә
4) Компенсация алымын кулланып тәрҗемә итү дә рус теленә тәрҗемә вакытында шактый еш файдаланыла. Әлеге ысул буенча авырлык тудырган яки бөтенләй тәрҗемә итеп булмый торган троплар төшереп калдырылалар һәм текстның башка бер өлешендә бу мәгънә төсмере өстәлү рәвешендә биреләләр. Мәсәлән, Ә.Еники

тасвирындагы *“**караңгы чырайлы**, сөмсер кеше”* (1: 112) образы тәрҗемә барышында текстның нәкъ шул өлешендә төшеп кала һәм бары бер сүз белән генә бирелә – *“унылый и **неприветливый** человек”* (2.: 177). Ләкин шул ук абзац ахырында тәрҗемәче оригинал текстта булмаган гыйбарә кертә: *“его лихорадочный огонь в глазах – от **пожирающей его болезни**”* (2.: 177) дип өстәп куя. Шулай ук, Салихны тасвирлаганда автор кулланган *“**Башлы** кеше”* (1: 114) ***(Очень умный)*** эпитеты тәрҗемәдә төшеп кала. Әмма фраза ахырында фразеологизм белән мәгънә төсмере өстәп әйтелә *“**сумел “сделать имя”*** (2: 181) дип бирелә.

Матур әдәбият тәрҗемәсе турында сүз алып барганда, тәрҗемә ысуллары буларак трансформацияләр турында да онытырга ярамый. Алар билгеле бер мәгънәне оригинал берәмлекләреннән тәрҗемә берәмлекләренә күчерергә ярдәм итә торган үзгәртүләр буларак тәрҗемә вакытында һәрчак кулланыла киләләр.

Ә. Еники әсәрләренең русча тәрҗемәләрендә дә төрле трансформацияләр файдаланылган. Мәсәләгн, *Гөләндәм – Гуляндам, мирза – мурза, “Дим буенда” – “У Дёмы”, егет – джигит, әфәнде – эфэнди, туташ, абый, Исхак, Салих, Сабира кебек м*исаллардан күренгәнчә, ялгызлык исемнәр, реалияләр транслитерация ысуллары ярдәмендә тәрҗемә ителгәннәр. *“Каракүл **бизәкләрен** “**укырга** өйрәндем”* (1.: 210) – *“научился...**читать узоры** каракуля”* (2: 300); *“хыял күге”* (1.: 131) – *“**небо мечтаний**”* (2: 203); *“Басыйр шикелле **бөҗәк**”* (1.: 181) – *“это недоступно пониманию такого **насекомого**, как Басыйр”* (2.: 256) кебек мисаллардан исә, тәрҗемәченең калькалаштыру ысулыннан файдалануын күрәбез.

Лексик-семантик алмаштырулар – оригиналдагы лексик берәмлекләрне тәрҗемә телендәге берәмлекләр белән алмаштыру дигәнне белдерә. Бу ысул, Ә.Еники тудырган образларны бирүдә аеруча уңышлы булып күренә. Мәсәлән: *“...ул кызый **күңелнең** әллә кайсы гына төшендә һаман **яшереп ята бирде**”* (1: 213). Тәрҗемәдә *“...она никогда **не покидала моего сердца**”* (2.: 303) дип кенә бирелә, образлылык дәрәҗәсе кими. Ә.Еники әсәрендәге Гөләндәмнең *“кара челтәр **шәлемне ябындым**”* (1.: 382) сүзләре исә тәрҗемәдә *“**голову повязал** черным гипюровым **шарфиком**”* дип тәрҗемә ителә. (2.: 164). Салих сөйләмендәге *“**Исте җилләр, күчте комнар**...”* (1.: 212) җөмләләре *“**Жизнь прошла**... **Много воды утекло**...”* (2.: 302) буларак тәрҗемә ителәләр.

Конкретлаштыру – киң логик мәгънәгә ия булган чыганак телдәге лексик берәмлекне тәрҗемәдә мәгънәсе таррак булган сүз яки сүзтезмә белән алмаштыру. *“Башта язганымча, максатым бер ел эшләп, **кая да***

булса *укырга китү иде”* (1: 211) җөмләсе*“Как я писал вначале, цель моя, была, проработав год, поступить учиться* ***в вуз”*** (2.: 300) дип тәрҗемә ителә. Мисалдан күренгәнчә, тәрҗемәдә кая да булса укырга керегә дигән гыйбәрә рус теленә конкретлаштырып, ВУЗга укырга керегә дип күрсәтелә. Икенче мисалда исә: *“Әмин әкә* ***таза гәүдәле, кызыл чырайлы****, аю сыман алпан-тилпән атлап, ашыкмыйча йөри торган бер кеше иде”* (1.: 116) җөмләсе *“****Богатырского сложения****, с* ***кирпично-красным лицом****, крепкой воловьей шеей, огромными ручищами с короткими толстыми пальцами и медвежьей походкой в развалку – ну прямо батыр из сказки.”* (2.: 184) дип бирелә. Конкретлаштыру мисалыннан тыш әлеге җөмләдә авторның үзеннән өстәгән чагыштырулар һәм эпитетлар күренә.

Шулай ук грамматик формаларны алыштыру, җөмләләрне берләштерү, антонимик тәрҗемә аша да Ә.Еники образлары рус телендәге текстларда гәүдәләнеш табалар. Шул рәвешле Ә.Еникинең самими Гөләндәме тәрҗемәдә күндәм, әмма үзсүзле туташ буларак, чын совет чыныгуы үткән Хәбиб Юлдашев – батыр комсомол егет итеп, талантлы Салих – үз дөньясына баткан егет буларак тасвирлана. Кызганычка каршы, тәрҗемәче үз эшен башкарганда төрле алымнар һәм ысуллар кулланса да, Ә.Еникинең үткен телен, образлар тудырудагы нечкәлекләрен биреп бетерә алмаган. Моны ике телдә төгәл тәңгәл килерлек эквивалент метафоралар булмау, бер телдәге сурәтләү чараларының башка телдә шундый ук эффект тудырмавы белән аңлатырга кирәк.

Гомумиләштереп әйткәндә, тәрҗемәчеләр никадәр генә Ә.Еники тудырган образларны төгәл бирергә тырышмасыннар, рус телендә алар башкача гәүдәләнә, сөйләмнәрендәге үзенчәлекләр, тышкы кыяфәтләренең тасвирланышы, үз-үзләрен тотышта аермалы юклар булу автор тудырган образларны төгәл күзалларга комачаулый.

Әдәбият

1. Еники Ә. Кичке шәфәкъ: Повестьлар – Казан: Тат. кит. нәшр., 1989. – 384 б.
2. Еники А. Совесть: Повести. Пер. с тат. – Москва: Сов. Россия, 1985. – 360 с.
3. Заһидуллина Д. Ф. 20нче гасыр татар әдәбияты тарихы: дәреслек. – Казан: Казан университеты, 2011. – 198 б.
4. Зиннатуллина Г.Х. Поэтическая ономастика прозы А. Еники: автореф.дис.... канд.филол.наук.- Казань, 2005. – 25 с.
5. Кашкин В. Б. Сопоставительная лингвистика. Учебное пособие для ВУЗов. – В.: Изд-во Воронежский государственный университет, 2007. – 88 с.
6. Мотигуллина Ә. Ә. Еники прозасында геройларның характерлары: Фил. фән. канд. диссер. – Казан, 2000. – 240 б.
7. Фатхрахманов Р. Г. Творческая лаборатория прозаика (на материале произведений А. Еники, М. Магдеева, А. Гилязова, Н. Фаттаха и др.): Автореф. дис.

канд филол. Наук.; ИЯЛИ им. Г.Ибрагимова АН Республики Татарстан. – Казань, 2000. – 32 с.

8. Хатипов Ф. М., Свергин Р. Х. Әмирхан Еники иҗатында поэтик аһәңнәре.– Казан: Татар.кит.нәшр., 2011. – 111б.

9. Юсупов Р. А. Тәрҗемә һәм сөйләм культурасы. – Казан: Татар. кит. нәшр., 2008. – 240 б.

10. Юсупов Р.А. Соотношение разноструктурных языков и вопросы перевода (на материале русского и татарского языков). – Казань: КГПУ, 2005. – 225 с.

PRESENTATION OF SPATIAL-TEMPORAL RELATIONS IN KYRGYZ LANGUAGE

Sonunbubu Karabaeva
Bishkek (Kyrgyzstan)
sonun2008@mail.ru, k.sonun@gmail.com

Статья посвящена восприятию таких понятий как «пространство» и «время» в кыргызском языке. Мы исследовали пространственные отношения на кыргызском языке [3], [4]. Здесь мы рассмотрим отношения между пространством и временем.

The article is devoted to the peculiarities of concepts perception such as "space" and "time" in the Kyrgyz language. We investigated spatial relations in Kyrgyz language [3], [4]. Here we consider relations between space and time.

Key words: time and space; culture and language; spatio-temporal relations in the Kyrgyz language.

The perception of time and space has an important place in the worldvision of each person.

Space and time are basic entities of matter, and it is quite natural that the interest in studying it unabated throughout history of natural sciences and humanities. National mindset directly impacts the structure of language.

Perception of space and time are experiencing such dramatic changes worldwide, that it is impossible to pass them by. It is necessary to develop a new concept of perception and secure a specific place on a global space as well as large scale of time stream.

Space and time are inextricably linked and have the following properties: they are inseparable from its material host; they are objective, universal, contradictory, final and infinite, absolute and relative, continuity and discontinuity.

Concepts of space and time change their contours depending on the world model which is generally typical for the society and remains at the intersection

of natural factors and cultural components. These concepts help to interpret the ways in which people unconsciously construct the world around them.

Perceptions of space and time form a complex progressive system, which is reflecting the diversity of spatial and temporal relations.

According to the linguistic view of the world in Kyrgyz, space can also be associated with the concept of real world within the meaning of the universe as well as the system of world itself.

The Kyrgyz philology not yet described detailed mathematical models of the semantic mapping of space and their linguistic expression.

In addition, it must be said that the concept of time is closely connected with the culture and language of a certain ethnic group. Language reflects the culture and it means that all changes in a culture entailing changes in time perception; and this is reflected in language. Language responds to these changes by basic metaphorical models of time. It should be noted, however, that change of basic metaphor does not alter the previous metaphorical models, since the status of basic metaphor receives only a metaphor, which plays an important role in society, where the meaning and purpose of which do not lose significance within new conditions. The combination of these basic metaphors consists of modern temporal views of the world, and partly of the linguistic view of the world. Language always models and reflect sour universe in a way how we see it today. The language forms its own space-time, even a set of space-time "put" into each other fractal.

The Kyrgyz literature, in comparison with other types of creativity, uses time and space in a maximum freeway.

Time and space in Kyrgyz can be multidimensional.

In Kyrgyz language cases show a space-time of objects position related to one another.

Within the frame of partial spatial relations; the following parameters of objective reality, passing linguistic conceptualization are considered: limitation of localized object with the frame of localizer (İÇİ/TIŞI, İÇKİ/SIRTKI), with/without contact, KESİLIŞ/AYRILIŞ, as well as positioning of localizing object regarding localizer within a three-dimensional grid coordinate, when the mutual position of objects is defined as vertical (COGORU/TÖMÖN; ÖYDÖ/ILDIY, ÜSTÜ/ASTY), sagittal (arrow-shaped) (ALDI/ARTI; ASTINKI/ARTKI, İLGERİ/KİYİN) or horizontal (ONG/SOL; ONG CAK/SOL CAK, ONG KAPTAL/SOL KAPTAL).

First peculiarity of the Kyrgyz language is that parts of space related to any object from the viewpoint of the subject and taking gravitation and direction of observation and motion in account are presented as nouns, with corresponding cases: BARIŞ [literally: going] – Dative, CATIŞ [lying] – Locative, ÇIGIŞ [going out] – Ablative, TABIŞ [finding] - Accusative.

The denotations of parts of space in the Kyrgyz language are follows:

ÜSTÜ – upper-space,
ASTI – before-and-lower-(observed)-space,
İÇ - interior,
SIRT-; TIŞ - exterior,
ÇEK – boundary-strip,
SOL – left-space,
ONG – right-space,
ORTO – middle-sport,
CAN – near-space,
ARA – between-space,
ALD(ALDI) – before-forward-space,
ART(ARTI) – behind-space,
KARŞI – opposite-space.
One exception is: BORBOR – central-point.

The relation of a part of space to a subject is expressed in Kyrgyz by means of İLİK – possessive case.

Examples: The butterfly is over the table:

KÖPÖLÖK ÜSTÖLDÜN ÜSTÜNDÖ - (literatim) the butterfly (is) in the table's upper-space.

Put the ball under the table:

TOPTU ÜSTÖLDÜN ASTINA KOY - (literatim) put the ball to the table's lower-space.

Second peculiarity is that the above-mentioned parts of space are fuzzy [1] what was demonstrated by experiments conducted by us.

Time as an objective form of matter existence is closely related to the space category. For instance, if we refer to the connection between the abstract concept of *time* and such realities as an *hour* or *month*, then language draws them as form and content: *hour, day* or *month*.

In Kyrgyz language the Locative (CATIŞ) case and some other grammatical forms are applied similarly both to space, to time and to circumstances:
KIŞINDA - in winter;
ANIN ALDINDA - before him;
UKKANDA – while listening.

It stresses similar conception of space and time reflected in the language.

Kyrgyz language provides measuring distances by time, for examples:

TÜŞTÜK COL - (literatim) midday way - (an average) distance a rider passes from dawn till midday;

KÜNDÜK COL - (literatim) daily way - (an average) distance a rider passes from dawn till evening.

"... BEŞ KÜNDÜK CERGE UGULDU" - (literatim) ... was listened [as far as] to five-daily ground" (epos "Manas").

ÜÇ AYLIK COL - (literatim) three-monthly way - (in opposition to preceding examples) a distance a caravan passes slowly with overnight stays etc.

The word KERE (fully) is also used (in the sense provided in [1]):

KERE KÜNDÜK COL - (literatim) maximal daily way - the distance a rider can pass during a day (epos "Manas").

On this base, the definition of a new mathematical object - kinematical space X (the metric $p_k(x,y)$ is the minimal time of passing between points x and y) with corresponding axioms was introduced and implemented in [2]:

Definition. A pair: a set X of points and a set K of routes is said to be a kinematic space (each route M, in its turn, consists of the positive real number T_M (time of route) and the function $m_M : [0, T_M] \to X$ (trajectory of route)) if the following conditions are fulfilled:

(K1) For each different $x, y \in X$ there exists such $M \in K$, that $m_M(0) = x$ and $m_M(T_M) = y$, and the set of values of T_M for all such M is bounded with a positive number from below (infinitely fast motion is impossible).

(K2) If $M = \{T_M, m_M(t)\} \in K$ then the pair $\{T_M, m_M(T_M - t)\}$ is also a route of K (the reverse motion is possible).

(K3) If $M = \{T_M, m_M(t)\} \in K$ and $T^* \in (0, T_M)$ then the pair: T^* and function $m^*(t) = m_M(t)$ $(0 \le t \le T^*)$ is also a route of K (one can stop at any desired moment).

It had justified established perception and aims of work at computer: passing from one image on display to other one as fast as possible.

We propose to search and examine all mentions of measuring distance by time in literature as a constituent of investigation of presentation of world in Kyrgyz language.

References

[1] Zadeh L.A. (1975) The concept of a linguistic variable and its application to approximate reasoning. Information Sciences, Vol. 8, pp. 199-249, 301-357; Vol. 9, pp. 43-80.

[2] Borubaev A.A., Pankov P.S. (1999) Computer presentation of kinematic topological spaces. The Kyrgyz State National University, Bishkek, 131 p. (in Russian).

[3] Karabaeva S., Dolmatova P. (2014) Mathematical and computer models of spatial relations in Kyrgyz language. Proceedings of V Congress of the Turkic World Mathematicians / Ed. by Acad. A. Borubaev. -Bishkek: Kyrgyz Mathematical Society, pp. 175-178.

[4] Karabaeva S. (2015) Peculiarities of spatial relations in Kyrgyz language. Abstracts of the Issyk-Kul International Mathematical Forum (Kyrgyzstan, Bozteri, 24-27 June, 2015) / Ed. by Acad. A. Borubaev. -Bishkek: Kyrgyz Mathematical Society, p. 79.

СОЗДАНИЕ БАЗЫ ДАННЫХ ЛЕКСИЧЕСКИЙ ФОНДА ТУВИНСКОГО ЯЗЫКА[1]

Б.Ч. Ооржак, А.Б. Хертек, М.А. Кужугет, А.Я. Салчак, В.С. Ондар, Е.Т. Чамзырын
Тувинский государственный университет
oorzhak.baylak@mail.ru

В статье описывается проводимая работа по подготовке базы данных лексического фонда тувинского языка для Электронного корпуса текстов тувинского языка. Электронная база данных лексического фонда тувинского языка будет представлять собой справочно-поисковую систему, при помощи которой можно будет находить в тексте необходимые для целей пользователя фрагменты с искомым значением.

***Ключевые слова**: тувинский язык, база данных, лексический фонд, лексико-семантические классы, лексико-семантические подклассы, лексическая сочетаемость, автоматизированная система поиска.*

Творческим коллективом в составе научных сотрудников и преподавателей Тувинского госуниверситета продолжается работа над разработкой Электронного корпуса текстов тувинского языка (ЭКТТЯ), начатого в 2011 г. при поддержке гранта Российского гуманитарного научного фонда. В настоящее время в ЭКТТЯ продолжает пополняться текстами тувинской художественной литературы и фольклора разных жанров в электронном формате. На данном этапе ведется работа по совершенствованию морфологической разметки корпуса. Параллельно с этим началась работа по разработке семантической разметки, которая позволит автоматизировать поиск необходимой семантической информации из текстов. Первым этапом работы в данном направлении является создание электронных баз данных по лексемам тувинского языка. Электронная база данных лексического фонда тувинского языка будет представлять собой справочно-поисковую систему, при помощи которой можно будет находить в тексте необходимые для целей пользователя фрагменты с искомым значением.

Разрабатываемая электронная база данных основывается на распределении всех полнозначных лексем тувинского языка на семантические разряды (классы) слов. Условно выделены в четыре базовые семантические классы: ЧЕЛОВЕК, ЖИВОТНОЕ, ПРЕДМЕТ, ПРИРОДНЫЕ ОБЪЕКТЫ И ЯВЛЕНИЯ, которые далее подразделяются на более

[1] Работа выполнена в рамках научно-исследовательских работ по проекту РГНФ №16-04-12020 «Создание базы данных лексического фонда тувинского языка».

дробные лексико-семантические подклассы. Лексико-семантические классы, подклассы и дескрипторы обозначаются тэгами на тувинском, русском и английском языках. Например, имена прилагательные разделяются на подклассы качественных и относительных. Пример распределения качественных имен, характеризующих человека приведен в нижеследующей таблице:

Таблица 1

Качественные имена, характеризующие человека

Кижи/Человек/Human		
Мага-боттуң шынарлары/ Физические качества / Physical quality	Мага-бот/ Тело, телосложение/ Body	*мөге* ‘большой и сильный’, *семис*‘полный, толстый’, *тырың* ‘крепкий, плотный’, *эът-ханныг* ‘здоровый, в теле’, *ээлгир* ‘гибкий’, *ыспан/ыспагар* ‘худой, тощий’.
	Дурт-сын/ Рост/ Growth	*бедик* ‘высокий’, *биче* ‘маленький’, *чавыс* ‘низкий’, *чавыссымаар* ‘низенький’, *чолдак* ‘невысокий, короткий’.
	Даштыкы хевир/ Внешность/ Appearance	*арыг-силиг* ‘чистоплотный’, *силиг* ‘аккуратный’, *чараш* ‘красивый’, *шевергин* ‘с правильными чертами лица’, *хөрлуг* ‘видный, представительный’.
	Баш/Голова/ Head	*течик/течигир* ‘с выпуклым затылком’, *дөңгүр* ‘лысый’, *тас* ‘лысый’, *моң* ‘с большой головой’.
	Баш дүгү/ Волосы/ Hair	*өгбегер/агбагар* ‘растрепанный’, *дыдыраш* ‘кудрявый’.
	Арын/Лицо/ Face	*дырышкак* ‘морщинистый’, *хылбаң* ‘худой, бледный’, *ыжык* ‘опухший, распухший’, *дүгдүнчек* ‘угрюмый’, *додуккан* ‘смуглый’.
	Карак/Глаза/ Eyes	*хапыгыр* ‘распухший, опухший’, *бүлдегер* ‘большой, блестящий’, *удумзургай* ‘сонный’.
	Кирбик/Брови/ Eyebrows	*чиңге* ‘тонкий’, *терең* ‘густой’, *дугаланчак* ‘дугообразный’.
	Кулак/Уши/ Ears	*делбиң/делбигир* ‘растопыренный’, *улуг* ‘большой’
	Думчук/Нос/ Nose	*коң/коңзагар* ‘нос крючком’, *кырлаң*‘прямой, правильной формы’.
	Чаак/Щеки/ Cheeks	*додуккан* ‘смуглый’, *бопугур* ‘круглощекий’, *хорлаңгы* ‘обветрившийся’.
	Эрин/Губы/ Lips	*дөрбегер* ‘большой’, *чиңгежек* ‘тонкий’.
	Хавак/Лоб/ Forehead	*кадыр* ‘крутой, высокий’.
	Хол/Руки/ Hand	*хаварык* ‘мозолистый, натёртый’, *хорлаңгы* ‘обветрившийся’.
	Оорга/Спина/ Back	*бүшкүгүр* ‘горбатый’, *доңгагар* ‘сутулый’, *ыргак* ‘сутулый’, *хөкпек/хөкпегер* ‘сутулый’.

	Бут/Ноги/Foot	*дойтуксумаар* ‘прихрамывающий’, *майышкак* ‘косолапый’, *майтак* ‘косолапый’.
	Ижин/Живот/ Abdomen	*дөртегер* ‘вздутый’, *хөртегер* ‘пузатый, брюхатый’, *шөртек/шөртегер* ‘пузатый, брюхатый’.
Угаан-медерел талазы-биле шынарлары/ Умственные качества/ Mental quality	Эки шынарлар/ Положительные Качества/ Positive qualities	*баштыг* ‘умный’, *угаанныг* ‘умный’, *бижик-биликтиг* ‘грамотный’, *бодангыр-сагынгыр* ‘сообразительный’, *сагынгыр* ‘сообразительный, смышленный’, *сарыылдыг* ‘разумный’.
	Багай шынарлар/ Отрицательные качества/ Negative qualities	*мелегей* ‘глупый’, *сээдең* ‘тупой’, *тудуу* ‘слабоумный’, *үзээргей* ‘глупый’, *ээдергей* ‘глупый’, *ээдергейзимээр* ‘глуповатый’
Сагыш шынарлары/ Психические качества и характер/ Mental qualities and character	Эки шынарлар/ Положительные Качества/ Positive qualities	*дидим* ‘смелый’, *биликсээчел* ‘любознательный’, *болгаамчалыг* ‘острожный’, *дөспес* ‘неспокойный’, *дузааргак* ‘отзывчивый’, *бүзүрээчел* ‘доверчивый’, *хайгаараачал* ‘наблюдательный, внимательный’.
	Багай шынарлар/ Отрицательные качества/ Negative qualities	*адыыргак* ‘тщеславый’, *былдаачал* ‘увиливающий, уклоняющийся’, *тураараргак* ‘высокомерный’, *туразында* ‘своевольный’, *туралыг* ‘своенравный’, *турамык*’ ‘высокомерный’, *тураараргак* ‘высокомерный’, *хараадаачал* ‘сожалеющий’, *хедер* ‘грубый’, *хирээннээчел* ‘сердитый, надутый’.
Сеткил шынарлары/ Душевные Качества/ Mental qualitys	Эки шынарлар/ Положительные качества/ Positive qualities	*ажык* ‘доброжелательный’, *чазык* ‘приветливый, открытый’, *амыр-мендилиг* ‘приветливый’, *арга-сүмелиг* ‘участливый’ *ажаанзырак* ‘заботливый’ *баштак* ‘веселый, любящий пошутить’.
	Багай шынарлар/ Отрицательные качества/ Negative qualities	*кортук* ‘трусливый’, *дерзии* ‘жестокий’, *хедер* ‘упрямый’, *чалгаа* ‘ленивый’, *халамыргай* ‘вялый, апатичный, слабый’, *аажылыг* ‘дерзкий’, *арга-*мегелиг ‘хитроумный’, *бак* ‘плохой’, *ылчың* ‘несерьёзный, легкомысленный, игривый, кокетливый’, *шыпыраңнааш* ‘вертлявый’.
Мөзүг-шынар/ Нравственные Качества/ Moral quality	Эки шынарлар/ Положительные качества/Positive qualities	*шынчы* ‘честный’, *томаанныг* ‘смирный’, *топтуг* ‘порядочный’, *топтуг-томаанныг* ‘порядочный’ *төлептиг* ‘достойный’, *чазыылдыг* ‘выдержанный, дисциплинированный’, *хумагалыг* ‘бережливый.
	Багай шынарлар/ Отрицательные качества/ Negative qualities	*мегечи* ‘нечестный’, *тоожок* ‘безнадёжный’, *байбаң* ‘болтливый’, *байыыргак* ‘кичащийся своим богатством’, *балалыг* ‘причиняющий вред’, *бараа* ‘странный’, *бачым* ‘спешный’, *бачыттыг* ‘грешный’, *бууулуг* ‘виноватый’, *ёзуургак* ‘манерный’, *ёзажок* ‘неразумный’.
Ниитилелде шынарлары/ Социальные Качества/ Social qualities	Хар-назынының аайы-биле шынарлар/ По возрастному признаку/ Age symptom	*чаш* ‘младенческий’, *чалыы* ‘молодой, юный’, *ушпа* ‘дряхлый, престарелый’, *хеймер* ‘самый младший в семье’,*хензиг* ‘маленький’, *хоочун* ‘старый’, *элээди* ‘подросткового возраста’, *ылбыс* ‘новорождённый’, *чөнүк* ‘дряхлый, престарелый’, *чедишкен* ‘достигший сорока пяти-пятидесяти лет’.

	Өнчү-хөреңги аайы-биле шынарлар/По имуществен-ному признаку/ By ownership	*бай* ‘богатый’, *тодуг* ‘сытый, богатый’, *тодуг-догаа* ′зажиточный, богатый, обеспеченный’, *түреңги* ‘нищий’, *чединмес* ‘нуждающийся’, *ядыы* ‘бедный’, *ядамык* ‘бедный’.
	Ниитилелде, өг-бүледе кижиниң туружунуң аайы-биле шынарлар/ Признание / не признание в обществе, семье, коллективе/ Recognition / recognition in society, the family, the community	*барыктыг* ‘сносный’, *үлегерлиг* ‘образцовый, примерный’, *чааскаан* ‘одинокий’, *чаңгыс* ‘одинокий’, *хүндүлүг* ‘уважаемый’, *хүндүткелдиг* ‘почётный, авторитетный’, *хүндүтен*‘почтенный’, *ынак* ‘любимый’, *эп-найыралдыг* ‘дружный, спаянный’, *эп-сеткилдиг* ‘единодушный, спаянный, дружный’.
Салым-чаяан/Способности, таланты/Abilities, talents		*салымныг* ‘способный, одарённый, талантливый’, *шевер* ‘искусный, умелый’, *чаяанныг* ‘способный, одарённый, талантливый’.

Кроме того, создаваемые базы данных полнозначной лексем тувинского языка будут служить для выявления их лексической сочетаемости. Так, в автоматизированную систему будут включены семантически допустимые сочетания лексем. Например, допустимые сочетания имени существительного и имени прилагательного:

	дээр ‘небо’
көк ‘синий’	*аржыыл* ‘платок’,
	бажың ‘дом’

Автоматизированной системой будут исключены семантически недопустимые сочетания, например: *көк аът* ‘синяя лошадь’. Таким образом, создание разных пользовательских запросов с учетом семантики позволит уточнить, выявить правила сочетаемости тех или лексических единиц.

УДК 004.912

ЭЛЕКТРОННАЯ БАЗА ДАННЫХ АТЛАСА РУССКИХ ГОВОРОВ

А.Г.Пилюгин, Ф.И. Салимов., В.Д. Соловьев
Казанский федеральный университет, Казань
pag@kcn.ru, Farid.Salimov@kpfu.ru, maki.solovyev@mail.ru

В статье рассмотрены вопросы, связанные с созданием электронной базы данных диалектологического атласа русских говоров (ДАРЯ).

Ключевые слова: *диалекты, русский язык, ДАРЯ, базы данных*

Одним из важных аспектов исследования языковых диалектов является изучение зависимости языковых явлений от их территориального размещения. Общая среда обитания формирует определенные устойчивые связи между языками народов, населяющих данное географическое пространство, часто объясняет сходство или различие языковых явлений, которые относятся к различным языковым группам, позволяет понять характер и скорость различных изменений, происходящих в языках и диалектах. Сбор информации для фиксации соответствующих языковых явлений проводится по заранее определенной единой программе исследований в рамках заданной географической территории в течение большого отрезка времени. На базе анализа собранных материалов издаются диалектологические атласы различных языков, которые включают в себя множество карт, демонстрирующих распространение определенных языковых явлений в рамках заданных территорий. Диалектологические атласы представляют своеобразные базы данных, отражающие зависимость языковых явлений от географического положения населенных пунктов, в которых они наблюдаются.

Объединение в едином виртуальном пространстве географической и лингвистической информации позволяет ставить и решать математические задачи изучения структуры элементов этого пространства, описывать меры сходства и различия системы лингвистических признаков, характеризующих различные географические точки. Подобный подход в настоящее время приобрел достаточно широкую популярность в западных странах под названием диалектометрия. Такие исследования, прежде всего, касаются задач описания диалектного членения языков (задачи кластеризации), описания географических границ (изоглосс), разделяющих географические зоны (ареалы) распространения диалектов и говоров, выявление системы базовых признаков, наиболее значимых при

построении таких границ, описание корреляционных связей между разными языковыми явлениями. Математические исследования подобного рода особенно необходимы при решении задач большой размерности, где объемы имеющейся информации не позволяют вручную обработать большие наборы данных. Конечно, лингвисты, решая подобные задачи вручную, упрощают набор данных, выделяют и подвергают обработке главные с точки зрения исследователя компоненты. Но при этом возникают вопросы обоснованности вводимых ограничений, а также вопросы оценки точности полученных решений. Исследования по диалектометрии проводятся в ряде Европейских стран: в частности выполнены для диалектов болгарского, голландского [7-9] и др. языков. В [10] приведен обзор современного состояния диалектометрии в мире.

В течение последних двух лет в Казанском Федеральном университете совместно с лингвистами Института русского языка РАН реализуется проект создания компьютерной фактографической базы данных по говорам русского языка. Диалектологический атлас русского языка (ДАРЯ) создавался, начиная с 40-х годов XX века, усилиями большого числа исследователей и географически охватывает территорию центральных областей Европейской части России. Атлас опубликован в виде трех альбомов и трех книг сопроводительных материалов, в которые вошли сведения по фонетике, морфологии и синтаксису русского языка [1-4]. Лингвистическая информация атласа собиралась по специальной разработанной программе [5] и охватывала 294 вопроса по различным разделам языка. Обследование проводилось в специально отобранных 4209 населенных пунктах, местоположение которых образовывала относительно равномерную сетку на карте Европейской части России с шагом 15 км.

Каждая карта атласа содержит информацию по распределению значений определенных языковых явлений по населенным пунктам, в которых проводилось анкетирование. В соответствие с этим база данных разбивается на две части: картографическую, содержащую информацию о населенных пунктах, и атрибутивную, содержащую информацию по распределению значений языковых явлений по населенным пунктам. Подобная база данных для диалектов и говоров татарского языка была создана в 2012 году [11]. Для ДАРЯ электронная база данных была создана под руководством Пшеничновой Н.Н. [6]. Однако в силу имевшихся в то время различных причин не использовалась реляционная модель баз данных, для представления данных применялась оригинальная кодировка, которая на настоящий момент времени утеряна. Это обстоятельство не позволяет использовать результаты Пшеничновой и ставит задачу создания базы данных ДАРЯ заново.

Картографическая часть базы данных должна содержать информацию по населенным пунктам, в которых производился сбор информации, их географических координат, административного подчинения. База данных может также содержать дополнительную информацию по истории и этнографии, по национальному и количественному составу населения в этих населенных пунктах.

Создание картографической базы данных требует точной локализации положения каждого населенного пункта на карте. К сожалению, подобная информация в опубликованных книжных изданиях атласов отсутствует: обычно в комментариях к атласу публикуется только список наименований обследованных населенных пунктов, их административная подчиненность, без указания точных географических координат. Такое положение дел было связано с тем, что в Советское время географические координаты населенных пунктов составляли предмет государственной тайны и не могли быть опубликованы в открытой печати. Поэтому одной из задач при создании картографической базы данных была реконструкция списка населенных пунктов, приведенного в комментариях к атласу. Эта задача осложнялась длительным отрезком времени, прошедшим со времени первых полевых экспедиций. За более чем полувековой период произошли значительные изменения в составе населенных пунктов, поменялись их наименования, административная принадлежность, произошло слияние некоторых населенных пунктов, многие населенные пункты пришли в запустение и исчезли из карт соответствующих регионов. Кроме того, некоторые регионы содержат в своем составе до десяти населенных пунктов с одним и тем же наименованием, что сильно затрудняет их географическую локализацию. Особенно сложные ситуации, связанные с определением географических координат возникают при изменении административного подчинения населенных пунктов.

Для проверки и уточнения списка населенных пунктов были просмотрены существующие базы данных в Интернет. База данных http://www.bankgorodov.ru представляет открытую энциклопедию регионов, муниципальных образований и населенных пунктов России. В этой базе данных дана подробная характеристика административного устройства Российских регионов. К сожалению, многие позиции в базе данных, такие как географические координаты, численность населения, историческая и этнографическая информация для многих населенных пунктов остаются незаполненными. Другим полезным источником является электронный архив старых карт населенных пунктов Российской Федерации, расположенный по адресу http://www.etomesto.ru. База данных позволяет вести поиск географических координат населенных

пунктов, на картах, изданных несколько десятков лет тому назад. Такой ресурс особенно полезен для поиска населенных пунктов, исчезнувших с современных карт. В качестве дополнительных источников можно использовать сайт http://foto-planeta.com, в котором наряду с видовыми фотографиями населенных пунктов, содержатся сведения об их географическом положении, сайт-путеводитель http://www.esosedi.ru/, содержащий различную географическую информацию, сайт http://uistoka.ru/, содержащий информацию о исторических местах Российских регионах, различные электронные энциклопедии https://ru.wikipedia.org/, http://wikimapia.org/. К сожалению, многие из перечисленных источников формируются коллективными усилиями неквалифицированных пользователей, не имеют официального статуса, часто приводимая в них информация является неполной и противо-речивой, не отслеживается ее актуальность. Очень немногие регионы имеют официальные источники информации.

Электронные базы данных, по сравнению с их книжным аналогом обладают расширенными возможностями по представлению информации о населенных пунктах. Картографическую часть базы данных можно рассматривать как распределенную, если включить в нее ссылки на информацию, которая хранится в различных ресурсах, размещенных в сети ИНТЕРНЕТ. Подобные ссылки могут указывать на исторические, этнографические, лингвистические материалы, которые хранятся в различных базах данных и представляют интерес для исследователей диалекта и языка. К сожалению, примеров систематического описания для отдельных территорий немного. Тем не менее, очень хороший ресурс по истории и этнографии Брянской области содержится на сайте http://www.kray32.ru, некоторые интересные факты по населенным пунктам Архангельской области содержатся по адресу http://www.russia29.ru/, по населенным пунктам Владимирской области – по адресу http://vladimirskaya-rus.ru/. Понятно, что создание подобных ресурсов дело хлопотное, связанные с большими затратами, но формирование таких энциклопедических баз данных, как диалектологические атласы для отдельных языков, может в значительной мере активизировать эти процессы

Атрибутивная база данных представляет собой отображение множества языковых явлений на множество обследуемых населенных пунктов. Первичная информация собиралась в течение длительного промежутка времени (более 50 лет) большим коллективом лингвистов по программе сбора сведений для составления диалектологического атласа, принятой в 1945 году. К сожалению, исходная картотека данных во время пожара в ИРЯ РАН была утеряна. В этих условиях при реконструкции атрибутивной части базы данных приходится ориентироваться на вторичные данные, которые опубликованы в картах атласа. Отметим, что

каждая карта ДАРЯ представляет результат определенной обработки первичных данных, в них значения лингвистических признаков приписаны не населенным пунктам, а целым областям, включающим по нескольку населенных пунктов. Понятно, что подобная факторизация данных может сказаться на точности обработки результатов анализа.

Для считывания информации из карт атласа была разработана специальная процедура, которая сканирует, а далее анализирует информацию результатов сканирования. При этом приходится решать задачи, связанные с идентификацией географических координат границ приведенных на картах областей, с определением типа использованной проекции на картах атласа.

К настоящему времени завершено создание атрибутивной базы данных по первому выпуску ДАРЯ (Фонетика). Просмотрена и отредактирована база данных из 2500 населенных пунктов (всего их – 4209).

Благодарности: Работа выполнена при финансовой поддержке РГНФ (проект № 15-04-12008)

Литература

1. Диалектологический атлас русского языка. Центр Европейской части СССР. Выпуск I: Фонетика / Под ред. Р. И. Аванесова и С. В. Бромлей. — М.: Наука, 1986.

2. Диалектологический атлас русского языка. Центр Европейской части СССР. Выпуск II: Морфология / Под ред. С. В. Бромлей. — М.: Наука, 1989.

3. Диалектологический атлас русского языка. Центр Европейской части России. Выпуск III: Карты (часть 1). Лексика. — М.: Наука, 1997.

4. Диалектологический атлас русского языка. Центр Европейской части России. Выпуск III: Карты (часть 2). Синтаксис. Лексика. — М.: Наука, 2005.

5. Программа собирания сведений для составления диалектологического атласа русского языка. М.-Л., 1947.

6. Пшеничнова Н.Н. Типология русских говоров. М.: Наука. 1996. 208 с.

7. John Nerbonne and Peter Kleiweg. Lexical Distance in LAMSAS. In: John Nerbonne and William Kretzschmar (eds.) Computational Methods in Dialectometry. Special issue of Computers and the Humanities, 37(3), 2003, 339-357.

8. *Nerbonne, J. y Kretzschmar, W., 2003 , "Introducing Computational Techniques in Dialectometry" ,* Computers and the Humanities *, vol. 37 , pp. 245-255 .*

9. Houtzagers, Peter , Nerbonne, John and Prokić, Jelena (2010) 'Quantitative and Traditional Classifications of Bulgarian Dialects Compared', Scando-Slavica, 56: 2, 163 — 188

10. John Nerbonne and William Kretzschmar, Jr.Dialectometry++. LLC: Journal of Digital Scholarship in the Humanities 28(1), 2013, pp.2-12. doi:10.1093/llc/fqs062

11. Салимов Ф.И., Рамазанова Д.Б., Пилюгин А.Г., Салимов Р.Ф. Электронная версия атласа татарских народных говоров // Вестник ТГПУ 4(26), 2011, Казань, изд-во КФУ, с.205-210.

УДК 004.912

ЭТНОЛИНГВИСТИЧЕСКИЙ ЭЛЕКТРОННЫЙ СЛОВАРЬ ТЕРМИНОВ ТАТАРСКОГО ЯЗЫКА

Ф.И. Салимов., Р.Ф. Салимов
Казанский федеральный университет, Казань
Farid.Salimov@kpfu.ru, Rust1k@gmail.com

В статье описан опыт создания электронного этнолингвистического словаря, построенного на базе материалов, собранных учеными ИЯЛИ АН РТ во время полевых экспедиций.

Ключевые слова: *этнолингвистика, диалекты, татарский язык, базы данных*

При создании лингвистических электронных ресурсов большое значение имеет коллективный опыт исследований различных архивных материалов, собранных в течение длительного времени, и опубликованный в виде традиционных форм представления информации: книг, словарей, различного рода научных публикаций. Конечно, идеальным вариантом при создании электронных баз данных представляется анализ и размещение материалов, почерпнутых из первичных источников информации, которые хранятся в виде бумажных картотек и в сжатом виде представляют экспериментальные данные, собранные коллективами ученых в различных лингвистических экспедициях. К сожалению, такие архивы по разным причинам являются малодоступными, подвержены различным внешним воздействиям, имеют высокую вероятность разрушения. В качестве таких примеров можно упомянуть картотеку данных русского диалектологического атласа (ДАРЯ), которая в большей своей части была утеряна в результате проведения ремонтных работ в институте Русского языка [1]. Подобная участь постигла картотеку атласа татарских народных говоров [2]. Одной из причин такого положения дел является падение интереса к сохранности исходных данных после их первичного анализа и выхода в свет книжных публикаций, в которых отражены результаты научных исследований.

При отсутствии первичных архивов приходится опираться на вторичные источники в виде опубликованных научных трудов, в которых исходные данные подвергнуты определенному анализу. Несмотря на то, что такой подход является менее информативным и содержит определенную субъективную окраску авторов публикаций, он имеет и некоторые положительные моменты, поскольку при отборе материалов в базу

данных могут быть учтены результаты его первичной обработки (особенно, если она сделана квалифицированно и качественно).

Начиная с конца 50-х годов XX столетия в институте языка, литературы и искусства им. Г. Ибрагимова Академии наук Республики Татарстан (ИЯЛИ АН РТ) проводится работа по сбору этнолингвистического материала по диалектам и говорам татар, проживающих в Республике Татарстан и других регионах России. Материалы собираются в архаически этнокультурном отношении диалектных зонах Сибири, регионах Урала, Среднего и Нижнего Поволжья. На основе анализа и систематизации собранных данных сотрудниками ИЯЛИ в различные годы было опубликовано более 20 монографий и около 300 научных статей.

В данной статье описан опыт по созданию электронного этнолингвистического терминологического словаря, включающего в себя в структурированном виде информацию, извлеченную из основных публикаций, изданных по результатам этнолингвистических экспедиций ИЯЛИ АН РТ. Источниками для словаря были выбраны объемные фундаментальные научные труды Баязитовой Ф.С., составляющие семейный цикл [3,4,5]. Тематика этих книг включает в себя описание лексики, обычаев, обрядов, связанных с рождением ребенка, свадебной церемонией, со смертью и погребением. Все книги Баязитовой изданы на татарском языке, имеют сходное строение, их содержание построено на тематико-гнездовом принципе, для большинства терминов дано их семантическое описание. Кроме того, каждая книга содержит большое количество образцов живой речи татар, которые приведены в качестве иллюстративного материала со ссылкой на диалекты и говоры, в которых они встречаются. Это обстоятельство, безусловно, повышает информативность публикаций и позволяет использовать их в виде источников для создания электронных информационных ресурсов.

Создание татарского этнолингвистического словаря на основе опубликованных печатных материалов состояло из нескольких этапов и включало в себя решение ряда задач:

1. Сканирование печатного материала книг с целью получения электронного образа;
2. Анализ и систематизация имеющихся в книгах материалов;
3. Сегментация содержимого книг с целью выделения фрагментов, которые составляют содержание статей электронного словаря;
4. Заполнение базы данных, создание библиотеки запросов;
5. Создание клиентской части программы с размещением ее в сети Интернет.

1. Реализация первого этапа носила в основном технический характер. Поскольку предполагалась автоматическая обработка полученного электронного образа, то особое внимание уделялось выбору формата хранения электронных данных, Одна из основных задач при конвертировании состояла в стремлении максимально сохранить стиль оформления текста сканируемого материала с целью его дальнейшего использования процедурами обработки текста. Наиболее подходящим оказался формат текстового процессора WORD. Поскольку существующие сканеры не всегда справлялись с поставленной задачей, потребовался просмотр и корректировка результатов сканирования в ручном режиме. С учетом объема исходного материала эта часть работы была достаточно затратной и потребовала больших усилий.

2. Книги Ф.С. Баязитовой имеют определенную структуру: любая книга состоит из многочисленных тематических групп (гнезд), в каждой из которых исследуется и систематизируется определенный набор терминов, относящийся к разделу раскрываемой темы. Содержание гнезда характеризует определенные ритуалы, связанные с рождением ребенка; описание лиц, принимающих участие в совершении обряда; описание предметов, которые используются при выполнении обрядовых действий; сами обрядовые действия; поверья, свадебные или погребальные ритуалы, характеристики персонажей, принимающих участие в церемониях, описание свадебной пищи, свадебной одежды, и пр. Каждое такое гнездо занимает определенный фрагмент в тексте книги. При этом границы соседних фрагментов порой сильно размыты и явно не выделены. Основная задача анализа текста книги состояла в выявлении системы признаков, характеризующих различные фрагменты текста с дальнейшим автоматическим выделением частей текста, относящихся к определенному термину. При анализе выделялся сам термин, его семантическое описание при его наличии в тексте, ссылки на диалекты и говоры, в которых этот термин используется, множество примеров употребления термина в различных диалектах.

3. Описание семантики терминов, приводимых в книгах Баязитовой, даже для опытных лингвистов представляет достаточно непростую задачу. Известно, что в этнолингвистических терминах отражается определенная «картина мира», которая формируется в этносе; в терминах прослеживается связь языка с элементами народной культуры, всех ее жанров и форм. Поэтому содержание того или иного термина может быть описано только в контексте определенных явлений, событий. Тем не менее, в рамках проекта по созданию терминологического словаря с учетом объема опубликованного материала была предпринята попытка предварительной автоматической обработки имеющихся текстов с последующей ручной корректировкой границ выделенных фрагментов. Была создана

процедура сегментации текста с выделением ключевых терминов, фрагментов, описывающих их семантику, выделения примеров употребления термина в различных диалектах. При разбиении текста на фрагменты процедура ориентировалась в основном на формат оформления соответствующих частей текста-источника, с дополнительным анализом материала на присутствие некоторых ключевых слов, приведенных в справочниках. К сожалению, этот подход позволял лишь приблизительно определять границы выделяемых фрагментов. Причина такого положения дел в первую очередь состояла в различии стилевого оформления одинаковых по смыслу фрагментов даже в пределах одного источника, не всегда поддерживался единый язык разметки различных частей текста, нередко в приводимых примерах информация по диалектам и говорам была неполной. При написании книг не предполагалась обработка текстов в автоматическом режиме, они, прежде всего, были написаны для специалистов. Поэтому после программного выделения фрагментов, точность определения их границ проверялась лингвистом, при обнаружении ошибок производилась ручная корректировка границ. Ниже приведен фрагмент промежуточной рабочей таблицы, которая сформирована в результате работы процедуры сегментирования.

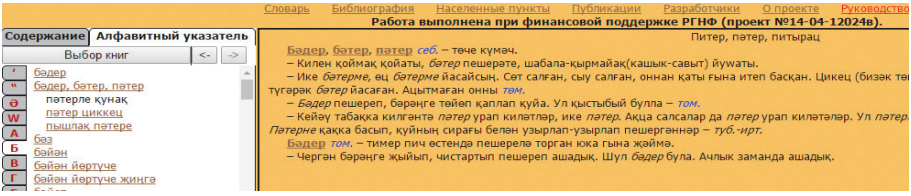
3	себ.	**Айшық**	**Айшық** *себ.* – гыйшык.
3	себ.	**Айшық уты**	**Айшық уты** *себ.* – гыйшык, мәхәббәт уты.
4	төм.		Қапқадин қарадым сәне, *Айшық уты* йандыра мәне – *төм.*
3	миш. д.	**Айшиклык йырлары**	**Айшиклык** [гар. айшик + афф. -лык] **йырлары** *миш.д.* – гыйшык, мәхәббәт җырлары. **Ашик тоткан кызлар** *чпр.* – гыйшык тоткан кызлар.
4	чпр.		– Кич утырганда *ашиклык йырлары* йырлыйбыз, кулга йаулык тотып. *Ашик тоткан кызлар* бик хуwат йырлыйлар ашиклык йырла--- рын– *чпр.*
4	күрш.		Әти дә кәбәмнең миленчасы Сәтенә йез пыт он тартадыр. Синдәй матурымны күргән сайын Йандин ашиклыгым артадыр – *күрш.*

В этой таблице третий столбец определяет выделяемый термин, второй столбец характеризует диалект, четвертый столбец содержит или семантическое описание термина, или примеры употребления термина в живой речи (характер записанной в строке информации определяется кодом, записанным в первом столбце). Зеленым цветом в тексте выделены диалектизмы, которые носят вспомогательный характер и расшифрованы в сносках книг. Эти термины были выделены программой в отдельный словарь.

4. В результате работы была создана база данных терминологического словаря, включающего в себя этнолингвистические термины, их описания, многочисленные примеры употребления терминов в живой речи татар с указанием на диалекты, в которых они встречаются. Словарь был дополнен информацией по фонетической транскрипции диалектизмов, а также для части терминов словаря, которые относятся к родильной и свадебной тематики был осуществлен перевод семантики на русский язык. При создании базы данных используемая в книгах Ф.С. Баязитовой облегченная транскрипция, была признана недостаточной: для каждого термина была предложена транслитерация терминов на письменные формы татарского языка с использованием символов Международного фонетического алфавита (МФА) [6]. При этом преследовалась цель использования терминов словаря в диалектических корпусах. К настоящему времени общий объем словаря составляет около 6000 словоформ и словосочетаний.

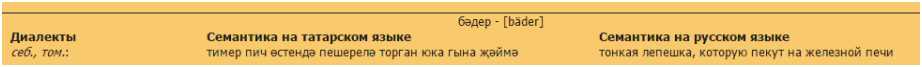
Дополнительно в базу данных включена информация по населенным пунктам, где собиралась информация, по информантам, по библиографии, использованной при создании книг. Кроме того, была реализована связь между терминами построенного словаря и картами атласа татарских народных говоров [7]. Такой результат, не давая точной картины географии распространения термина, используя диалектное членение татарского языка, позволяет приблизительно определить населенные пункты, где может употребляться данный термин.

5. Программа размещена в сети Интернет по адресу ethnoling.antat.ru. Ниже на рисунке показан вид основного экрана программы.



В левой части экрана визуализируется список терминов, начинающихся на определенный символ, выбираемый пользователем. При этом пользователь также может выбирать источник (книгу), может выбирать способ представления информации (тематический в виде определенных гнезд или алфавитный). В правой части экрана показывается подробная информация по выбранному термину - его семантика, примеры употребления в языке, указания на диалекты, в которых этот термин употребляется. Названия диалектов выделены синим цветом. Выбор диалекта мышью позволяет активизировать карту регионов

РФ с визуализацией географического расположения населенных пунктов, в которых распространен указанный диалект. В нижней части экрана приведены фонетическая транскрипция термина, его семантическое описание на татарском и русском языках:



Создаваемый терминологический словарь может быть использован при обучении татарскому языку. Кроме того большой набор примеров употребления различных терминов в живой речи материал может служить ресурсом для создания диалектологических подкорпусов татарского языка, этнолингвистических словарей.

Благодарности: Работа выполнена при финансовой поддержке РГНФ (проект № 14-04-12024)

Литература

1. Диалектологический атлас русского языка. Центр Европейской части СССР. Выпуск I: Фонетика / Под ред. Р. И. Аванесова и С. В. Бромлей. — М.: Наука, 1986.

2. Атлас татарских народных говоров. 2-е изд. – доп./ Под ред. Д. Б. Рамазановой, Т. Х. Хайрутдиновой - Казань, ИЯЛИ, 2015 - 631 с.

3. Баязитова Ф.С. Туй йолалары лексикасы (җирле сөйләш һәм фольклор текстлары яссылыгында). – Казан: Паравитта, 2011. – 976 б.

4. Баязитова Ф.С. Халык традицияләре лексикасы: бишек туе (йола һәм фольклор текстлары яссылыгында). – Казан: Алма-Лит, 2012 – 331 б.

5. Баязитова Ф.С. Халык традицияләре лексикасы: соңгы туй (дини фольклор һәм җирле сөйләш текстлары яссылыгында). – Казань, 2015. – 710 б.

6. У.Ш.Байчура "Звуковой строй татарского языка", изд-во КГУ, 1959г., Ч. 1. 183 с.

7. Салимов Ф.И., Рамазанова Д.Б., Пилюгин А.Г., Салимов Р.Ф. (2012) Электронная версия атласа татарских народных говоров// Вестник татарского государственно-гуманитарного педагогического университета, Казань, с.205-210

УДК 81.25

НАРОДНЫЕ ТРАДИЦИИ СКВОЗЬ ПРИЗМУ ЯЗЫКА

Ф.С. Баязитова, Ф.И. Салимов, Л.Г. Хабибуллина
ИЯЛИ АН РТ, Казанский федеральный университет, Казань
fbajazit@gmail.com, farid.salimov@kpfu.ru, valievalg@mail.ru

В научный оборот вводится новый материал "памятники" народной культуры – свадебные обрядовые тексты. Речь идет о диалектологии языка и культуры, в частности о свадебной обрядовой терминологии и диалектно-фольклорных текстах. Материалы собраны и систематизированы по всем регионам, где компактно проживает татарское население.

Ключевые слова: *этнолингвистика, этнография, диалекты, татарский язык*

В настоящее время в мире наблюдается повышенный интерес к языку народной духовной культуры. Записи живой диалектной речи, собранные в полевых условиях представляют для исследователей языка ценнейший материал. Диалектологические тексты фиксируются либо вручную, либо с помощью различных записывающих устройств, которые позволяют точно воспроизводить многогранное своеобразие и богатый колорит народной речи. Такая речь, перенесенная на бумагу, является прекрасным иллюстративным материалом, используемым в диалектологических словарях, в электронных корпусах. Диалектологические тексты, собранные по определенной тематике, являются основным источником в изучении региональной обрядовой лексики, относящейся к традиционной культуре. На основе анализа таких текстов проводятся различные исследования диалектного языка на всех его уровнях. Именно на этих примерах можно наблюдать и анализировать конкретные языковые факты, на основании которых исследователь может судить об общеязыковых явлениях в системе татарского диалекта. Собранный в экспедициях языковой этнографический материал подтверждает довольно хорошую сохранность традиционной народной культуры в памяти татарского народа.

В татарском языке особым богатством обрядовой лексики выделяются говоры мишарского диалекта, говоры сибирских, астраханских, нукратских, касимовских, пермских и крещенных татар. При изучении собранных материалов основное внимание было сконцентрировано на

терминологии, которая является “частью плана выражения народной духовной культуры, а не просто фактом языка” [1,2].

Многочисленные экспедиционные записи, сделанные в различных регионах, где компактно проживает татарское население, позволяют восстановить корни одного из сложных по своему составу обрядов – народной свадьбы. Свадьба, как по своей значимости, так и по обрядовому содержанию занимает одно из центральных мест в жизни человека. Ни одна деревенская свадьба не обходится без широко разработанного ритуала, включающего в себя множество различных обычаев, берущих начало из глубины веков и связанных с языческими поверьями. Свадебный обряд долго и упорно хранил созданные многовековые традиции, одновременно допуская при этом сосуществование элементов разных эпох. Конечно, различные элементы свадебного обряда в разные исторические эпохи по-разному реагировали на социальные и общественные явления. Бурная ломка традиционной свадебной обрядности, особенно заметно начавшаяся со времен коллективизации, привела к тому, что многие звенья свадебного обряда практически прекратили свое существование, но сохранились в народной памяти, что позволяют исследователям и в настоящее время воссоздавать достаточно полное описание старинной свадьбы.

Экспедиционные материалы по традиционному свадебному обряду татарского народа позволяют делать интересные наблюдения и выводы, касающиеся структуры и территориальных особенностей самого обряда и обрядовых действий, участвующих в них лиц, реалий и т.д., особенностей распространения и функционирования свадебной терминологии. Подход к чисто этнографическому материалу сквозь призму языка со специальным вниманием к терминологии как к источнику реконструкции духовной культуры может быть эффективен только при условии, что эта терминология изучается системно [3].

При полевом исследовании традиционных обрядов большое внимание уделяется коллективному опросу, беседе одновременно с несколькими информантами, которые, отвечая на тот или иной вопрос, обычно дополняют и уточняют ответы друг друга. Коллективный опрос активизирует память информантов, быстрее настраивает их на разговор. Очень содержательными оказались записи обрядов от главных действующих лиц – исполнителей и участников. Во время экспедиций были выявлены наиболее знающие, талантливые знатоки обрядов и обычаев, говорящие богатым, сочным деревенским языком.

Лексика свадебных обрядов, как и всякая обрядовая терминология, может быть описана только в контексте всей обрядовой реальности. Параллельное исследование явлений языка и культуры в данном случае

особенно продуктивно: этнографические данные дополняют и завершают лингвистические, и наоборот лингвистические сведения дают прочную базу этнографическим исследованиям.

В собранных в экспедициях материалах содержатся явления из жизни народа, в них как в зеркале отображается его внутренняя история, народный быт, нравы, обычаи, поверья, разговорные диалоги и множество других этнографических и языковых сведений. Поэтому иллюстрированные материалы, записанные от информаторов старшего поколения, являются надежной и незаменимой базой, основой для выявления обрядовой терминологии и фиксации бесценных богатств народно-разговорной речи, которые могут в будущем обогатить литературный язык народными словами. Поскольку традиционный свадебный обряд состоит из народных обычаев, то свадебная обрядовая лексика включает в себя фольклорно-этнографические общеупотребительные и диалектные термины.

Приведем некоторые примеры из раздела «Свадебные персонажи». Как известно, в свадебных церемониях были специально назначенные люди, знатоки и распорядители всего торжества или отдельных его звеньев. Ниже приводятся названия некоторых свадебных персонажей, принимавших участие в свадебном обряде, как со стороны жениха, так и со стороны невесты. В качестве источников использованы экспедиционные записи, сделанные Ф.С. Баязитовой. Термины приводятся в той форме, в которой они встречаются в диалектах:

Аргыш – в заказанском говоре крещенных татар, а также – в чистопольском говоре крещенных татар означает: распорядитель, самый почетный гость на свадьбе; обычно в этой роли выступает дядя или тетя или родной брат невесты, а со стороны жениха – его друзья.

Безнең әнилəр аргыш ыйы, аргыш дигәнем –туй башлыгы инде. Аргышқа туйда җырлыйлар:

Бийекəй тауларның башында
Өлгерə микән сабан ашлыгы.
Сиңа җырламыйчы кемгə җырлыйм,
Син бит бу туйларның башлыгы.

Наши мамы были аргышем. Аргыш – это значит глава свадьбы. Аргышу поют:

На макушке высоких гор
Созревают ли яровые хлеба
Если не тебе, то кому не петь
Ведь ты же глава этой свадьбы.

В говоре пермских татар (*аргыщ*) – это приглашенные на свадьбу супружеские пары. Когда свадьба идет в доме невесты, таковыми обычно бывают родители, родственники жениха, и наоборот, когда свадьба идет

в доме жениха – родственники невесты. В других тюркских языках: Хакассия - *аргыс* (товарищ, друг); Ойрот - *аргыш* – товарищ, попутчик. В удмуртском языке: *аргыш* – один из устроителей обряда, почетная должность при большом жертвоприношении во время встречи весны [7]

Туй йегетләр, туй кызлар в нукратском говоре: парни и девушки, принимающие участие в свадьбе. *Туй йегетләр, туй кызлар булгалады туйда. Алар кызны йәрәшкәчен, сөлге биргәчен дә туйны хәстәрләмә йабышалар. Туй йегетләр туй алашаны кийендермә йабышалар, туй кызлар кызашка йермә йабышалар. (На свадьбе присутствовали (свадебные) парни и девушки. Они после сговора, после получения полотенца сразу начинали украшать свадебных лошадей, девушки начинали ходить на свадебные угощения)*

Известно, что границы между материальной и духовной сферами культуры весьма расплывчаты и размыты. Поэтому являются целесообразными и актуальными исследования материального оформления обрядов и обычаев, а также относящейся сюда лексики (обрядовая одежда, пища и т.д.). Она дает возможность увидеть то ценное, что выработано народом на протяжении веков и в какой-то мере может быть использовано в наши дни[4,5,6].

Собранные лингвистические и этнографические материалы показывают, что одежде отводилась важная роль в свадебных обрядах. Свадебная одежда была яркой и праздничной. В терминологии одежды и обрядов свадебного периода отображаются диалектные черты и особенности разных групп татарского народа. Одежда тесно связана с народными обычаями, с древнейшими истоками культуры. Народный костюм превратился в звено, которое связывает художественное прошлое нашего народа с его настоящим и будущим. Многие старинные наряды уже вышли из употребления, но в деревнях некоторые женщины бережно хранят их в сундуках как память о старших поколениях и иногда надевают эти наряды во время праздников, свадеб. Иными словами, функции этих вещей изменились, теперь они удовлетворяют эстетические потребности людей и тем самым как бы перешли в сферу духовной культуры. Изучение подобных явлений в полевых условиях дает возможность собрать сведения о них, сохранившиеся в народной памяти, фольклорных сюжетах.

Как известно, головные уборы наиболее тонко и ярко отражали возрастные и социальные изменения в жизни девушек. Обряд смены девичьего головного убора на женский во время свадьбы существовал во всех группах татарского народа. Например *тастар сару*, *баш сару* в мишарском диалекте и *баш бәйләү* у крещенных татар[6].

Термины свадебных обрядов, сохраняют в своем составе целый ряд архаических языковых особенностей, относящихся к различным

историческим временным отрезкам. Это не удивительно, поскольку термины свадебных обрядов не являются продуктом современной эпохи, и большая часть их возникла в древности и существует по настоящее время. В составе терминов свадебных обрядов выявляются как слова, так и словообразовательные средства, восходящие к глубокой древности. Наряду с древними морфологическими формами в терминах свадебных обрядов сохранились также элементы древнетюркского синтаксиса.

Основным в изучении и реконструкции древнейшей татарской духовной культуры является продолжение экспедиционных сборов информации о самых различных этнографических, фольклорных, мифологических и других элементах, где они в той или другой степени (в реликтовой форме) дошли до нашего времени. Такое же значение имеет проводимая параллельно научная систематизация всех материалов, которыми располагает современная наука.

Благодарности: Работа выполнена при финансовой поддержке РГНФ (проект № 14-04-12024)

Литература

1. Толстые Н.И. и С.М. О задачах этнолингвистического изучения Полесья / Н.И. и С.М.Толстые // Полесский этнолингвистический сборник. – М.: Наука, 1983., с. 3-20.
2. Толстой Н.И. Язык и народная культура (очерки по славянской мифологии и этнолингвстике) / Н.И.Толстой. – М., 1995. – 509 с
3. Толстая С.М. Терминология обрядов и верований как источник реконструкции древней духовной культуры / С.М.Толстая // Славянский и балканский фольклор. Реконструкция древней славянской духовной культуры: источники и методы. – М.: Наука, 1989., с.215-229
4. Баязитова Ф.С. Аш-су һәм халык традицияләре лексикасы / Ф.С.Баязитова. – Казан: Дом печати, 2007. – 480 б.,
5. Баязитова Ф.С. Халык традицияләре лексикасы: бирнә һәм кием-салымнар (этнолингвистик, культуралогик, диалектологик концептлар) – Казан, 2009, 412 б.
6. Баязитова Ф.С. Туй йолалары лексикасы (җирле сөйләш һәм фольклор текстлары яссылыгында) – Казан: Паравитта, 2011, - 976б.
7. Владыкина Т.Г., Глухова Г.А.Обряды и праздники удмуртского календаря, - Ижевск, 2011, 320с.

РАЗДЕЛ V | E-LEARNING

ЭФФЕКТИВНОСТЬ ИСПОЛЬЗОВАНИЯ ИНТЕРНЕТ ПЛАТФОРМЫ ДУОЛИНГО ПРИ ОБУЧЕНИИ ИНОСТРАННЫМ ЯЗЫКАМ

А.И. Абдуллин
Казанский федеральный университет, Казань
Aidar-abd@mail.ru

В статье описывается опыт использования интернет платформы дуолинго в начальной и средней школе. Раскрывается механизм внедрения в учебный процесс, особенности использования. Проведенный эксперимент наглядно доказывает эффективность использования интернет платформы при обучении английскому языку.

Ключевые слова: *дуолинго, английский, интренет, платформа, эффективность*

Сколько учителей иностранного языка используют IT технологии на своих уроках? И как выбрать интернет платформу обеспечивающую мотивацию учащегося, системность и контроль учебного процесса? На уроках английского языка с помощью Интернета можно решать целый ряд дидактических задач: формировать навыки и умения чтения, используя материалы глобальной сети; совершенствовать умения письменной речи школьников; пополнять словарный запас учащихся; формировать у школьников мотивацию к изучению английского языка. Исходя из этих критериев, нами была выбрана наиболее популярная платформа (более 110 млн. учащихся) Дуолинго для определения эффективности использования в начальной и средней школе.

Предметом нашей работы является технология интернет платформы Дуолинго в процессе изучения иностранных языков. Объектом работы является процесс обучения английскому языку. Целью работы является определения эффективности использования интернет платформы в начальной и средней школе при обучении английскому языку.

Задачами работы является:

- исследование способов и механизмов внедрения технологии Дуолиго в процесс изучения английского языка;
- экспериментальное подтверждение положительной динамики в обучении иностранных языков при использовании платформы Дуолинго.

Внедрение и экспериментальное подтверждение проходило на базе МАОУ «Гимназии №19» в 7 классах с численностью от 14 до 18 человек в классе, всего в эксперименте приняло участие более 100 человек. Организационный момент внедрения длился 1,5 месяца. Что включало в себя регистрацию каждого ученика, получения логина и пороля, добавление в систему «дуолинго для школ». Сам эксперимент длился с середины октября 2015 года до февраля 2016 года включительно.

Эффективность подхода Дуолинго, основанного на анализе статистики, была проверена сторонним исследованием по заказу компании. Исследование, проведенное профессорами Городского университета Нью-Йорка и Университета Южной Каролины, показало, что 34 часа на Дуолинго дают столько же навыков чтения и письма, сколько даёт начальный семестровый курс в американском высшем учебном заведении, занимающий около 130.

Образовательная модель Дуолинго предлагает многочисленные письменные уроки и диктанты, однако разговорным навыкам уделяется меньше внимания. В Дуолинго есть игровое дерево навыков, по которому продвигаются пользователи, и словарный раздел, где можно практиковать уже изученные слова. Пользователи получают «очки опыта» (монеты) по мере изучения языка, например, после прохождения урока. Навыки считаются изученными, когда пользователи выполняют все связанные с ними уроки. За один урок можно заработать 10 очков. В Дуолинго также есть функция тренировки на время, когда пользователям дается 30 секунд и двадцать вопросов. За каждый правильный ответ дается одно очко опыта и семь или десять дополнительных секунд (время зависит от длины вопроса). За один курс пользователь может изучить до 2000 слов.

За прохождение всех уроков в навыке выдаётся 2 лингота, внутренняя игровая валюта. Существуют и другие способы приобретения линготов. Линготы можно тратить в игровом магазине или дарить пользователям, оставившим полезный комментарий на форуме. На любом уровне изучения языка можно купить тест на знание языка за 25 линготов, по прохождению теста будет выдан электронный сертификат. Тест можно проходить неоднократно.

Дуолинго использует для обучения подход, основанный на анализе большого количества статистических данных. На каждом этапе система запоминает, какие вопросы вызвали у пользователей трудности и какие

ошибки были совершены. Затем она агрегирует эти данные и использует для машинного обучения. Таким образом формируются индивидуальные уроки.

В эксперименте так же присутствовал соревновательный аспект: ученики видели друг друга в системе, наблюдая за прогрессом своих одноклассников и в школе на стене прогресса у каждого класса был свой плакат, где отмечались успехи каждого ученика. После прохождения уроков и по достижению уровня 500 очков опыта происходило торжественное вручение 2 сертификатов (один выдавался на руки ученику, второй помещался на стене прогресса рядом с таблицей достижений класса ученика). В течении 4.5 месяца эксперимента из 103 учеников уровня 500 очков достигли 66 человек, 1000 очков 42 человека, более 2000 очков достигли 27 человек. В результате использования платформы повышается общая языковая эрудиция, увеличивается словарный запас. По результатам итогового теста в классах, где внедряли интернет платформу Дуолинго качество на 37 % больше чем в классах не практикующие данную платформу.

В заключении можно сделать вывод о несомненном положительном эффекте внедрения платформы Дуолинго в учебный процесс. Так как предыдущие исследования проводились только среди студентов ВУЗа и взрослых людей желающих изучить язык, нами была предпринята попытка изучения эффекта в начальной и старшей школе. Опыт использования позволяет прийти к выводу о том, что для успешного внедрения интернет платформы необходима качественная работа с родителями и администрацией школы по популяризации идей учителя предметника. Данный эксперимент доказывает эффективность внедрения платформы, показывает возможность масштабирования практики использования и включиения платформы Дуолинго в учебный план школ Российской федерации.

Благодарности. Работа выполнена при поддержке научного руководителя д-р. пед. наук, доцента Абдрафиковой А.Р.

Литература

1. Бершадский, М. Информационная компетентность.//Народное образование. - 2009 - №4. - с.139.

2. Белкова М. М. Информационные компьютерные технологии на уроках английского языка // Английский язык в школе. 2008.

3. Интернет-ресурс: Duolingo Effectiveness Study. – URL: https://s3.amazonaws.com/duolingo-papers/other/vesselinov-grego.duolingo12.pdf (дата обращения: 25.03.2016).

4. http://www.sdkrashen.com/content/articles/krashen-does-duolingo-trump.pdf (дата обращения: 24.03.2016).

Интернет-ресурс: Duolingo Review: The Quick, Easy and Free Way to Learn A Language. – URL: http://www.fluentin3months.com/duolingo/ (дата обращения: 24.03.2016).

УДК 37.022

ПРИМЕНЕНИЕ ИКТ В ПРОЦЕССЕ ОБУЧЕНИЯ ДИСЦИПЛИН НА ИНОСТРАННОМ ЯЗЫКЕ В ВУЗЕ

М.А. Романова, Р.Р. Зарипова, Л.Л. Салехова
Казанский федеральный университет, Казань
romanova.maria.rus@yandex.ru, rinata-z@yandex.ru,
salekhova2009@gmail.com

В статье рассматриваются вопросы, связанные с применением ИКТ в процессе обучения предметному содержанию на иностранном языке в вузе. Дается характеристика предметно-языкового интегрированного обучения и лингво-информационной компетенции, формируемой в результате внедрения ИКТ в процесс обучения.

Ключевые слова: *предметно-языковое интегрированное обучение, лингво-информационная компетенция, информационно-коммуникационные технологии, иностранный язык.*

В условиях интернационализации российского высшего образования повышается роль иностранного языка не только как средства коммуникации, но и как средства познавательной деятельности. В связи с этим актуальным и значимым представляется разработка и внедрение образовательных программ на иностранных языках. Однако анализ практического опыта европейских университетов показал, что недостаточно перевести имеющийся учебный курс на иностранный язык, необходимо использовать новые технологии к обучению предметному знанию на иностранном языке. Одной из таких технологий является предметно-языковое интегрированное обучение (Content and Language Integrated Learning - CLIL), который предоставляет студентам возможность изучать дисциплину и иностранный язык одновременно, что способствует интенсификации профессиональной подготовки в вузе.

Предметно-языковое интегрированное обучение, будучи одной из форм билингвального образования, стало весьма популярным в последние несколько лет. Исследования, проведенные в Европе, показывают, что такой вид обучения внедряется в подавляющем большинстве европейских

стран в различных формах. Оно представляет собой двунаправленный образовательный подход, при котором иностранный язык используется одновременно как цель и средство обучения [4]. Аббревиатура CLIL стала популярной и широко используемой уже в 1990-х. Это не абсолютно инновационный подход в обучении иностранному языку и предмету. Такой тип обучения был известен прежде как «иммерсионное обучение» и очень успешно использовался во многих странах, например, в Канаде. Что отличает CLIL от других форм билингвального обучения, так это то, что он не фокусируется только на изучении языка, но уделяет равное количество внимания одновременно двум областям – языку и неязыковой предметной области. При этом преподавание неязыкового предмета происходит не на иностранном языке, а через него, и прогресс в обеих областях имеет одинаковую важность. Таким образом студентам не преподают иностранные языки как на обычных уроках, они погружены в языковую среду, подобную той, в какую попадает ученик, оказавшийся в иностранной школе на типичном уроке. Это означает что после начальной стадии иностранный язык становится языком обучения неязыкового предмета, что позволяет студентам познакомиться с культурной средой, частью которой этот иностранный язык является [5].

В соответствии с моделью 4C [4], успешное занятие при предметно-языковом интегрированном обучении должно объединять в себе следующие элементы:

- содержание (Content) – овладение знаниями и умениями, понимание специфики предметной области;
- общение (Communication) – использование иностранного языка как средства общения и обучения;
- познание (Cognition) – развитие мыслительных навыков, формирование концептуальных понятий;
- культура (Culture) – (формирование альтернативных перспектив и национального мышления, что позволит углубить понимание культурного феномена «свой-чужой»).

Несмотря на то что эти элементы возможно осуществить индивидуально, при предметно-языковом интегрированном обучении они функционируют в единой системе. Это взаимодействие объединяет теорию обучения, теорию языкового обучения и межкультурное понимание.

Эффективное изучение языка на основе предметной области и изучение предмета с помощью иностранного языка, предполагаемое при предметно-языковом интегрированном обучении, обусловлено:

а) прогрессом в знаниях, навыках и пониманием предметной области;

б) включением в ассоциативный мыслительный процесс;

в) коммуникативным взаимодействием;

г) развитием соответствующих языковых знаний и навыков;

д) приобретением углубленных межкультурных знаний посредством противопоставления «свой» - «чужой».

Информационно-коммуникационные технологии (ИКТ), определяемые как разнообразный набор инструментов и ресурсов, используемых для создания, распространения, хранения и управления информацией, представляются исключительным помощником в образовательном процессе на иностранном языке и обуславливают более содержательное обучение, повышение интереса и мотивации, критический анализ информации.

ИКТ открывают для студентов и преподавателей более быстрый доступ к информации, которая может быть получена не благодаря текстам, но и через аудиовизуальные средства. С использованием ИКТ учебный процесс перестает быть простым восприятием и сохранением информации, полученной в аудитории. Так, студенты не только получают информацию, а оказываются вовлеченными в ее поиск. ИКТ предоставляет множество ресурсов для самостоятельного оценивания знания и немедленной обратной связи [5].

Существует два аспекта использования ИКТ в учебном процессе. С одной стороны, они являются поисковым инструментом, а с другой – средством взаимодействия и общения. ИКТ можно использовать как для индивидуальных заданий, так и для совместной групповой работы студентов. При планировании занятия с использованием ИКТ необходимо не только четко определить тематическую наполненность урока, но технические средства, которые не должны идти отрыве от учебного процесса и изучаемого предмета, а быть интегрированы в него.

Основной вопрос, возникающий на основе данной интеграции, - как на ИКТ основе можно достичь целей, которые ставит CLIL обучение. Взаимодействие ИКТ и CLIL предполагает развитие у учащихся как минимум двух компетенций: языковой и информационной.

Языковая компетенция предполагает использование языка как инструмента устной и письменной коммуникации, так и средства регулирования поведения и эмоций. Коммуникация на иностранных языках требует навыков межкультурного взаимодействия и понимания. Развитие этой компетенции - ключ к решению различных конфликтов.

Работа с информацией предусматривает ее поиск, получение, обработку, передачу и трансформацию в знания. Различные аспекты этих процессов, от поиска информации, ее передачи на различных носителях до использования вместе с коммуникационными технологиями при обучении, составляют информационную компетенцию, которая вкупе с языковой образует лингво-информационную компетенцию студента.

В исследованиях отечественных ученых-педагогов (В.П. Беспалько, Б.С. Гершунский, А.П. Ершов, И.Г. Захарова, Е.С. Полат, А.Ю. Уваров и др.) основательно разработана теория и методика использования информационных и коммуникационных технологий в системе образования. Формирование лингво-информационной компетенции как лингво-образовательной инновации отражено в работах Рыбалко Т.Г.[2], Салеховой Л.Л. [3], Зариповой Р.Р. [1], Ступиной Т.Л. и др. Однако подавляющее количество работ по интеграции ИКТ и CLIL принадлежит зарубежным ученым, где степени разработанности этой проблемы намного выше в силу развития компьютерных технологий.

Изабель Перес [6] выделяет различные причины использования ИКТ в предметно-интегрированном языковом обучении:

1. ИКТ подразумевает новые пути обучения и преподавания;
2. ИКТ и CLIL имеют принципиальные методологические сходства, фокусируясь одновременно на учебном процессе и заданиях;
3. ИКТ способствует укреплению и взаимодействию компонентов модели 4С;
4. Существует большое количество разнообразных ИКТ, чтобы работать с материалом на иностранных языках;
5. ИКТ позволяет применять активные совместные стратегии.

ИКТ имеет широкий спектр ресурсов, которые могут быть интегрированы в предметно-языковое интегрированное обучение. Большую часть из них предоставляет интернет-сфера, в которой становится возможным не только поиск информации, но и ее обработка, а также создание собственного продукта.

В настоящее время в сети появляется все больше ресурсов, готовых предложить услуги подкастинга. Подкастинг — процесс создания и распространения звуковых или видеофайлов (подкастов) в стиле радио-и телепередач в Интернете. Подкасты имеют определенную тематику и периодичность издания, являясь выгодная альтернативой классическому радиовещанию, так как подкаст доступен к прослушиванию в любое время и любое количество раз. Подкасты на иностранных языках дают возможность студентам не только прослушивать записи многократно, но и комментировать их, обсуждать онлайн, дополнять содержание текстами, видео и фотоматериалами, создавать собственные работы [5]. В качестве примера можно привести ресурс Podomatic (www.podomatic.com), на котором можно создавать собственные работы, загружать аудио, фото и текстовые материалы, задания к ним и делиться ими.

Одной из наиболее популярных технологий является Веб-квест. Веб-квест - это сайт или задание в сети Интернет, с которым работают студенты, выполняя ту или иную учебную задачу. Веб-квесты

оформляются таким образом, чтобы нацелить студента не только на поиск информации, но и на ее анализ. Все материалы, с которыми работают учащиеся, размещены в сети Интернет. Они охватывают отдельную проблему, учебный предмет, тему, могут быть межпредметными. В результате выполнения веб-квеста студенты создают продукт, который может быть в виде веб-страниц, блогов, видео, презентации и т.д.

Веб-квест способствует достижению нескольких задач:

- повышение мотивации к самообучению, поощрение учеников учиться независимо от учителя;
- формирование новых компетенций на основе использования ИКТ для решения учебных задач, умений находить несколько способов решений проблемной ситуации, определять наиболее рациональный вариант, обосновывать свой выбор;
- реализация творческого потенциала;
- развитие коммуникативных умений и умений работы в группе; (планирование, распределение функций, взаимопомощь, взаимоконтроль);
- развитие мышления;
- повышение словарного запаса.

Одним из ресурсов сети Интернет, дающих возможность создавать собственные веб-квесты, является Zunal (http://zunal.com/). Он удобен тем, что рубрики в нем уже готовы и автор должен их заполнить, а не создавать заново. Веб-квест является наиболее распространенным примером задания на основе Интернет ресурсов. К ним также относятся:

1. Тематический список ссылок (Hotlist) - «список по теме» представляет собой список Интернет сайтов по изучаемой теме.

2. Коллекция мультимедийных файлов (MultimediaScrapbook) - представляет собой своеобразную коллекцию мультимедийных ресурсов, которые могут быть скачаны студентами и использованы в качестве информационного и иллюстративного материала при изучении темы.

3. Поиск сокровищ (Treasure/ScavengerHunt) - содержит ссылки на различные сайты по изучаемой теме, направляющие поисковую деятельность студентов.

4. Коллекция примеров (SubjectSampler) - здесь содержатся ссылки на текстовые и мультимедийные материалы сети Интернет. Главной особенностью этой технологии является то, что получение информации должно строиться на эмоциональном уровне. Необходимо не просто ознакомиться с материалом, но и выразить и аргументировать свое собственное мнение по изучаемому вопросу.

Веб-квест включает в себя все компоненты четырех указанных выше материалов и предполагает проведение проекта с участием всех студентов.

Итак, технологии дают возможность студентам и преподавателям работать с богатым мультимедийным материалом, а компоненты модели CLIL будут улучшаться, если студенты будут способны создавать файлы, объекты, программы, презентации, проекты на иностранных языках, что приведет к естественному сочетанию разговорного и письменного языка, звуков и изображений. Внедрение ИКТ в предметно-языковое интегрированное обучение будет способствовать не только ознакомлению или обзору изучаемой информации через аудио или видеоматериалы, но расширению и углублению содержания через веб-исследования [5].

Таким образом, фундаментальные компьютерные навыки, развитые на основе средств ИКТ, могут помочь студентам овладеть новым материалом. Знание их является немаловажным и подготавливает учащихся к различным областям образования в целом. CLIL создает лучшую среду для объединения обучения на иностранном языке и ИКТ при взаимосвязанном изучении контента и языка, целью которого является формирование лингво-информационной компетенции как интегрированного целого, включающего знания иностранного языка и знания информационных технологий, умения и навыки, способствующие формированию готовности к их практическому применению в профессиональной деятельности и являющиеся средством профессионального развития и самосовершенствования [2].

Литература

1. Зарипова Р.Р. Компьютерные технологии в инновационном обучении иностранным языкам / Система Moodle. – 2014.

2. Рыбалко Т. Г. Формирование лингво-информационной компетентности как лингвообразовательная инновация. - Вестник Нижегородского университета им. Н.И. Лобачевского - № 6 / 2008.

3. Салехова Л.Л. , Хакимуллина Н.И. Формирование лингво-информационной компетенции школьника в процессе билингвального обучения информационно-коммуникационным технологиям // Современные проблемы науки и образования. – 2013. – № 1.;

4. Coyle, D., Hood, P., and Mash, D. (2010). CLIL: Content and Language Integrated Learning, Cambridge: Cambridge University Press.

5. Lidia Wojtowicz, Mark Stansfield, Thomas Connolly, Thomas Hainey. The Impact of ICT and Games Based Learning on Content and Language Integrated Learning. – International conference “ITC for Learning Language”, 4th Edition , Florence, Italy, 20 - 21 October, 2011.

6. Pérez Torres, I. Apuntes sobre los principios y características de la metodología AICLE en V. Pavón, J. Ávila (eds.), Aplicaciones didácticas para la enseñanza integrada de lengua y contenidos. Sevilla: Consejería de Educación de la Junta de Andalucía-Universidad de Córdoba, 2009.

УДК 378.147

ФОРМИРОВАНИЕ ИНФОРМАЦИОННОЙ КОМПЕТЕНЦИИ СТУДЕНТОВ ГУМАНИТАРНЫХ СПЕЦИАЛЬНОСТЕЙ (на примере курса «Информационные технологии»)

М.А. Лукоянова
Казанский федеральный университет, Казань
marina-lkn@yandex.ru
Р.Р. Ибрагимова
Казанский федеральный университет, Казань
ibragimova1492@gmail.com

В статье описывается решение проблемы формирования информационной компетенции студентов гуманитарных специальностей, способных использовать информационные технологии в профессиональной деятельности.

Ключевые слова: *информационная компетенция, информационные технологии, студенты гуманитарных специальностей.*

Глобальная информатизация общества предполагает подготовку специалистов, способных решать профессиональные задачи на основе использования различных источников информации и современных информационных технологий. Практически каждому современному профессионалу необходимо ориентироваться в тенденциях информационного развития, жить и работать в постоянно меняющемся информационном обществе, овладевать новыми способами информационного обмена.

Поэтому требования к уровню подготовки специалистов гуманитарных специальностей, способных решать профессиональные задачи с применением информационных технологий, возрастают. В Федеральном государственном образовательном стандарте высшего образования (ФГОС ВО) нового поколения, отличительной чертой которого является направленность на компетентностный подход, уровень подготовки студентов-гуманитариев определяется сформированностью их информационной компетенции.

Понятие «информационная компетенция» в настоящее время не имеет однозначного толкования. Так, доктор педагогических наук С.Д. Каракозов считает, что «информационная компетенция заключается в способности человека обеспечить себе открытый доступ к информации с возможностью публикации собственной информации и свободного выбора источников информации» [4, с. 122]. А.В. Хуторской трактует данное

понятие как «умение самостоятельного поиска, отбора и анализа необходимой информации при помощи реальных объектов и информационных технологий» [8]. О.Б. Зайцева считает, что под информационной компетенцией следует понимать «сложное индивидуально-психологическое образование на основе интеграции практических умений и теоретических знаний в область инновационных технологий и определённого набора личностных качеств» [3, с. 14].

Информационная компетенция студентов гуманитарных специальностей во ФГОС ВО 3+ определяется, как способность решать стандартные задачи профессиональной направленности на основе информационной культуры с использованием информационно-коммуникационных технологий (ИКТ) и с учетом основных требований информационной безопасности [5].

В работах отечественных и зарубежных ученых С.Г. Воровщикова, А.Н. Дахина, Дж. Равена и др., выдвигающих методологию компетентностного подхода, информационная компетенция выделяется как «ключевая» [6, с. 3]. «Именно новые потребности общества и личности определили информационную компетенцию как одну из базовых, ключевых» [1, с. 113].

Формирование информационной компетенции студентов гуманитарных специальностей рассмотрим на примере курса «Информационные технологии». Данный процесс является достаточно специфичным, так как:

1) Студенты-гуманитарии обладают лингвистическим типом интеллекта, имеют грамотную речь, эрудированы, но, как правило, категорично относятся ко всем естественнонаучным и математическим дисциплинам.

2) Школьная подготовка студентов-гуманитариев в области информатики и ИКТ является недостаточной для быстрого усвоения предоставляемого материала.

3) Обычно работа студентов над заданиями сводится преимущественно к автоматическому выполнению действий, описанных в методичке, поэтому запоминания пройденного материала не происходит. «Как правило, в них нужно строить графики надуманных и зачастую излишне усложненных функций, которые редко встречаются не только в гуманитарных, но даже в точных науках» [2, с. 172].

Для формирования информационной компетенции студентам-гуманитариям нужно иметь представление о теоретических и практических аспектах информационной грамотности. Также стоит учесть, что информационные компетенции студентов-гуманитариев должны быть специализированы и помимо обобщенных знаний должны включать знания и умения актуальные для конкретной специальности.

Учитывая особенности процесса формирования информационной компетенции студентов гуманитарных специальностей, отметим, что технические аспекты работы с компьютером лучше запоминаются, если они имеют для студентов надпредметный смысл, доступную ассоциацию и яркий образ. Поэтому формирование информационной компетенции студентов-гуманитариев должно быть основано на следующих принципах: аттрактивного целеполагания, целостности и комплексности заданий, выполняемых с помощью информационных технологий, аутентичности и уникальности решения задания.

В связи с выявленными трудностями при формировании информационной компетенции студентов-гуманитариев необходимым, на взгляд автора, становится выделение единой основной цели, объединяющей изучение прикладных аспектов курса «Информационные технологии», которые могут быть рассмотрены при решении профессионально-значимых для гуманитариев задач. Основная цель позволит обеспечить целостность изучаемого курса, состоящего из разных тем, обусловленных его прикладными аспектами, и получить требуемый образовательный результат у студентов-гуманитариев. Определение единой цели основано на теории педагогического целеполагания Я.С. Турбовского, в которой учебный процесс рассматривается не только как абстракция, раскрывающая сущностные характеристики определённой совокупности явлений и связей между ними, но и как непрерывная практическая деятельность, призванная реализовать формирующие возможности представленных в программах знаний в процессе осуществляемой преподавателем целеполагающей деятельности [7].

Для обеспечения целостности курса, повышения мотивации студентов и эффективности учебного процесса, направленного на формирование информационной компетенции студентов-гуманитариев, были объединены прикладные аспекты курса «Информационные технологии» (операционная система компьютера, текстовый редактор, электронная таблица, программа для создания презентаций) и технология связывания и внедрения объектов в документы различных типов основной целью – решением профессионально-значимых для студентов гуманитарных специальностей задач с использованием инструментов автоматизации обработки различных видов информации. В дальнейшем закрепление полученных знаний и приобретение определенного опыта информационной деятельности, направленной на формирование информационной компетенции студентов-гуманитариев, может осуществляться при оформлении, обработке результатов исследования и разработке презентации при создании и представлении курсового или дипломного проекта.

Студенты, работая в среде текстового редактора, изучают основные приемы и специальные средства создания, редактирования, форматирования и автоматизации обработки документа, что помогает им в обработке многостраничных документов сложной структуры, содержащих внедренные электронные таблицы, диаграммы и графические объекты.

В электронных таблицах студенты изучают приемы редактирования и форматирования текстовой и числовой информации, формул, автозаполнения ячеек, работают с различными типами ссылок, средствами редактирования таблиц, создания диаграмм и их представлением на листах диаграмм. Работа с электронной таблицей позволяет студентам освоить основные операции по обработке результатов исследовательской деятельности и их наглядному представлению в виде таблиц и диаграмм. Внедряя динамические таблицы и диаграммы в текстовый документ и презентацию, для обеспечения их модификации в случае изменения данных в исходном файле электронной таблицы, студенты изучают и применяют на практике технологию связывания и внедрения объектов.

Изучение студентами программы для создания и демонстрации презентаций направлено на освоение приемов и средств создания, редактирования, оформления и настройки презентации для публичных выступлений. Студенты осваивают различные способы создания и оформления презентации, вставки и форматирования различных объектов, включая внедренные из электронной таблицы, вставки и настройки навигации в презентации, настройки управления демонстрацией для показа презентации.

Процесс осмысления студентами гуманитарных специальностей основной цели изучения курса «Информационные технологии» связан с решением профессионально-значимых задач в рамках его основных тем, отражающих изучение прикладных аспектов. Отметим, что в результате подобной информационной деятельности у студентов-гуманитариев складывается положительное отношение к изучаемой дисциплине, что является значимым фактором при формировании у них информационной компетенции.

Для оценки уровня сформированности информационной компетенции студентов-гуманитариев были проанализированы результаты текущего и итогового контроля успеваемости контрольных и экспериментальных групп за 2012-2015 гг. Студенты контрольных групп осваивали лишь прикладные аспекты курса. В экспериментальных группах осваивались прикладные аспекты, связанные основной целью – решением профессионально-значимых задач с использованием инструментов автоматизации обработки различных видов информации. Мониторинг резуль-

татов был осуществлен по средним значениям от набранного каждым студентом количества баллов за весь курс на основе бально-рейтинговой системы. В контрольных группах студенты показали средний балл 76, в экспериментальных группах – 84. Полученные результаты позволяют говорить о положительной динамике обучения информационным технологиям студентов гуманитарных специальностей, когда темы курса связаны основной целью.

Таким образом, выделение основной цели курса «Информационные технологии» позволяет обеспечить его целостность, повысить уровень эффективности учебного процесса, получить практический результат, связанный с решением профессионально-значимых задач, что в конечном итоге, способствует формированию информационной компетенции студентов гуманитарных специальностей, соответствующей уровню требований современного информационного общества.

Литература

1. Войнова Н.А., Войнов А.В. Особенности формирования информационной компетентности студентов вуза // Инновации в образовании. 2004. № 4. С. 111-118.

2. Жук О.Л., Сиренко С.Н., Колесников А.В. Формирование общепрофессиональных компетенций студентов-гуманитариев в процессе изучения информационных технологий на основе междисциплинарной интеграции // Информатизация образования – 2014: педагогические аспекты создания и функционирования виртуальной образовательной среды: материалы междунар. науч. конф., Минск, 22-25 окт. 2014г. / редкол.: В.В. Казаченок (отв. ред.) [и др.]. – Минск: БГУ, 2014. С. 170-176.

3. Зайцева О.Б. Формирование информационной компетентности будущих учителей средствами инновационных технологий: автореф. дис. ... канд. пед. наук. Брянск, 2002. 19 с.

4. Каракозов С.Д. Развитие предметной подготовки учителей информатики в контексте информатизации образования: дис. … докт. пед. наук. Барнаул, 2005. 427 с.

5. Об утверждении федерального государственного образовательного стандарта высшего образования по направлению подготовки 45.03.01 Филология (уровень бакалавриата) от 7.08.2014 г. – URL: http://www.edu.ru/db/mo/Data/d_14/m947.pdf.

6. Табачук Н.П. Развитие информационной компетенции студентов в образовательном процессе гуманитарного вуза: автореф. дис. ... канд. пед. наук. Хабаровск, 2009. 17 с.

7. Турбовской Я.С. Взаимодействие педагогической науки и практики: диагностический аспект. М., 1993. 194 с.

8. Хуторской А.В. Ключевые компетенции и образовательные стандарты // Интернет-журнал «Эйдос». 2002. URL: http://www.eidos.ru/journal/2002/0423.htm.

УДК.004

ИНТЕЛЛЕКТУАЛИЗАЦИЯ ПРОЦЕССОВ ТЕСТИРОВАНИЯ В ЭКСПЕРТНОЙ СИСТЕМЕ ОБУЧЕНИЯ ИНОСТРАННОМУ ЯЗЫКУ

Мамедова М.Г.
Институт Информационных Технологий НАНА,
г. Баку, Азербайджан
depart15@iit.ab.az
Кулиева З.Ю.
Институт Информационных Технологий НАНА,
г. Баку, Азербайджан
guliyeva_z_y@hotmail.com

В статье в рамках предложенного концептуального подхода к проектированию экспертной системы обучения иностранному языку рассмотрены вопросы разработки диагностического тестирующего блока (ДТБ). На примере грамматического модуля ДТБ приведены архитектура, принципы функционирования, структурные компоненты модели представления знаний, экспертные требования к формированию тестовых заданий, включенных в учебный контент последнего.

Ключевые слова: *экспертная обучающая система, диагностический тест-блок, модуль грамматики, база знаний, лингвистическая переменная.*

Введение

Информатизация общества, глобализационные процессы, высокие темпы развития технологий, стремительное устаревание знаний обусловливают жесткую конкуренцию на рынке труда и предопределяют постоянно растущие требования к непрерывному обновлению знаний, умений и навыков. Основу непрерывного образования составляют компьютерные системы обучения, предоставляющие равные возможности доступа к образовательным ресурсам различным категориям пользователей и удовлетворяющие их потребности в повышении уровня образования, в самореализации и саморазвитии [1]. Наиболее востребованными в последние годы являются интеллектуальные обучающие системы [2-6], и в том числе получившие широкое распространение экспертные обучающие системы (ЭОС), способные имитировать работу человека-эксперта в определенной предметной области. Эта особенность ЭОС позволяет реализовать личностно-ориентированный подход

к обучению без непосредственного участия преподавателя и открывает широкие возможности для решения задачи самообучения, предполагающей учет целей обучающегося, а также оптимизацию траектории его обучения.

Среди областей применения компьютерных обучающих технологий значительное место занимает изучение иностранных языков (ИЯ) [7,8]. Интерес к самостоятельному компьютерному изучению ИЯ вызван как развитием аппаратных и программных средств информационных технологий, предоставляющих возможности включения в единое информационно-коммуникационное пространство все большего количества естественных языков, так и потребностями в обеспечении успешной профессиональной коммуникации. Не случайно широкие исследования и разработки по компьютерному изучению ИЯ за рубежом стимулировали появление для обозначения этого направления специального термина “CALL” – “computer assistant language learning” (компьютерное изучение языков), в рамках которого разработано большое количество CALL-технологий, успешно реализуемых в создаваемых обучающих программах [9-11]. Разработчики последних также связывают будущее развитие компьютерных технологий обучения с разработкой интеллектуальных обучающих систем и, в частности, с ЭОС.

Концептуальный подход к проектированию ЭОС

Предлагаемый в работе концептуальный подход к проектированию экспертной системы обучения иностранному языку базируется на следующих принципах:

1) построение интеллектуальной обучающей среды для изучения иностранного языка различными категориями пользователей, в основе которой лежат знания экспертов-учителей, отражающие наилучшие методики преподавания набора учебных курсов - разделов ИЯ (фонетики, грамматики, лексики и т.п.), а также различные учебные ситуации, представленные правилами адаптации учебных курсов к усвоению предлагаемых материалов.

2) индивидуализация обучения, требующая адаптацию интеллектуальной обучающей среды к индивидуальным особенностям (уровню знаний и когнитивным способностям по усвоению ИЯ) каждого конкретного пользователя;

3) разработка модели пользователя (обучающегося), модели «электронного преподавателя», модели процесса обучения, модели учебного курса, модели тестирования, контроля и оценки знаний, а также механизмов манипулирования данными о начальном уровне знаний и текущем состоянии обучения отдельного пользователя;

4) формирование баз знаний и данных, поддерживающих модели «электронного преподавателя», пользователя, учебного курса, процесса обучения, тестирования, контроля и оценки знаний пользователя, а также позволяющих имитировать экспертные рассуждения, оценку и заключения учителя, делать выводы и принимать решения;

Основная цель, которую преследуют авторы при создании экспертной системы обучения иностранному (в данном случае, английскому) языку, заключается в предоставлении пользователю возможности самостоятельно изучить иностранный язык при содействии «электронного преподавателя», в качестве которого выступает ЭОС.

Диагностический тест-блок ЭОС

Традиционно одним из важных аспектов многогранного процесса обучения иностранному языку является тестирование и контроль преподавателем уровня знаний, практической и теоретической базы обучаемого, по результатам которого осуществляется выбор соответствующего уровня и траектории обучения. На сегодня компьютерные технологии, и в особенности, экспертные обучающие системы, играют ведущую роль в выявлении как начального уровня знаний пользователей (лиц, принявших решение относительно самостоятельного обучения иностранному языку с использованием ЭОС), так и для самоконтроля усвоения материала и выявления степени соответствия пользователя требуемому эталону [12-14].

Преподаватель при оценке уровня владения иностранным языком оперирует такими понятиями, как «среднее владение языком», «свободное владение языком», «слабое знание грамматики» и т.п., которые фактически представляют собой нечеткие понятия, однако оцениваются в рамках четкой балльной системы. Так, например, перечисленные вербальные параметры могут быть идентифицированы как значения лингвистической переменной «уровень владения иностранным языком». Использование теории нечетких множеств и модели нечеткого логического вывода [15,16] может позволить сократить неопределенность вербально выраженных параметров, критериев и показателей посредством их формализации в виде лингвистических переменных и соответствующих значений функций принадлежности, а также учесть степень сложности и значимости каждого тестового задания. Это, в свою очередь, даст возможность повысить степень объективности результатов тестирования.

В настоящей работе рассматривается структура одного из модулей диагностического тест- блока (ДТБ) экспертной системы, разработанного для оценки знаний грамматики английского языка. Контент модуля грамматики ДТБ состоит из набора тестовых заданий разной сложности, позволяющих оценить знания грамматики английского языка в рамках

5 уровней (Beginner, Elementary, Pre-Intermediate, Intermediate, Upper-Intermediate). При этом для каждого из грамматических уровней разработана собственная модель, включающая соответствующие базы знаний и данных, а также правила формирования порогового результата для перехода к следующему (более высокому) уровню (набору тестовых заданий/вопросов). Разработка тестовых заданий осуществлялась при непосредственном участии экспертов-преподавателей английского языка и, естественно, сформированные и включенные в базу знаний правила отражают логику рассуждений последних. При этом эксперты руководствовались следующими соображениями:

1. Классификация и отбор грамматических категорий английского языка по уровням *(Level gradation)*.
2. Составление вопросов с учетом всех категорий, соответствующих рассматриваемому уровню.

Оценка вопросов по следующим критериям: коэффициент сложности вопроса *(Expert evaluation of question gravity coefficient)*, количество грамматических категорий в вопросе (*Number of level categories*), вербальная оценка степени сложности вопроса (*Expert verbal evaluation*).

3. Создание базы знаний для каждого уровня.

Модуль грамматики ДТБ имеет иерархическую структуру, в соответствии с которой база знаний тестовых заданий в соответствии с уровнями иерархии разделена на подмодули, состоящие из определенного количества тестовых вопросов. При этом каждому подмодулю соответствует своя подбаза знаний, включающая правила тестирования, содержательно зависящие от назначения данного уровня. Тестовые задания в базе знаний описаны в виде продукционных правил типа: «Если условие1 и/или условие2 … и/или условие К, то действие1 и … действие М». Следует отметить, что подготовленные экспертом (учителем) тестовые задания неодинаковы по значимости и в зависимости от степени сложности имеют различные веса. Доступ к выполнению тестовых заданий следующего уровня разрешается только после завершения тестирования и получения результатов на текущем уровне. В зависимости от количества правильных ответов и их сложности устанавливается уровень знаний пользователя и ДТБ принимает решение, в соответствии с которым пользователь переводится в блок обучения, где на основе экспертных знаний система определяет индивидуальный учебный курс, т.е. грамматический уровень, с которого необходимо начать процесс обучения. Например, для определения уровня знаний пользователя по уровню Beginner экспертом-преподавателем составлены восемь возможных альтернативных варианта ответов и определены правила вывода, в которых степень тяжести вопроса вербально определена как *легкий, средней сложности и высокой сложности*.

Ниже приведены примеры правил из базы знаний грамматического модуля ДТБ ЭОС.

Правило 1. *ЕСЛИ* *X_{11}(пользователь ответил на не менее 5 вопросов по уровню Beginner*

ИЛИ *X_{12}(пользователь ответил на 7 легких/2средней сложности вопросов по уровню Beginner)*

ИЛИ *X_{13}(пользователь ответил на 4 легких/средней сложности/ 2 сложных вопросов по уровню Beginner)*

ИЛИ *X_{14}(пользователь ответил на 5 легких/средней сложности/1 сложный вопрос по уровню Beginner)*

ТО *Y_1 (он должен начать обучение с первой части уровня Beginner)*

Правило 2. *ЕСЛИ* *X_{21}(пользователь ответил на 6 легких/средней сложности/1 сложный вопрос по уровню Beginner)*

ИЛИ *X_{22} (пользователь ответил на 3 легких/средней сложности/3сложных вопросов по уровню Beginner)*

ИЛИ *X_{23}(пользователь ответил на 5 легких/средних/2 сложных вопросов по уровню Beginner)*

ТО *Y_2 (он должен начать обучение со второй части уровня Beginner)*

Правило3. *ЕСЛИ* *X22 (пользователь ответил не менее на 8 вопросов по уровню Beginner)*

ИЛИ *X_{23} (пользователь ответил на 4 легких/средней сложности / 3 сложных вопросов по уровню Beginner)*

ИЛИ *X_{24}(пользователь ответил на 5 легких/средней сложности /2 сложных вопросов по уровню Beginner)*

ТО *$У_3$ (он должен начать обучение со второй части уровня Beginner*)

Таблица 1

Информационная модель знаний, включенных
в учебный контент модуля грамматики английского языка в ДТБ

Уровень	*Число грамматических категорий, учитываемых в рамках одного уровня*	*Количество вопросов*	*Число категорий, знание которых проверяется в вопросах*		*Диапазон изменения весов (коэффициента трудности) вопроса*	*Вербальные (нечеткие) экспертные оценки лингвистической переменной «сложность вопроса»*	
			Число категорий, отраженных в вопросе	*Кол-во вопросов*		*Вес вопроса*	*количество вопросов*
Beginner	17 категорий	10	1 категория 2 категории 3 категории	2 4 4	0,3-0,6	Легкий Средней сложности Высокой сложности	2 5 3
Elementary	22 категории (14 категорий, повторяющихся на предыдущем уровне)	10	1 категория 2 категории 3 категории	1 4 5	0,3-0,7	Легкий Средней сложности Высокой сложности	1 4 5
Pre-intermediate	28 категории (11 категорий, повторяющихся на предыдущем уровне)	10	1 категория 2 категории 3 категории	1 3 6	0,5-0,8	Легкий Средней сложности Высокой сложности	1 6 3
Intermediate	32 категории (13 категорий, повторяющихся на предыдущем уровне)	10	1 категория 2 категории 3 категории	0 2 8	0,7-0,9	Легкий Средней сложности Высокой сложности	0 3 7
Upper-intermediate	37 категорий (8 категорий, повторяющихся на предыдущем уровне)	10	1 категория 2 категории 3 категории	0 4 6	0,8-0,9	Легкий Средней сложности Высокой сложности	0 4 6

После ответа на все вопросы по уровню Beginner на основе полученных данных и вышеуказанных правил определяется уровень знаний пользователя и принимается одно из двух нижеуказанных действий:

1. Если фактические данные соответствуют правилу1 или правилу 2, то выполнение программы приостанавливается и выдается сообщение, соответствующее данному правилу;

2. Если фактические данные соответствуют правилу 3, выполнение программы продолжается и пользователю предоставляется возможность для проверки знаний на более высоком уровне.

Оптимальное число выбранных категорий в рамках одного уровня может варьировать в зависимости от целей обучения. Одна и та же грамматическая характеристика может проверяться как в отдельности, так и в комплексе с другими грамматическим категориями, что находит свое отражение в увеличении значения веса (коэффициента сложности) предлагаемого вопроса. Нечеткие значения экспертной оценки весов лингвистической переменной «сложность вопроса» изменяются в интервале [0.3-0.9]. В рамках одного вопроса может диагностироваться от одного до трех грамматических категорий, что находит свое отражение в вербальной (нечеткой) оценке лингвистической переменной «сложность вопроса».

Заключение

Разработанная на основе нечеткого моделирования база знаний модуля грамматики диагностического тест-блока, позволяющая осуществить автоматизацию процесса тестирования как начальных знаний пользователей, так и усвоения ими учебного материала на основе анализа допущенных ошибок, дает возможность адаптировать учебный материал к индивидуальным характеристикам пользователя.

Для практической реализации процесса тестирования на базе грамматического модуля ДТБ в ЭС обучения иностранному языку, в частности, грамматике английского языка, были разработаны алгоритмы обработки результатов тестирования и соответствующее программное обеспечение на языке Delphi. ДТБ, апробированный среди сотрудников Института Информационных Технологий НАНА, позволил определить их уровень знания языка и планировать дальнейшую индивидуальную траекторию обучения.

Литература

1. Сулейманов Д.Ш., Гильмуллин Р.А., Хасанова Л.Р. Интерактивный Интернет-учебник по татарскому языку «Татар теле онлайн» // Эл. журнал "Образовательные технологии и общество", № 1, 2011. Спец. раздел выпуска под ред. акад. АН РТ, дир. НИИ «Прикладная семиотика» АН РТ, проф. КФУ Д.Ш. Сулейманова /http://ifets.ieee.org/russian/depository/v14_i1/pdf/10r.pdf

2. Изучение принципов работы и поддержки обучающих систем. http://www.soljah.narod.ru/1semestr.htm

3. Трембач В.М. Основные этапы создания интеллектуальных обучающих систем// Программные продукты и системы, №3, 2012, стр.147-151.

4. Брусиловски П. Адаптивные и интеллектуальные технологии для сетевого обучения. http:Новости искусственного интеллекта, 2001, №4, с.3-13.

5. Клюдкин В.Э. Web-ориентированные интеллектуальные обучающие системы на основы нечеткого деятельностного подхода в обучении. Наука и Образование, Электронный Научно-Технический журнал .№ 1, ноябрь, 2012.http://technomag.edu.ru/doc/489620.html

6. Kudryavcev V.B. and others. Modeling educational process using expert system. html//intsys.msu.ru/en/staff

7. Шкурская Н.М. Использование компьютерных программ в обучении иностранным языкам//Язык, речь, общение в контексте диалога языков и культур: Сборник научных трудов.- Минск: Изд.центр БГУ,2012.-С.177-183.

8. Сосновкий С.А. ByHeart: Система для самообучения иностранным языкам (обзор программного обеспечения). - http://ifets.ieee.org/russian/depository/v3_i3/html/6.html

9. Xin LU Expert Tutoring and Natural Language Feedback in Intelligent Tutoring Systems Department of Computer Science, University of Illinois at Chicago, U.S. http://nlp.cs.uic.edu/PS-papers/ICCE06DSC14-XinLu.pdf

10. Alla Anohina Advances in Intelligent Tutoring Systems: Problem-solving Modes and Model of Hints. International Journal of Computers, Communications & Control Vol. II (2007), No. 1, pp. 48-55

11. Archana K Rane Intelligent Tutoring System For Marathi Dissertation.2005/ http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.7770&rep=rep1&type=pdf

12. Красильникова В.А. Теория и технологии компьютерного обучения и тестирования. - Москва-2009. 331 с.

13. Бухараев Р.Г., Сулейманов Д.Ш. Семантический анализ в вопросно-ответных системах. - Казань: Изд-во Казан. ун-та. - 1990. -124 с.

14. Сулейманов Д.Ш. Методология и принципы создания интеллектуального агента в системе анализа вопросно-ответных текстов // Системный анализ и семиотическое моделирование: материалы первой всероссийской научной конференции с международным участием (SASM-2011). – Казань: Изд-во «Фэн» Академии наук РТ, 2011. – С. 18-27.

15. Zadeh L.A. Fuzzy logic and approximate reasoning // Synthese, 1975. V. 80. P. 407–428.

16. Гогунский В.Д., Вишневская В.М., Буслаев А.Г. Анализ внедрения адаптивной обучающей программы на основе нечеткой логики. – Труды Одесского политехнического университета, 2007, вып.2(28), с. 127-128.

УДК 81’33

К ВОПРОСУ ОБ ИСПОЛЬЗОВАНИИ КОРПУСНО-ОРИЕНТИРОВАННОГО ПОДХОДА В ПРЕПОДАВАНИИ ТАТАРСКОГО ЯЗЫКА

А.А. Мубаракшина
Казанский федеральный университет, Казань
mubarakshinaaa@poelidovolen.ru
Б.Э. Хакимов
Казанский федеральный университет, Казань
НИИ “Прикладная семиотика”
Академии наук Республики Татарстан, Казань
khakeem@yandex.ru

Статья посвящена проблеме обучения национальным языкам на основе корпусно-ориентированного подхода на примере татарского языка. Рассматриваются общие методические аспекты использования корпусов текстов в обучении языку и возможности использования корпусов текстов при изучении татарского языка. Приведены примеры корпусных исследовательских заданий для студентов.

Ключевые слова: *корпус текстов, татарский язык, обучение языку, национальное образование.*

В современном процессе образования на первое место ставятся компьютерные и информационные технологии, которые стали неотъемлемой частью нашей жизни. День за днем люди пытаются упростить и ускорить различные процессы посредством использования современных технологий. Наряду с модернизацией, большое место также занимает изучение и использование различных языков. Также, на сегодняшний день, одним из самых актуальных вопросов является вопрос сохранения национальных языков, к которым относится и татарский язык.

В последние десятилетия в науке сформировалось понятие билингвального или полилингвального образования, под которым подразумевается обучение на двух и более языках. Тем не менее, стоит отметить, что двуязычное образование, ставшее столь популярным и актуальным в последние годы, уже давно практикуется в некоторых учебных заведениях Республики Татарстан. Татарстан является уникальным местом, где уже несколько веков представители более ста национальностей живут в согласии и в мире, наличие двух государственных языков – татарского и русского является – благоприятным фактором для развития билингвизма.

Современные исследователи, которые активно изучают как положительные, так и отрицательные стороны двуязычного образования, пришли к выводу, что преимущества билингвального обучения покрывают его недостатки и издержки. Учитывая тот факт, что в сегодняшнем мире двуязычному образованию уделяется очень много внимания, преподаватели и методисты ищут новые способы упрощения, совершенствования и повышения эффективности процесса билингвального образования. Компьютерные технологии также стали активно применяться в этой области. Все больше и больше понятий компьютерной лингвистики стали использоваться наряду с понятием «билингвальное образование». Среди таких понятий мы сегодня можем выделить и корпусную лингвистику.

Будучи сложным словесным единством, корпус может включать в себя не только разнообразную информацию о составе и структуре речевого материала, но и другие формализованные методы его представления. Исходя из данных фактов, мы можем сказать, что корпусы также можно рассматривать и как специально построенную семиотическую систему.

Так как корпус является уменьшенной моделью языка или подъязыка, его репрезентативность определяет достоверность данных, которые были получены. Репрезентативность – это не только объем полученных данных, но и пропорциональность представления отображаемого фрагмента речевой деятельности. Таким образом, мы с уверенностью можем сказать, что увеличение объема исследуемого корпуса не может означать увеличение достоверности. Самым важным фактом в данном случае является тщательная выборка текста и при планировании корпуса, и его использовании. Немаловажно учитывать все эти факты при дальнейшей работе с лингвистическими корпусами [1].

Именно сегодня, когда мы с уверенностью говорим о глобализации образовательных процессов и об эффективности двуязычного образования, необходимо подойти к этому вопросу со всей ответственностью и определить преимущества и недостатки. При отсутствии систематизированного изучения особенностей корпусно-ориентированного подхода в двуязычном образовании, его ценность для педагогической деятельности не будет полностью раскрыта и реализована, поэтому необходимо исследовать все аспекты данной проблемы, в том числе потенциальные преимущества и возможные недостатки.

Можно сказать, что до настоящего момента тема использования лингвистических корпусов в обучении татарскому языку не была глубоко исследована и все работы были направлены на частичный анализ того или иного конкретного аспекта. Однако разработанные лингвистические корпусы татарского языка могут уже сейчас использоваться педагогами, которые ведут занятия в двуязычных группах. Разработанные в ходе исследования практические рекомендации могут помочь при обучении татарскому языку.

Основные преимущества корпусно-ориентированного подхода в преподавании языка заключаются в следующем:

1) Корпус является источником реальных языковых примеров, «живого» языка.

2) Корпус дает возможность осуществлять обучение через исследование.

3) Использование корпусов способствует развитию общей информационной и лингвоинформационной компетенции обучающихся.

Создание текстовых корпусов татарского языка требует много времени, сил и исследований. С одной стороны, для создания лингвистического корпуса на татарском языке требуется кропотливая работа над текстом, с другой стороны, требуется создание и специальных компьютерных программ. Важную роль играет лингвистическая разметка, в процессе которой каждому слову присваивается информация о морфологических, семантических, либо других его особенностях. Благодаря этому пользователь автоматически находит нужную ему информацию.

В Казанском федеральном университете с 2012 года осуществляется преподавание лингвистических дисциплин на основе корпусного подхода. В частности, работа со студентами отделения татарской филологии Института филологии и межкультурной коммуникации проводится с использованием Национального корпуса татарского языка «Туган тел», а также других корпусов, таких как, например, Письменный корпус татарского языка. В учебный план студентов, обучающихся по направлению «Прикладная филология» включен специальный курс – «Корпусная лингвистика».

Ниже приведен пример исследовательского задания по корпусу татарского языка:

Задание. Найти в корпусе прилагательные-синонимы ***матур-гүзәл-чибәр*** *и исследовать их сочетаемость с определяемыми существительными.*

При выполнении этого задания в ходе поиска по корпусу студенты получают, например, такие результаты:

Матур (‘красивый’) – 22174 вхождений

Гүзәл (‘прекрасный’) – 6111 вхождений

Чибәр (‘красивый (о человеке)’) – 4921 вхождений.

На основе этих результатов делается общий вывод о частоте употребления разных синонимов, экспериментально подтверждается в принципе очевидное суждение о доминанте данного синонимичного ряда (*матур*).

Далее студентам предлагается исследовать сочетаемость рассматриваемых синонимов с различными существительными. Пример результатов по этому заданию приведен в таблице 1.

Таблица 1

Сочетаемость синонимов ***матур-гүзәл-чибәр***
по корпусу татарского языка

Синоним	Кеше (‘человек’)	Кием (‘одежда’)	Күлмәк (‘рубашка, платье’)	Йорт (‘дом’)	Көн (‘день’)
Матур	88	84	105	70	180
Гүзәл	41	0	0	6	5
Чибәр	43	0	2	10	3

Аналогичным образом, могут быть разработаны задания на исследование частотности и сочетаемости других классов лексики, например, наименований цветов (колоронимов) и др.

В завершение необходимо отметить, что дальнейшие разработки и методические изыскания в области корпусного преподавания татарского языка будут способствовать реализации указанных преимуществ корпусно-ориентированного подхода и повысят качество и результативность изучения языка.

Литература

1. Баранов А.Н. Введение в прикладную лингвистику. М: Эдиториал УРСС, 2001. – 360 с.

2. Добровольский Д.О. Корпус параллельных текстов как инструмент сопоставительного описания языков / Д.О. Добровольский // Русская и сопоставительная филология: состояние и перспективы: Международная научная конференция, посвященная 200-летию Казанского университета (Казань, 4-6 октября 2004 г.): Труды и материалы: / Под общ. ред. К.Р. Галиуллина.– Казань: Изд-во Казан. ун-та, 2004.

3. Салехова Л.Л. Когнитивные издержки билингвального обучения // Филология и культура. Philology and Culture. – 2015. – № 2 (40). – С. 314–319.

4. D.Biber, S. Conrad, R. Reppen (2001). Corpus Linguistics: Investigating language structure and use. - Cambridge University Press, 2001.

5.Национальный корпус татарского языка «Туган тел». Интернет-ресурс: http://corpus.antat.ru.

6. Письменный корпус татарского языка. Интернет-ресурс: http://corpus.tatfolk.ru.

УДК 372.3/.4

ИСПОЛЬЗОВАНИЕ ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ В СПЕЦИАЛЬНОМ ДОШКОЛЬНОМ ОБРАЗОВАНИИ

М.М. Романенко, Р.Р. Зарипова, Л.Л. Салехова
Казанский федеральный университет, Казань
MMRomanenko@stud.kpfu.ru**,** rinata-z@yandex.ru,
salekhova2009@gmail.com

В статье рассматриваются различные ИКТ, разработанные отечественными и зарубежными учеными, предназначенные для детей дошкольного возраста с различными нарушениями в развитии (нарушение слуха, нарушения зрения, дети с синдромом дефицита внимания и гиперактивности и др.).

Ключевые слова: *ИКТ, дошкольные организации, специальное образование, логопеды.*

Информационно-коммуникационные технологии (ИКТ) в настоящее время широко применяются в дошкольных организациях. В современных условиях большинство дошкольных организаций оснащено современными техническими средствами, с помощью которых происходит реализация компьютерных технологии обучения. Они усиливают мотивацию детей к усвоению новых знаний, а также открывают неограниченные возможности для самостоятельной и совместной творческой деятельности воспитателей, детей и их родителей. Кроме того, занятия, проводимые с использованием ИКТ, позволяют детям овладеть навыками чтения, рисования и письма [1,28-30]. Немаловажную роль ИКТ играют в специальном образовании детей с ограниченными возможностями здоровья, т.е. детей с речевыми патологиями, нарушениями опорно-двигательного аппарата, зрения, слуха, интеллекта. Применение ИКТ позволяет активизировать компенсаторные механизмы и достичь оптимальной коррекции нарушенных функций. Многообразие дефектов, их клинических и психолого-педагогических проявлений предполагает применение разных методик коррекции, а, следовательно, и использование различных ИКТ. Поэтому разработка новых приёмов, методов и средств коррекционного обучения детей представляется нам актуальным и значимым.

В последнее время все больше внимания уделяется обучению детей дошкольного возраста, которые нуждаются в специальном образовании. Исследования методик обучения на основе информационно-коммуника-

ционных технологий обладают необходимым потенциалом. Позволяют значительно повысить эффективность коррекционно-образовательного процесса [2, 234-246]. Вопросами использования ИКТ в процессе специального обучения в дошкольных организациях занимались как отечественные, так и зарубежные ученые. В нашей стране внедрение ИКТ в дошкольное специальное образование ведется с 1987 года. Так, научно-исследовательский центр «Дошкольное детство» им. А.В. Запорожца занимается вопросами реабилитации глухих и слабослышащих детей и их интеграции в общество слышащих, развитием речи речевых технологий дошкольников на основе ИКТ. Исследования Е.В. Зворыгиной, Н.Ф.Талызина, Н.Н. Малофеев, Н.П. Чудова, С. Пейперт, Б. Хантер, Е.Н. посвящены разработке и применению развивающих компьютерных игр для развития мыслительных операций в специальном дошкольном образовании. Психолого-педагогические и дидактические аспекты использования компьютерных технологий в процессе общего образования (Я.А. Ваграменко, А.А. Кузнецов, Е.И. Машбиц, Е.С. Полат, И.В. Роберт), специального образования (В.П. Беспалько, Л.Р. Лизунова, Гончарова Е.Л., Кукушкина О.И., Королевская Т. К.) дошкольного образования в коррекции нарушений речи рассмотрены в работах Ю.Б. Зеленской, Т.К. Королевской, О.И. Кукушкиной, Л.Р. Лизуновой, И.А. Филатовой [3,126-156].

Таким образом, различные специалисты, занимающиеся проблемами специального образования, сходятся во мнении, что ИКТ способствуют развитию учебных навыков ребенка, а также могут помочь в создании развивающей образовательной среды в дошкольном учреждении. Проведенные исследования подтверждают, тот факт, что ИКТ могут помочь детям с трудностями обучения, с сенсорными и физическими нарушениями. Кроме того, одаренные и дети с несбалансированным билингвизмом (вид индивидуального двуязычия, характеризующийся различным уровнем языковой компетенции билингва) могут также испытывать трудности в обучении, которые могут быть преодолены с помощью ИКТ. Существующие исследования доказывают, что использование ИКТ в специальном дошкольном образовании предоставляют детям дополнительные возможности заключающиеся в повышении наглядности, разнообразия содержания и формы подачи материала [1,47-50].

В результате, ИКТ играют существенную роль в решении задачи создания благоприятных условий развития детей в соответствии с их возрастными и индивидуальными особенностями и склонностями, развития способностей и творческого потенциала каждого ребенка как субъекта отношений с самим собой, с другими детьми, взрослыми и миром, указанной во ФГОС дошкольного образования (Приказом МОиН РФ №1155 от 17 октября 2013 года) в качестве одной из приоритетных [4]. Кроме того, дети с особыми педагогическими потребностями имеющие возможность обучаться по специальным образовательным программам

в дошкольных учреждениях, испытывают меньше затруднений при их дальнейшем обучении в начальной школе. Положительный эффект от внедрения ИКТ в специальное образование убедил преподавателей начать их широкое использование в дошкольных учреждениях. Рассмотрим разработанные отечественными и зарубежными учеными средства ИКТ для детей с различными нарушениями (Таблица 1).

Таблица 1

Программное обеспечение для детей
с различными нарушениями в развитии

Категории детей с различными нарушениями в развитии	Название программы и её автор(ы)	Описание программы
Нарушения слуха	«Мир за твоим окном» О.И. Кукушкина, Т.К. Королевская, Е.Л. Гончарова, 1997; О.И. Кукушкина	Программа состоит из пяти частей: «Четыре времени года»; «Погода», «Одежда»; «Рассказы о временах года»; «Календарь». Предназначена детям, испытывающим трудности в обучении, детям с различными нарушениями.
	“Sign my World” «Мой мир» (Auslan).	Приложение представлено в качестве мобильной версии видеоигры в целях ознакомления с существительными и глагольными признаками.
	«Digital interactive storybook» Интерактивный сборник рассказов; Yen(Йен) и Lee(Ли)	Основан на голосовой жестовой конструкции. Используя преимущества портативного цифрового устройства с сенсорным экраном, включены обучение и проектирование в сюжетную линию.
Нарушения зрения	«Сиолл» Айдар Фахрутдинов	Информация в учебник загружается через usb-привод и преобразуется на экране в текст шрифтом Брайля. «Ввод текста осуществляется с помощью особого стилуса».
	«Multimodal computer system» «Мультимодальная компьютерная система» Raisamo	Эта система обучения, состоящая из шести микро-слов, которые представляют астрономическое явление, кото-рое учащиеся могут изучить самостоятельно.
Нарушения письменной речи	«PHAES» (Phonological Awareness Educational Software) Фонологическая	Фонологическая осведомленность образовательного программного обеспечения гипермедиа-приложение разработан как инструмент вмешательства для учащихся с дислексией, а также

	осведомленность образовательного программного обеспечения	используется для оценки и успешного сопоставление букв и соответствующих им звукам.
	«MAPS» (Mental Attributes Profiling System) Loizou и Laouris	Программа психические признаки профильной системы позволяет оце-нить познавательные способности. Он состоит из восьми независимых языко-вых тестов, которые измеряют различ-ные аспекты обучения.
СДВГ Синдром дефицита внимания и гиперактивности	«Внимание» Разработчик Effecton Studio	Программа включает уникальную коллекцию из 14 тестов и 15 упраж-нений, позволяющих детально исследо-вать и развивать все основные свойства внимания.
	CAI Keller & Keller	Программа позволяет каждому учаще-муся работать с модулем в течение двадцати минут за сеанс, а также записывает физические и словесные реакции каждого учащегося.
Autistic Spectrum Disorders (ASD) С аутистическим спектром	«Аутизм: Общение» Компания Game Garden	Приложение содержит: 1. Коммуни-катор, при помощи которого ребенок может обозначать предметы, состав-лять полноценные предложения-просьбы. 2. Галерею карточек, содер-жащую более 150 качественных изображений, которые ребенок учится различать, наименовать и соотносить с различными категориями.
	«Let's Face It!» «Давайте посмотрим правде в глаза» Tanaka и соавторы	Программа состоит из семи интерак-тивных компьютерных игр, которые направлены на конкретные лица, связанные с заболеванием аутизма.
Одаренные дети	«Кирилл и Мефодий» энциклопедия	Компьютерная энциклопедия содержит разнообразные справочные сведения о различных сферах, принципы работы, англо-русский словарь компьютерных терминов и многое другое.
	«RAPYHT» Clark	В программе проверяется реакция на яркие настенные дисплеи, мультимеди-йные технологии.
Билингвы	«Супердетки» MultiSoft	Программа содержит множество увлекательных заданий и упражнений, которые помогают всесторонне развить ребенка.
	«Pacific Island people» Жители тихого океана Samoans	Инновационный продукт обучения способствует приобретению цифровых, лингвистических и работу с культур-ными традициями конкретных языковых сообществ.

Таким образом, рассмотренные прикладные компьютерные программы могут быть использованы в дошкольных организациях, и могут быть применимы на всех этапах обучения (актуализации прежних знаний, объяснения нового материала, закрепления нового материала). Эффективность применения ИКТ с её мультимедийными возможностями в специальном дошкольном образовании позволяет сделать процесс обучения и развития ребенка достаточно успешным, открывая новые возможности взаимодействия в образования не только для самого ребенка, но и для педагога.

Литература

1. Захарова И.Г. Информационные технологии для качественного и доступного образования // Педагогика. - 2012.- №1.- С. 27-33

2. Калинина Т.В. Управление ДОУ. «Новые информационные технологии в дошкольном детстве», М.: Сфера, 2011.- С.360

3. Кукушкина О.И. Информационные технологии в специальном образовании: концептуальные идеи и их практическая реализация // Хрестоматия к курсу "Информационные технологии в специальном образовании". Разд.I,2013.- С.440

4. Приказ Минобрнауки России № 1155 от 17 октября 2013 г. «Об утверждении федерального государственного образовательного стандарта дошкольного образования»

УДК 378.147:811.11

СЕРВИСЫ WEB 2.0 В ОБУЧЕНИИ ИНОСТРАННЫМ ЯЗЫКАМ

Л.Л. Салехова
Казанский федеральный университет, Казань
salekhova2009@gmail.com
К.С. Григорьева
Казанский национальный исследовательский технический университет им. А.Н. Туполева-КАИ, Казань
grigks@yandex.ru

В статье описываются возможности использования сервисов web 2.0 в образовательном процессе высших учебных заведений. Представлена классификация сервисов web 2.0 и опыт их использования в процессе иноязычной подготовки студентов в техническом вузе.

Ключевые слова: *web 2.0 технологии, сервисы второго поколения, технология социальной сети, обучение иностранным языкам, иноязычная подготовка в техническом вузе, CLIL.*

В настоящее время информационные технологии, в частности, сервисы web 2.0 находят все большее применение в образовательном пространстве вуза. В последнее время система образования России становится все более оснащенной эффективными средствами телекоммуникации. Все участники образовательного процесса имеют доступ к глобальной сети Интернет. В связи с развитием сервисов второго поколения сеть Интернет становится не только источником разного рода информации, но и средой взаимного общения преподавателей и студентов.

Сервисы Web 2.0 ли Интернет-технологии второго поколения представляют собой Интернет-ресурсы, где каждый из пользователей может принять активное участие в формирование контента. Для этого не нужно обладать специальными знаниями в области программирования. Достаточно лишь выработать несколько навыков (например, создание онлайн презентации, записи аудио- или видеофайла и т.д.).

Возможности, которые открывает использование данных технологий в образовательном процессе вуза, рассматриваются в работах многих российских и зарубежных авторов [1-7]. Однако, несмотря на большой интерес ученых к данной проблеме в научно-методической литературе нет единой формулировки данного понятия применительно к сфере образования. Так, например, С.В.Титова и А.В.Филатова[5]трактуют web 2.0 как «платформу для социального взаимодействия». ТимО'Рейли [2], исследователь, благодаря которому термин «web 2.0» получил широкое распространение, полагает, что данное понятие не имеет четких границ. Следовательно, дать ему четкое определение не представляется возможным.

Различные характеристики сервисов web 2.0 предложены в работах P.Anderson, R.Dawson, Е.Д.Патаракинаи т.д.[6, 7, 4]. В частности, R.Dawson выделил семь базовых социальных характеристик web 2.0, а именно: участие, стандарты, децентрализация, открытость, модульная структура, контроль со стороны пользователей, идентичность [7].

Интерес представляет классификация, разработанная исследователями университета г. ХьюстонB. Robinи S. McNeil [3]. Они классифицируют сервисы web 2.0 в зависимости от целей коммуникации, которые ставит перед собой преподаватель в ходе конкретного занятия.Так согласно данной классификации все сервисы web 2.0 можно разделить на сервисы, предполагающие:

- одностороннюю коммуникацию – обратная реакция на представленный контент не предполагается (GoogleSites, PodOmatic, Feedly и т.д.);
- двустороннюю коммуникацию – пользователь может делиться информацией и получать ответ (PollEverywhere, Blogger, Boardhost, YahooMailи т.д.);

- комплексную коммуникацию – предполагает синхронную и асинхронную коммуникацию между большим количеством людей – (GoogleHangout, GoogleDocsи т.д.).

Применение в образовательном процессе сервисов web 2.0, открывающих большие возможности для осуществления комплексной коммуникации в формате «преподаватель – студент», «студент – студенты», «преподаватель – студенты» и т.д., приобретает особую актуальность в связи с постоянным сокращением аудиторных часов, выделяемых на изучение предметов гуманитарного цикла, в том числе «Иностранный язык» в техническом вузе.

Идея использования социальных сетей в процессе преподавания гуманитарных дисциплин, особенно иностранных языков, в основе преподавания которых лежит принцип коммуникативной направленности, является перспективным направлением. Современные Web 2.0 технологии, на базе которых строится образовательная социальная сеть, позволяют осуществлять аудио- и видеодиалог между преподавателем и студентом в режиме реального времени, создавать собственные текстовые, аудио- и видеоматериалы, проводить обсуждение пройденного материала и различного рода проблем, а также осуществлять общение с носителями языка в on-line режиме[1].

Опыт использования сервисов web 2.0 в процессе иноязычной подготовки в техническом вузе.

Образовательная социальная сеть «Country Study» организована на базе студенческого лингвострановедческого кружка с одноименным названием, функционирующего на кафедре иностранных языков КНИТУ-КАИ им. А.Н. Туполева. Общение происходит на английском языке.

Цель деятельности образовательной социальной сети «CountryStudy» – способствовать формированию и развитию иноязычной компетенции КГТУ-КАИ им. А.Н. Туполева, необходимой для осуществления коммуникации в профессиональной деятельности, а также информационной компетенции, которая предполагает развитие следующих навыков и умений:

- ориентации в информационных потоках современного общества;
- эффективной работы по поиску информации;
- критической оценки и отбора информационных ресурсов;
- координирования совместных действий в процессе компьютерной коммуникации;
- наличие навыков компьютерно-опосредованной коммуникации [5].

Преподаватель, являясь основным модератором поступающего контента, обладает правом принимать или отклонять новых пользователей,

что, в свою очередь, ограничивает количество участников социальной сети студентами определенного ВУЗа или группы. Таким образом, преподаватель выполняет функции цензора.

Студенты выступают в качестве участников, каждый из которых, получает доступ ко всем ресурсам проекта, а также возможность оставлять собственные комментарии, публиковать статьи, видео-, аудиоролики, презентации и оформить собственную веб-страницу в данной социальной сети. Основная задача, поставленная в процессе работы надИнтернет ресурсом «Countrystudy», – организация иноязычной образовательной среды для дополнительного, внеаудиторного общения между преподавателем и студентами на учебные темы. Подобное взаимодействие становится возможным благодаря использованию Web 2.0. технологий, лежащих в основе Интернета второго поколения. Благодаря различным сервисам и приложениям студенты имеют возможность активно участвовать в наполнении сайта. Наибольшую активность пользователей вызывает просмотр видеофайлов и их дальнейшее комментирование.

Следует отметить, что возможность оставить собственный комментарий позволяет, в первую очередь, преподавателю внести коррективы в комментарий студента, исправив имеющиеся ошибки и дав подробное объяснение, что не всегда возможно сделать в рамках аудиторного занятия, где преподаватель ограничен во времени. Для преподавателя опубликованные работы студента - это возможность осуществлять контроль за их самостоятельной работой. В свою очередь, студент–пользователь получает возможность анализировать и развивать навыки письменной речи, публиковать свои мысли во всемирной сети, продолжить начатую на практическом занятии дискуссию.

Учебная группа на базе социальной сети «В контакте»

С целью организации дополнительного образовательного простран-ства для осуществления дополнительного иноязычного общения на базе популярной социальной сети «В контакте» была организована специаль-ная учебная группа «EnglishforstudentsofKAI» https://vk.com/ club58822551. Данная страница выступает в качестве своеобразной информационной площадки для студентов, изучающих дисциплину «Профессиональный английский язык». В ходе изучения дисциплины перед нами стояла задача не только сформировать у студентов иноязычную компетенцию в сфере профессиональной коммуникации, но и развить умения и навыки использования ИКТ, в частности сервисов web 2.0, в образовательных целях.

На странице «English for students of KAI» преподаватель размещает изучаемый учебный материал, выкладывает методические рекомендации и примеры выполнения того или иного задания, а также работы студентов. Примером подобного взаимодействия является задание на создание

с помощью ресурса Voicethread презентации по пройденному лексическому материалу. Для выполнения данного задания студенту необходимо зарегистрироваться на сайте https://voicethread.com (подробные инструкции о выполнении задания размещены в группе «English for students of KAI»), посмотреть пример выполнения подобного задания (пример, созданный преподавателем https://voicethread.com/share/409/), создать собственнуюVoicethread, в которой рассматривается изучаемая лексика, согласно следующему плану:

1. дефиниция слова или словосочетания на английском языке;
2. примеры использования изученной лексики в контексте;
3. возможные варианты перевода слова или словосочетания на русский язык.

В рамках составления Voicethread студенты используют все возможности данного ресурса (запись видео-, аудиокомментария, письменное комментирование и т.д.). Ссылка на выполненное задание размещается в группе в vk.com. Таким образом, преподаватель получает возможность проверить навыки произношения, перевода, степень усвоения изученного лексического материала и т.д.

Еще одно задание, которое так же, как и предыдущее, может быть использовано для контроля усвоения и оценки полученных знаний, направлено на проверку не только языкового аспекта, но и степени усвоения контента. Поскольку изучаемый контент содержит новый с точки зрения содержания материал, объясняющий такие понятия, как «аэродинамические силы», силы, оказывающие влияние на летательный аппарат во время полета, центр тяжести, поверхностное трение, индуктивное сопротивление и т.д., возникает необходимость оценить степень его понимания и усвоения. С этой целью студентам предложено составить презентацию в программе PowerPoint, где они подробно должны рассмотреть и описать воздействие аэродинамических и механических сил на летательный аппарат, используя изученную лексику, а также дополнительный материал и ссылки по теме, размещенные на страницах группы в социальной сети. (Например, ссылка на фильм «The Aerodynamics of Flight» (http://www.youtube.com/watch?v=5ltjFEei3AI), а также текст фильма (аудиоскрипт)). Презентация и последующее обсуждение задания проводятся на аудиторном занятии.

Отметим, что постоянная работа и общение в социальной образовательной сети делает непрерывным процесс пополнения знаний студентов, а также процесс развития навыков и умений общения на иностранном языке. Следует также отметить, что использование новых информационных технологий в процессе обучения иностранным языкам

является сильным мотивирующим фактором для студентов с различным уровнем иноязычной подготовки.

Литература

1. Григорьева К.С. Использование социальных сетей в обучении английскому языку студентов неязыковых специальностей / К.С. Григорьева // Информатика и образование.- М.: Изд-во «Образование и Информатика». – 2011. – №4. – С.57-61.
2. О'Рейли, Т. Что такое Веб 2.0 // Компьютерра. 2005. №37 (609), №38 (610).
3. Робин, Б.,МакНейл,С. Мощныеинструментыдляпреподаванияобучения: инструментыweb 2.0// ЭлектронныйресурсURL: https://www.coursera.org/ course/ newtechtools
4. Патаракин, Е.Д. Формирование личного учебного пространства в сети электронных коммуникаций // EducationTechnologyandSociety. 2008. № 11(2). С. 416-425.
5. Титова,С.В., Филатова,А.В. Технологии Веб 2.0 в преподавании иностранных языков / С.В. Титова, А.В. Филатова; Мос. гос. ун-т им. М.В. Ломоносова, Факультет иностран. яз.и регионоведения. – 2-е изд., перераб. и доп. – М.: Издательский дом «Квинто-Консалтинг», 2010. – 100 с.
6. Anderson, P. What is Web 2.0? Ideas, technologies and implications for education //JISC Technology and Standards Watch. 2007. URL:http://www.jisc.ac.uk/media/ documents/techwatch/tsw070lb.pdf
7. Dawson, R. (2007) Launching the Web 2.0 Framework. URL: http://www. rossdawsonblog.com/weblog/archives/2007/05/launching_the_w.html

УДК 372.881.1

ВНЕДРЕНИЕ WEB 2.0-ТЕХНОЛОГИЙ НА БАЗЕ ПРОГРАММЫ NEARPOD В ОРГАНИЗАЦИЮ САМОСТОЯТЕЛЬНОЙ РАБОТЫ УЧАЩИХСЯ СРЕДНЕЙ ШКОЛЫ ПО АНГЛИЙСКОМУ ЯЗЫКУ

А.Н. Ульянова
Казанский федеральный университет, Казань
anna.ulyanova.92@mail.ru

В статье рассматриваются технологии Web 2.0, а именно Web 2.0-платформы Nearpod в процессе организации самостоятельной работы учащихся средней школы во время проведения занятий по изучению английского языка. Нами описываются преимущества данного вида самостоятельной работы, а также результаты проделанной экспериментальной работы.

Ключевые слова: *Web 2.0-технологии, Nearpod, самостоятельная работа учащихся*

С развитием процессов глобализации перед средним образованием ставятся новые цели - подготовка выпускников, которые в будущем будут способны эффективно трудиться в кардинально изменившихся условиях глобального рынка. Важнейшим инструментом обновления знаний служит глобальная сеть Internet, которая предоставляет широкие возможности для организации самостоятельной работы учащихся. В настоящее время сложно представить современную глобальную сеть без блогов, поисковых систем, социальных сетей. Таким образом, появляется группа сервисов, разработанных на основе Web 2.0-технологий, где мы можем наблюдать активное участие пользователей в формировании контента.

Современные учащиеся средней школы быстро адаптируются к изменениям в компьютерных технологиях, следовательно, этот факт предполагает прекрасную возможность использования Web 2.0-технологий в образовательных целях.

Для успешного осуществления учащимися самостоятельной работы на уроках английского языка понадобится не только доступ к сети Internet, но также и использование смартфонов, планшетных компьютеров и других приспособлений для обучения. Таким образом, это означает использование потенциала имеющихся у учащихся гаджетов для работы с обучающими приложениями на английском языке. Следовательно, на настоящий момент можно констатировать готовность социума к использованию Web 2.0 технологий не только для повседневных, но и для образовательных целей.

Одной из программ, позволяющей учащимся активно осуществлять самостоятельную работу на уроках английского языка, является программа Nearpod. Nearpod является Web 2.0-платформой, которая позволяет учителю создавать презентации и подбирать материал к занятиям по английскому языку и делиться ими с учащимися во время урока. Вы высылаете по электронной почте или называете ученикам код доступа к вашей презентации, и учащиеся подключаются к вашему устройству со своих смартфонов или планшетных компьютеров. Вы листаете слайды презентации или текстового документа, сами задаёте и контролируете темп урока, вовлекаете учащихся в выполнение заданий по английскому языку и в реальном времени отслеживаете результаты каждого ученика. Для контроля результативности урока учителю достаточно лишь загрузить с помощью программы на своё устройство какой-либо заранее составленный тест, грамматическое задание, задание на сопоставление и т.д. Также с помощью программы Nearpod учащиеся могут добавлять в общую презентацию свой контент: видеофайлы, изображения, ссылки на различные сайты.

Имея чёткое представление о достоинствах Web 2.0-платформы Nearpod, нами было проведена экспериментальная работа по использованию данной

программы в рамках самостоятельной работы учащихся средней школы на уроках английского языка. Экспериментальная работа проводилась нами в три этапа: констатирующий, формирующий и контрольно-оценочный. В реализации эксперимента участвовали учащиеся II подгруппы 6 Б класса, МБОУ СОШ №27 города Чебоксары (13 учеников).

Основной целью констатирующего этапа (2014 - 2015 учебный год) экспериментальной работы явился анализ организации самостоятельной работы учащихся и изучение Web 2.0-ресурсов, в частности программы Nearpod. В ходе этого этапа нами были использованы такие методы исследования, как изучение и анализ научной, психолого-педагогической и методической литературы.

Формирующий этап экспериментальной работы проводился так же в 2014 - 2015 учебном году (с 6 апреля по 22 мая 2015 года). На данном этапе нами было проведено внедрение Web 2.0-платформы Nearpod в организацию самостоятельной работы учащихся средней школы, проводилась аудиторная работа с учениками, рассматривались критерии оценивания результатов внедрения.

В ходе проведения формирующего этапа учащимся было предложено разделиться на две группы (по 7 и по 6 человек в каждой группе). Первой группе было предложено использовать собственные смартфоны и планшетные компьютеры на уроках английского языка для просмотра подготовленных нами презентаций и выполнять задания к ним в режиме реального времени, второй группе было дано задание – просмотр тех же презентаций на экране, а затем выполнение заданий к ним (письменно или устно). Следует отметить, что начальный уровень мотивации учащихся в первой группе был намного выше, поскольку им был предложен новый вид работы, к тому же у них в руках были собственные гаджеты. Во второй группе учащиеся откликнулись на задание без энтузиазма, поскольку этот способ усвоения новой информации повторяется на уроках английского языка довольно часто.

Анализируя результаты самостоятельной работы учащихся в течение первого месяца (апрель), их оценку (по 100-балльной системе), можно выделить следующие данные об успешности обучения учащихся (Таблица 1).

Таблица 1

Результаты самостоятельной работы учащихся за апрель месяц

№ п/п	Контрольные точки	Средний балл по группам	
		1 группа (7 учащихся)	2 группа (6 учащихся)
1	Результаты выполнения грамматических заданий (максимум 20 баллов)	18 баллов	14 баллов
2	Результаты групповых обсуждений (максимум 20 баллов)	16 баллов	14 баллов
3	Результаты составления диалогов по обсуждаемой теме (максимум 20 баллов)	17 баллов	16 баллов
4	Результаты самостоятельного поиска дополнительного материала по пройденному разделу (максимум 20 баллов)	20 баллов	12 баллов
5	Результаты итогового группового задания – создание презентации по пройденному разделу (максимум 20 баллов)	20 баллов	17 баллов
Итого		91 балл	73 балла

Через 4 недели мы поменяли способ демонстрации нового материала для учащихся. Теперь с программой Nearpod работали учащиеся второй группы (6 человек). Результаты самостоятельной работы учащихся представлены ниже (Таблица 2).

Таблица 2

Результаты самостоятельной работы учащихся за май месяц

№ п/п	Контрольные точки	Средний балл по группам	
		1 группа (6учащихся)	2 группа (7учащихся)
1	Результаты выполнения грамматических заданий (максимум 20 баллов)	20 баллов	16 баллов
2	Результаты групповых обсуждений (максимум 20 баллов)	17 баллов	15 баллов
3	Результаты составления диалогов по обсуждаемой теме (максимум 20 баллов)	18 баллов	17 баллов

4	Результаты самостоятельного поиска дополнительного материала по пройденному разделу (максимум 20 баллов)	19 баллов	15 баллов
5	Результаты итогового группового задания – создание презентации по пройденному разделу (максимум 20 баллов)	20 баллов	20 баллов
Итого		94 балла	83 балла

Выполнив математические подсчёты, мы можем отметить, что результаты самостоятельной работы учащихся с использованием Web 2.0-платформы Nearpod оказались выше на 15,7 %.

Контрольно-оценочный этап нашей экспериментальной работы проводился в 2014 - 2015 учебном году (май 2015 года). Также нами был проведен анонимный опрос учащихся. Более 90% учащихся предпочли работу с Web 2.0-платформой Nearpod обычному просмотру презентаций на экране и устной (или письменной) работе над просмотренной презентацией.

Результаты экспериментальной работы полностью подтвердили предположение о том, что внедрение Web 2.0-платформы Nearpod в организацию самостоятельной работы учащихся средней школы на занятиях по английскому языку позволит повысить ее эффективность.

Таким образом, можно сделать вывод, что в рамках самостоятельной работы учащихся Web 2.0-платформа проявила большой потенциал и нашла положительный отклик у учащихся средней школы.

Литература

1. Патаракин, Е.Д. Образовательные возможности Веб 2.0. Веб 2.0- сервисы Интернета - новые формы коллективного педагогического взаимодействия. [Электронный ресурс] / Е.Д. Патаракин // Новые возможности в обучении. - 2008. - URL: http://eelmaa.net/dld/web20.pdf. (дата последнего обращения: 13.10.15);

2. Патаракин, Е.Д. Социальные сервисы Веб 2.0. в помощь учителю: практическое руководство / Е.Д. Патаракин. - М.: Интуит.ру. - 2007. - 71 с.

3. Hargadon, S. Web 2.0. Is the Future of Education / S. Hargadon. - URL: http://wvvw.tecMeaning.com/bto.php. (дата последнего обращения: 13.10.15);

4. O'Reilly, Tim What is web 2.0? [Электронный ресурс] / Tim O'Reilly. - URL: http://www.computerra.ru/think/234100.html. (дата последнего обращения: 13.10.15);

5. https://www.nearpod.com/ (дата последнего обращения: 03.10.15)

УДК 519.683.8

РАЗРАБОТКА WEB-ПРИЛОЖЕНИЯ ДЛЯ УДАЛЕННОГО РЕШЕНИЯ МАТЕМАТИЧЕСКИХ ЗАДАЧ

Л.Э. Хайруллина, М.М. Загидуллин
Казанский федеральный университет, Казань
liliya-v1@yandex.ru, Zagmysa@mail.ru

В статье описывается способ разработки Web-приложения для удаленного решения математических задач средствами MATLAB и Asp.net.

Ключевые слова: *Web-приложение, MATLAB, Asp.net*

Введение. В современном темпе жизни все большей актуальности и популярности приобретает движение в области разработки и создания Web-приложений, приходящие на смену классическим программам. Web-приложение - это компьютерная программа, которая размещена в сети Интернет и доступна пользователям через браузер в их компьютере, или даже мобильном телефоне. Многие из Web-приложений прочно вошли в нашу жизнь, например, приложения для общения по сети, для купли/продажи товаров, для перевода текстов и другие. Данная работа посвящена описанию способа разработки Web-приложения для удаленного решения математических задач средствами MATLAB и Asp.net.

Краткая характеристика MATLAB и Asp.net. MATLAB - это высокоуровневый язык и интерактивная среда для программирования, численных расчетов и визуализации результатов. MATLAB представляет собой основу всего семейства продуктов MathWorks и является главным инструментом для решения широкого спектра научных и прикладных задач. Продукты MathWorks для развертывания приложений позволяют избежать затратной и ненадежной реализации алгоритмов MATLAB на другом языке программирования. Пользователь, работая с исходным кодом в MATLAB, с легкостью может разрабатывать и обновлять алгоритмы, автоматически создавать на их основе самостоятельные приложения или компоненты программ и встраивать их в такие среды, как C, C++, Java ™, .NET, и Excel [3].

В мире программирования на данный момент существует технология создания Web-приложений и Web-сервисов от компании Майкрософт, которая называется ASP.NET (Active Server Pages). ASP.NET - технология создания веб-приложений и веб-сервисов от компании Microsoft. Она яв-

ляется составной частью платформы Microsoft.NET и развитием более старой технологии Microsoft ASP. На данный момент последней версией этой технологии является ASP.NET5 [1, 4].

Описание разработки Web-приложения для удаленного решения математических задач. Разработка Web-приложения начинается с создания М-файлов в MATLAB [2].

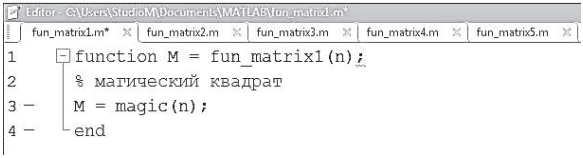


Рис. 1. Пример создания М-файла

Файлы сохраняются в рабочем каталоге. Созданную функцию можно использовать так же, как и встроенные функции. Вызов собственных функций может осуществляться из файл-программы или из другой файл-функции.

Для того, чтобы поделиться своим приложением MATLAB в виде исполняемого приложения или разделяемой библиотеки, необходимо воспользоваться MATLAB Compiler, которая позволяет запускать приложения без установленной среды MATLAB. Такая архитектура значительно уменьшает время разработки приложения, т.к. исчезает необходимость ручного перевода созданного кода на другой язык программирования. В случае создания независимого приложения MATLAB Compiler создает исполняемый файл для использования конечным пользователем.

Для создания библиотеки .dll необходимо выбрать Library Compiler и в появившемся графическом окне выполнить следующие шаги:

1. выбрать .NET Assembly;
2. выбрать ранее созданный m-файл (например, у нас это fun_matrix1.m);
3. написать название библиотеки *fun_matrix1*;
4. написать класс *Class_fun_matrix1*;
5. запустить упаковку библиотеки.

Теперь, когда создана библиотека с функциями MATLAB, можно приступить к созданию проекта ASP.NET. В Microsoft Visual Studio 10.0 создадим Web-приложение ASP.NET. Выберем имя и место расположения проекта. Перед использованием методов проекта, необходимо добавить ссылки на скомпилированную библиотеку *fun_matrix1.dll* и на библиотеку MWArray.dll, которую можно найти по адресу \MATLAB\R2010a\toolbox\dotnetbuilder\bin\win32\v2.0.

Создадим страницу сайта. Для этого в проект добавим форму Web Form и дадим ей имя.

Алгоритм работы приложения должен быть следующим:

1. Получение функции в символьном виде с TextBox1.
2. Вызов метода *fun_matrix1* из класса *Class_fun_matrix1*.
3. Получение выходного массива *output* (тип MWNumericArray).
4. Вывод массива *output* в TextBox1

Для использования библиотек в проекте необходимо добавить описание пространства имен:

```
using fun_matrix1;  //скомпилированная библиотека
using MathWorks.MATLAB.NET.Arrays; //библиотека самого MATLAB Compiler
using MathWorks.MATLAB.NET.Utility; //библиотека самого MATLAB Compiler
```

Инициализируем событие Button1_Click и напишем код:

```
protected void Button1_Click(object sender, EventArgs e)
 {
 MWNumericArray output = new MWNumericArray(); //массив возвращаемого параметра
 MWArray[] result = new MWArray[100]; //выходной массив метода fun_matrix1
 Class_fun_matrix1 obj1 = new Class_fun_matrix1(); //экземпляр класса компонента
 Label1.Text = " ";
try
 {
 int n = Convert.ToInt32(TextBox1.Text); // в числовую переменную n записываем значение из TextBox1
 result = obj1.fun_matrix1(1, n); // обращаемся к библиотеке с переменным n, через экземпляр класса компонента
output = (MWNumericArray)result[0]; // вывод вычисления в массив
 TextBox2.Text = Convert.ToString(output); // выводим результат в TextBox2
 }
catch (Exception)  //обработка исключения
 {
 Label1.Text = "Проверьте правильность ввода данных"; // сообщение об ошибке
 TextBox2.Text = " ";
 }
}
```

Запустим проект, браузер откроет страницу нашего сайта.

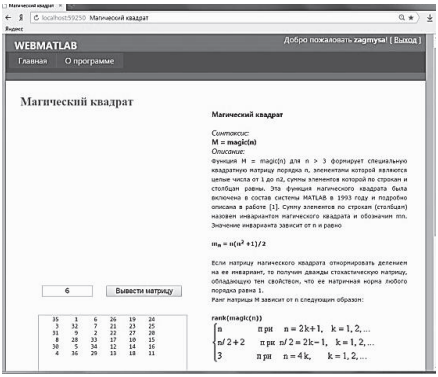


Рис. 2. Страница Web – приложения

Как видим, математические вычисления производятся в среде MATLAB и выводятся на страницу приложения.

Подобным образом можно скомпилировать библиотеку, с помощью которой можно будет, например, находить значения функция, производить действия над матрицами, строить 3D-график заданной функции.

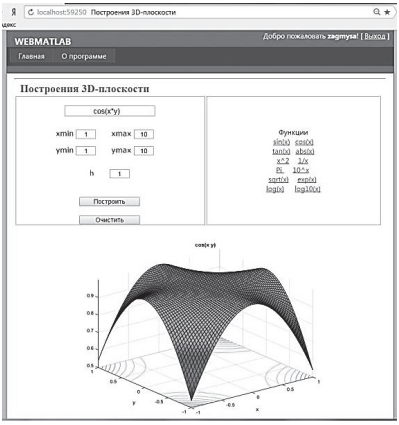


Рис. 3. Страница «Построение 3D-плоскости»

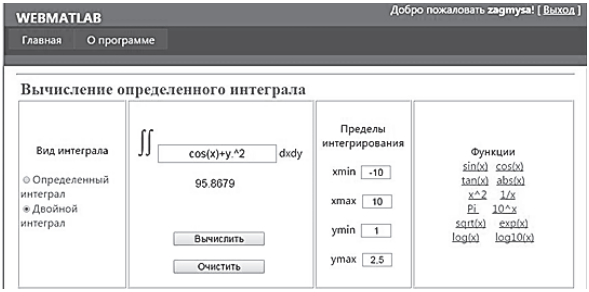


Рис. 4. Страница «Вычисление определенного интеграла»

Вывод. Созданное приложение может использоваться конечным пользователем независимо от MATLAB. MATLAB Compiler дает возможность интегрироваться и с другими языками разработки. Для этого существует ряд дополнительных продуктов (MATLAB Builder JA, MATLAB Builder NE, MATLAB Builder EX), позволяющих упаковать приложения MATLAB в виде программных компонентов для использования в других приложениях.

Литература

1. ASP.NET [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/ASP.NET
2. Видео и вебинар на тему «Развертывание приложений с помощью MATLAB» [Электронный ресурс]. URL: http://www.mathworks.com/videos/application-deployment-with-matlab-81715.html?s_iid=main_custom_MN_cta1
3. МАТЛАБ [Электронный ресурс]. URL: http://matlab.ru/products/matlab
4. Официальный сайт компании «Mathworks» [Электронный ресурс]. URL: http://www.mathworks.com

ОБЛАЧНЫЕ ТЕХНОЛОГИИ КАК ВЕДУЩИЙ ИНСТРУМЕНТ SMART- ОБУЧЕНИЯ

А.Х. Хусаинова
Казанский федеральный университет, Казань
alfirahamzovna@gmail.com

В статье описывается методика создания образовательной среды учебной дисциплины на основе принципов Smart Education – современного метода обучения, базирующегося на облачных технологиях и обеспечивающего интерактивность учебного процесса, свободный доступ ко многим источникам информации, возможность создания максимально комфотных условий для построения индивидуального

образовательного маршрута, способствующего развитию навыков общения, сотрудничества и творческого подхода к решению проблем.

Ключевые слова: *Smart Education, облачные технологии, образовательная среда учебной дисциплины.*

На сегодняшний день мы живем в условиях стремительно меняющегося информационного общества. Важно, в рамках компетентностного образования организовать обучение студентов таким образом, чтобы он на выходе из университета был обладателем профессиональных компетенций, адаптированных к быстро меняющейся информационной среде, актуальных знаний и прикладных навыков.

«Новые ИКТ ставят под сомнение эффективность традиционных автономных систем организации и управления электронным обучением (LMS / электронных учебных оболочек, с которыми работает вуз, приобретенных на коммерческой основе или находящихся в свободном доступе). По мнению специалистов [1,3], будущее электронного обучения – за разработкой специальных сервисов, которые интегрируют LMS с социальными сетями, облачными вычислениями и обеспечивают студентам доступ к обучению с помощью мобильных устройств и мобильных приложений.

Проблемы разработки и приобретения специальных приложений для реализации электронного обучения посредством новых перспективных ИКТ решаются на институциональном уровне. Педагоги же, помимо или вместо использования в учебном процессе автономных LMS, проводят эксперименты по применению бесплатных общедоступных образовательных интернет - инструментов, комбинируя сервисы Web 2.0 и облачных вычислений» [4]. Мы видим смену парадигмы обучения: элементы LMS интегрированы в учебный материал, а не наоборот[6].

Развитие электронного образования можно условно разделить на этапы (рис.1), сроки прохождения которых отличаются в зависимости от национальной системы образования. На смену прежним технологиям электронного обучения пришло «умное» smart – обучение.

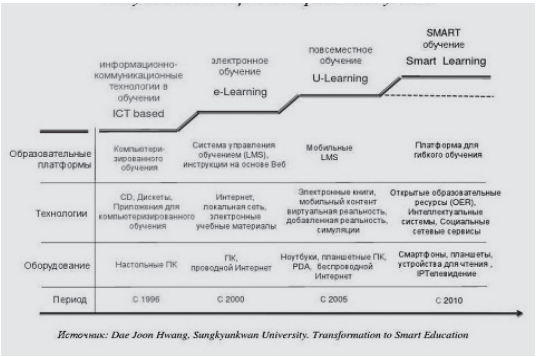


Рис. 1. Этапы развития электронного образования

Smart Education («умное обучение») – это обучение в интерактивной образовательной среде с наличием доступа к источникам информации, находящимся в свободном доступе; легко адаптируемое под потребности каждого студента. Цель использования smart – технологий состоит в том, чтобы обеспечить доступность образования и максимальную индивидуальность траектории обучения для каждого[5].

Роль преподавателя при этом сдвигается в сторону организации и управления учебным процессом. Он теперь не единственный источник информации для студента и нет необходимости писать лекцию под диктовку. Все чаще применяется технология «перевернутого обучения», когда студентам предлагается до занятия ознакомиться с текстом лекции, а в аудитории идет непосредственное обсуждение темы, попытка найти решение каких-то проблем, создание творческих проектов и т.д.

Для реализации данной методики необходим быстрый доступ к интернету и устройство для просмотра информации (компьютер, ноутбук, планшет, смартфон и др.).

Все разнообразие доступных ресурсов должно быть объединено на основе какой – либо платформы, выбор которой зависит от потребностей и предпочтений организатора учебного процесса. В нашем случае это сервисы Google. На выбранной платформе создается образовательная среда учебной дисциплины (рис.2).

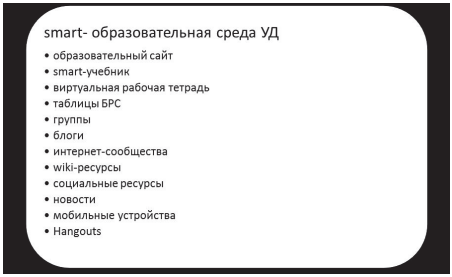


Рис. 2. Структура smart образовательной среды учебной дисциплины

Образовательный сайт дисциплины содержит избыточный учебный контент, ссылки на внешние ресурсы, на документы, созданные участниками образовательного процесса, картинки, видео, анкеты, тесты т.д. Учебник, созданный преподавателем, может быть опубликован в одном из социальных сервисов. Наш учебник размещен в сети Интернет через сервис issuu.com. Данный сервис позволяет студенту конспектировать текст и сохранить выбранные фрагмент, отправляя их через почту, социальные сервисы (G+, например). Smart – учебник, созданный студентом - это конспект учебного материала в виде схем, картинок, ссылок на созданные им в облачных сервисах работы. Студент создает собственный учебный контент, который легко контролируется и оценивается преподавателем, у которого есть возможность просматривать виртуальную рабочую тетрадь студента и комментировать записи в ней. При необходимости студентом вносятся коррективы в работу. В этом случае, цель преподавателя не найти ошибки и снизить оценку, а указать на недостатки и дать возможность исправить их и достичь лучшего результата.

Рис. 3. Smart-учебник студента по изучаемой дисциплине

Данный метод предполагает объединение огромного числа ресурсов, в том числе и социальные сервисы. Существует огромное количество библиотек, в которых сосредоточены книги различной тематики, которые

доступны с любых устройств, имеющих доступ в интернет. Мы имеем возможность формировать собственные коллекции книг, например, «полки» в сервисе Книги Google. Книги с этих полок доступны для просмотра в любое время. На базе Google Диска можно организовать репозиторий, где будут храниться материалы курса, найденные или созданные как преподавателем, так и студентами. Здесь же есть возможность регулировать уровни доступа к документу, от просмотра до редактирования. Источники информации должны быть привлекательны для студента, должны мотивировать к изучению нового. Следовательно, целесообразно находить новые формы подачи учебного контента – это интернет-телевидение учебного назначения (проект «Академия» на tvkultura.ru, univertv,..), дистанционные образовательные ресурсы типа MOOC (intuit.ru, coursera, …), учебное видео на YouTubeEDU, и другие. Студент должен иметь доступ к избыточному количеству источников информации разного жанра, это позволит ему выбрать для себя наиболее доступный и привлекательный контент.

Кроме учебного контента важное место в smart-образовательной среде уделяется практико-ориентированным заданиям. В начале обучения по дисциплине проводится анкетирование на определение уровня ИКТ – компетентности и в зависимости от него студенту предлагаются задания разного уровня сложности. Задания, чаще всего, носят характер проекта, предусматривается разный уровень выполнения заданий. Студент вправе выбрать тот, что ему по силам. Однако, необходимо мотивировать его на выбор более сложного уровня через бальную систему оценок, через соревновательный момент или через совместную деятельность вместе с преподавателем и другими студентами. Для создания условий для общения, обмена информацией, консультаций применяются сервисы электронной почты, чаты в документах Google, позволяющие обмен сообщениями при совместной работе над документом. Hangouts, встроенный в учебный сайт, дает возможность онлайн-консультации при изучении нового материала или при выполнении задания. Также пространством для общения служит Группа Google.

Контроль и самоконтроль результатов обучения легко организуется через Таблицы Google и таблицы результатов Форм Google, доступ к которым студент получает в любое время с любого устройства. Важно дать студенту возможность презентовать результаты своей работы. Одним из вариантов могут быть специально созданные коллекции работ студентов на базе сайтов Google или через облачные Доски заметок. Так, зачетные работы (веб квесты) студентов ИФМК КФУ по курсу «ИКТ в КПД» размещены на облачном ресурсе Padlet [2].

Для успешной реализации метода smart- образования важно преодолеть психологический барьер, чтобы студенты привыкли «иметь дело» с облачными сервисами, использовали их как удобный и полезный инструмент, средство обучения. Тогда не возникнет противоречия с традиционными методами обучения. Облачные сервисы позволяют организовать образовательную среду учебной дисциплины с учетом особенностей преподаваемой дисциплины и с учетом уровня подготовки студентов, что в конечно итоге, позволяет повысить эффективность учебного процесса. За период использования данного метода отмечено ускорение формирования у студентов практических навыков в решении реальных профессиональных задач.

Литература

1. Воронкин А.С. Социальные сети: эволюция, структура, анализ [Электронный ресурс] // Образовательные технологии и общество. 2014. № 1. С. 650–675. URL: http://ifets.ieee.org/russian/depository/v17_i1/pdf/21.pdf (дата обращения: 07 04 2016).

2. ИКТ в ОКПД. Веб квесты к зачету. (http://padlet.com/alfirahamzovna/ICT Дата обращения: 07 04 2016)

3. Соловов А.В. «Золотые клетки» виртуальных учебных сред [Электронный ресурс] // Высшее образование в России. 2012. № 11. С. 133–137. URL: http://cyberleninka.ru/ article/n/zolotye-kletki-virtualnyh-uchebnyh-sred (дата обращения: 07 04 2016).

4. Фомина А.С. Учебное проектирование с применением Google Диск (Drive) в высшем учебном заведении // Теория и практика общественного развития . 2015. №11. С.281-289.

5. Smart обучение Өрлеу (https://issuu.com/zkoipk/docs/smart-_____ Дата обращения: 07 04 2016)

6. Smart-учебники в smart-образовании. Новая парадигма контента. (http://www.slideshare.net/pnevostruev/smart-congress Дата обращения: 07 04 2016)

УДК 13.00.02

МУЛЬТИМОДАЛЬНОСТЬ В ОБРАЗОВАНИИ

А.Ф. Хусаинов, Д.Д. Якубова
Казанский федеральный университет,
НИИ «Прикладная семиотика» АН РТ, Казань
khusainov.aidar@gmail.com, suleymanovad@gmail.com
Ж.Е. Вавилова
Казанский государственный энергетический университет, Казань
zhannavavilova@mail.ru
А.В. Паркалов
Белорусский государственный университет информатики и радиоэлектроники, Минск
a.parkalov@gmail.com

В статье описывается реализация мультимодального подхода в образовании. Авторы дают определение понятию мультимодальной грамотности, приводят особенности мультимодального текста и характеризуют обучающую программную систему, адаптирующуюся под возможности обучаемого.

Ключевые слова: *мультимодальность, мультимодальная грамотность, мультимодальный текст, компьютерная обучающая система*

XXI век предъявляет новые требования к образованным личностям. Традиционное понимание грамотности как умения читать и писать уступила новому, учитывающему широкий спектр социальных, экономических и технологических факторов. Одним из нововведений системы образования стала интеграция принципа мультимодальности в образовательный процесс, что привело к изменению ряда традиционных аспектов образовательной системы.

Под мультимодальной грамотностью мы вслед за Г. Крессом и К. Джувитт понимаем разнообразие способов представления знания и создания смысла, где модальность рассматривается как результат социально-исторического оформления средств, выбираемых обществом для репрезентации, или социально обусловленный семиотический ресурс смыслообразования [1]. При этом модальности постоянно трансформируются пользователями в зависимости от коммуникативных нужд общества или отдельных социальных групп; периодически создаются новые модальности.

Несмотря на то что мультимодальность лишь недавно стала частью образовательного процесса, вся практика коммуникации, грамотности и создания произведений (текстов, картин, музыки, фильмов и др.) всегда была мультимодальной [2]. К тому же мультимодальность свойственна самой человеческой природе: так, способами самовыражения маленьких детей являются речь, рисование, жесты, знаки и пр. Как правило, учащиеся знакомятся с мультимодальными технологиями не в стенах школы, а в неформальных условиях, однако формирование в них мультимодальной грамотности является задачей института образования.

Мультимодальный текст

Традиционно основной модальностью в обучении являлся текст, однако новый контекст предполагает смещение письменного текста с центральной позиции. Особенно это касается учебных книг для дошкольников и младших школьников, где важную роль играют, с одной стороны, картинки и графические элементы, а с другой – взаимодействие с учителем и одноклассниками вокруг текста в процессе прочтения его смысла.

Следует отметить, что в наше время само понятие текста расширилось и включает в себя разные виды представления знаний. Компьютерная верстка текста позволяет создавать смысл самыми разными способами, к примеру, при помощи шрифта и его размера, цвета, использования изображений, которые дополняют или заменяют печатный текст. В результате мультимодальные тексты имеют разные способы прочтения. Способ презентации знания является важным элементом конструирования знания и определяет, что должно быть выучено и как это должно быть выучено [3]. Уже на уровне начальной школы учителям следует обучать детей «читать» такие тексты, а значит декодировать текстовую информацию в сочетании с прочими культурно и социально обусловленными способами выражения.

Еще одной важной характеристикой цифровой эры является смещение центральной роли от книги к экрану. Экранные тексты – это сложные мультимодальные сочетания изображения, звука, анимации и других форм представления и коммуникации [4]. В них грань между собственно текстом и изображением зачастую стирается; более того, не имеет смысла их разделение, так как в соответствии с мультимодальным подходом сами пространственные отношения между текстом и изображением являются смыслообразующим ресурсом [4].

Естественно, что подход к работе с такими «текстами», включая конфигурацию мультимодальных данных при их создании и формулировку заданий, отличается от подхода к работе над традиционными текстами, и переход является энерго- и времязатратным. Дополнительная

сложность заключается в необходимости подготовки и переподготовки учителей, многие из которых чувствуют себя неуютно в эру цифровых технологий и испытывают сложности в связи с необходимостью интеграции новых принципов в свою каждодневную работу. Однако это является приоритетной задачей в эру информационных технологий и должно рассматриваться как часть непрерывного образования учителей и преподавателей.

Мультимодальные технологии

В последние десятилетия активно создаются компьютерные обучающие программы, в которых знания представлены в различных модальностях. При этом открытым остается вопрос о том, должен ли каждый квант знания быть представлен в разных модальностях, и если да, то в каких именно.

Если все модальности в коммуникативной ситуации способствуют созданию смысла, логично предположить, что модальности взаимодополняемы, а не взаимозаменяемы; следовательно, для полного раскрытия смысла необходимо использование нескольких модальностей. Однако это не означает, что чем больше модальностей используется, тем эффективнее процесс обучения. Ведущими принципами при выборе модальности для каждого случая являются принципы достаточности и избыточности. То есть, с одной стороны, процесс обучения проходит более эффективно, если задействованы несколько каналов восприятия учащегося, а с другой – одновременное использование многих модальностей может отвлекать, приводить к когнитивной перегрузке или потере времени вследствие повторной обработки аналогичных знаний.

Для достижения оптимального результата при создании компьютерных обучающих систем целесообразно предусмотреть возможность адаптации контента учебного курса под особенности восприятия каждого конкретного обучаемого, как это делает разрабатываемая авторами система генерации мультимодальных обучающих материалов [5]. С этой целью в начале курса программа путем тестирования определяет доминирующий канал восприятия студента и на основе результатов формирует подходящий ему способ отображения обучающего материала. Преподавателю необязательно вводить один и тот же факт в разных формах, так как система осуществляет автогенерацию модальностей, но ему необходимо указать значимость данного кванта знания в настройках учебного курса. Так, особо значимый материал «визуалу» будет представлен одновременно в текстовой форме и в виде изображения, а менее значимый – только в виде текста.

Таким образом, вышеописанная программа реализует мультимодальный подход, который соответствует сложившейся в последние десяти-

летия образовательной тенденции учитывать индивидуальные характеристики учащихся, в том числе их нейрофизиологические особенности. Данный подход позволяет приблизить технологии к человеку, оптимизировать процесс обучения и, соответственно, помочь учащимся реализовать свой потенциал.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проект № 15-57-04085 «Модели и средства мультимодального синтеза текстов для интеллектуальных обучающих систем»).

Литература

1. Jewitt C., Kress G. (Eds.). Multimodal literacy. New York: Peter Lang, 2003.
2. Kress G. Multimodality: A Social Semiotic Approach to Contemporary Communication. New York: Routledge, 2010.
3. Hassett, D.D., Curwood, J.S. Theories and Practices of Multimodal Education: The Instructional Dynamics of Picture Books and Primary Classrooms // The Reading Teacher. 2009. Vol. 63 (4). Pp. 270-282.
4. Jewitt, C. Multimodality, “Reading”, and “Writing” for the 21st Century // Discourse: Studies in the Cultural Politics of Education. 2005. 26(3). Pp. 315-331.
5. Вавилова Ж.Е., Русецкий К.В., Хусаинов А.Ф., Якубова Д.Д. Прототип адаптивной системы генерации мультимодальных обучающих материалов // Казанская наука. 2015. №11. С. 246-249.

УДК 37.378.147.88

ЭЛЕКТРОННО-ОБРАЗОВАТЕЛЬНЫЕ РЕСУРСЫ В НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ И ПРОЕКТНОЙ ДЕЯТЕЛЬНОСТИ СТУДЕНТОВ

И.Э. Ярмакеев
Казанский (Приволжский) федеральный университет
ermakeev@mail.ru
А.Р. Абдрафикова
Казанский (Приволжский) федеральный университет
albina112000@mail.ru
Т.С. Пименова
Казанский (Приволжский) федеральный университет
pimenova_ts@mail.ru

В статье рассматриваются формы, способы организации и этапы проведения научно-исследовательских и проектных работ студентов с использованием электронно-образовательных ресурсов (Е-ресурсов). Определены общие и специфические требования к содержанию

и оформлению научных исследований, структура научно-исследовательской работы и ее основные элементы, а также обозначены критерии их оценивания. Разработанный авторами Е-ресурс «Научно-исследова-тельская и проектная деятельность студентов» включает образцы научной статьи, научных тезисов, научного доклада и научного проекта с полным описанием, а также ориентирует студентов, как определить актуальность темы исследования, поставить цель, провести различие между предметом и объектом исследования; определить гипотезу, объяснить теоретическое и практическое значение, выбрать современные методы исследования. В Е-ресурсе представлены основные направления исследований в области теории обучения иностранным языкам; предложены методические рекомендации для преподавателей и студентов и списки литературы. В статье представлены результаты эксперимента, которые позволяют авторам констатировать, что применение электронно-образовательных ресурсов способствует повышению научной культуры студентов при проведении научно-исследовательских и проектных работ. Е-ресурс «Научно-исследовательская и проектная деятельность студентов» был создан для студентов-бакалавров.

Ключевые слова: *научно-исследовательские, деятельность по проекту, научная статья, критерии оценки, теоретическая и практическая значимость, научная риторика.*

INTRODUCTION

A constant search for towering teaching tools in scientific research and project activities of students has been undertaken by numerous educators (Burgin & Kuznetsov, 1994; Zimniya, 2001; Bidenko & Van Santvoort, 2003; Halskova, 2004; Mikeshina, 2005; Kodzhaspirova & Petrov, 2006; Mikheeva, 2010; Krasilnikova, 2012; Rusavin, 2012; Pasechnaya, Skomorokhova & Yurtaev, 2013; Trainev & Teplyshev, 2013). None attempts to implement E educational resources in scientific research and project activities of students have been made though. In this paper, the authors claim that E-educational resources can be effective for conducting scientific research and project activities of students. The E-educational resource “Scientific research and project activities of students” was created for senior students obtaining a bachelor’s degree.

The following study materials include current scientific topics:

- scientific research and project activity;
- a scientific article and its main elements;
- a scientific project, types of scientific projects;
- a scientific thesis;
- a scientific experiment;

- a student scientific paper;
- basic rules of scientific rhetoric;
- assessment criteria of research and project work.

The efficiency of students' scientific research and project activities was supported by the methodical guidance worked out by the authors.

This E-educational resource provided high-leveled scientific research and project activities of students.

MATERIALS AND METHODS

Students' Background

Number, age and gender characteristics: 49 students, 20-22 years old, mixed.

Students' status: bachelor students.

The languages spoken: most students are from two cultural and linguistic backgrounds: Tatar and Russian.

Research Sites

Leo Tolstoy Institute of Philology and Intercultural Communication, Kazan Federal University.

Study instruments

The following study instruments were suggested to implement the E-educational resource "Scientific research and project activities of students":

- Discussion
- Forum
- Power Point presentation
- File document
- Chat
- Public speech
- Essay
- Video-conferences.

RESULTS

The evaluation table presents the progress students made while acquiring general and specific requirements for the content and design of scientific research.

Date	Research problem formu-lation	Pithiness of the submit-ted materials	Theoretical and practical signify-cance of the study	Findings' validity	Innovative research forms and methods	Presen-tability	Issues' disclosure	Use of references
Sept. 2015	0	0	0	0	0	12	0	0
Oct. 2015	12	12	10	9	15	30	14	21
Nov. 2015	28	45	27	19	25	45	36	40
Dec. 2015	40	49	39	40	34	49	40	49

Table 1. Students' progress evaluation table

The challenging issues were as follows: to choose innovative research forms and methods; to define theoretical and practical significance of the study; to formulate the research problem; to prove the findings' validity; to find a creative approach to issues' disclosure.

DISCUSSION

Students perform scientific research and project activities while:

- writing term and final papers with the elements of scientific research;
- reading for practical and seminar classes;
- performing specific tasks with the elements of scientific research during compulsory school teaching practices;
- planning and conducting a scientific experiment and processing data;
- participating in students scientific and educational conferences.

The main types of research and project activities

The following main types of research and project activities of students are suggested:

• writing an essay;

• compilation of an annotated list of articles from relevant journals in the proposed field of knowledge (pedagogical, psychological, methodological, etc.);
• preparation of an article and/ or a textbook review;
• performing micro scientific analysis;
• writing a scientific thesis/ a short article;
• making a list of references.

Organization of research and project activities of students

The following subjective factors of the intensification of research and project activities of students caused by the psychological specifics of students should be taken into account:

1. Knowledge of the program material.
2. The availability of a strong system of knowledge required for mastering basic university courses.
3. The availability of skills, experience of mental work.
4. The specificity of cognitive mental processes of the student: attention, memory, language, observation, intelligence and thinking.
5. Good performance, which is provided by a normal physical condition.
6. Compliance of the chosen activity, profession with the individual abilities.
7. The ability to self-regulate your emotional state and eliminate the circumstances that violate the business spirit, preventing the intended target.
8. Mastering the best work style ensuring the success in activities.
9. The level of self-demand, determined by the current self-esteem.
10. The adequate assessment of knowledge, strengths, weaknesses that is an important component of the human's self-organization, which can not be successful without managerial work on your behavior and activity.

Stages of research and project activities

We can distinguish the following stages of research and project activities of students:

The first stage - preparatory. It should include the drafting of a working program with the allocation of topics and assignments; intermediate planning for the semester; preparation of teaching materials; diagnostics of students' efficiency.

The second stage - organizational. The main objectives of individual and group work of students are determined at this stage as well as the introductory lecture is read and the individual-group set consultation is held during which

the forms of research and project activities of students and assessment criteria are discussed; new terms and forms of presentation of interim results are set.

The third stage - motivational. The supervisor at this stage should provide a positive motivation of individual and group activities; verify intermediate results; organize self-control and self-correction, interchange and mutual test in accordance with the set goal.

The fourth stage - assessing. It includes individual and group reports' evaluation. The results can be presented as a final project, a term paper, an essay, a report, diagrams, tables, oral presentations, models, reports, etc.

Assessment criteria of research and project activities

1. The level of research problem formulation.
2. Relevance and originality of the theme.
3. The logical sequence of presentation.
4. Pithiness of the submitted material.
5. Theoretical and practical significance of the study.
6. The validity of the findings.
7. Summary of the innovative forms and methods of research.
8. Culture of speech (speech literacy).
9. Presentability.
10. A creative approach to disclosure issues.
11. The correctness of the use of references.

CONCLUSION

It is common knowledge that the process of introduction of research and project activities of students in the learning process should be staged and justified. The effective implementation of this type of activities needs effective tools. The authors advocate the idea that E-educational resources can meet the requirements.

The implementation of the E-educational resource "Scientific research and project activities of students" was held with the aim in view to conduct and write research papers on theoretical and practical issues in teaching foreign languages. Students' works were assessed in accordance with the above mentioned criteria. The high-leveled results obtained by students proved the validity and reliability of the E-educational resource "Scientific research and project activities of students." This E-educational resource is available in open access at http://do.kpfu.ru/course/view.php?id=1684.

References

1. Bidenko,V. I., Van Santvoort J. (2003). Modernization of Vocational Education: Modern Stage. 2-nd ed. Moscow: Research Center for Specialists Training.

2. Burgin, M. S., Kuznetsov, V. I. (1994). Introduction to Modern Methodology of Science: Structures of Knowledge. Manual for Students. Moscow: Aspect Press.

3. Halskova, N. D. (2004). Modern Methods of Teaching Foreign Languages. Manual for Teachers. 3-d. ed. Moscow: ARKTI.

4. Kodzhaspirova, G. M., Petrov, K. V. (2006). Training Tools and Methods. Manual for Students of Higher Educational Institutions. 3-d ed. Moscow: Academia.

5. Krasilnikova, V. A. (2012). The Use of Information and Communication Technologies in Education. Tutorial. 2-nd ed. Orenburg: Orenburg State University. Retrieved from http://www.bibliorossica.com/book.html?currBookId=7901

6. Mikheeva, N. F. (2010). Methods of Teaching Foreign Languages. Training Manual. Moscow: Russian University of Friendship of Peoples. Retrieved from http: //www.bibliorossica.com/book.html?currBookId=10371

7. Mikeshina, L. A. (2005). Philosophy of Science: General Problems of Cognition. The Methodology of the Natural Sciences and Humanities. Reader. M.: FLINTA. Retrieved from http://www.gumer.info/bogoslov_Buks/Philos/mik_filn/index.php

8. Pasechnaya, I. N., Skomorokhova, S. V., Yurtaev, S. V. (2013). Culture of Speech. Manual. M.: FLINTA. Retrieved from http://znanium.com/bookread.php?book=466248

9. Rusavin, G. I. (2012). The Methodology of Scientific Knowledge. Training Manual. Moscow: Unity-Dana. Retrieved from http://www.biblioclub.ru/book/115020/

10. Trainev, V. A., Teplyshev, V. Yu. (2013). New Information and Communication Technologies in Education. 2-nd ed. Moscow: Publishing and Trading Corporation "Dashkov and K". Retrieved from http://znanium.com/bookread.php?book=430429

11. Zimniya, I. A. (2001). Lingo-Psychology of Speech Activity. Moscow-Voronezh.

ENGLISH ABSTRACTS. SECTION 1

AUTOMATED PROCESSING OF ARCHIVAL COLLECTIONS OF THE SCIENTIFIC JOURNAL "DIGITAL LIBRARY"

D.Y. Akhmetov
Kazan Federal University
akhmetov.dy@gmail.com

Developed and tested methods of automatically extracting metadata from the archival collection of the electronic scientific journal "Digital Libraries" are presented. Software package selection and processing of metadata magazine articles, implemented in PHP using CURL technology, html dom developed.

Keywords: *automatic extraction of metadata, Semantic Web, xml-format*

APPLICATION DEVELOPMENT SUPPORT OF MANAGEMENT DECISION MAKING

A. Bakunina, K. Tsybenko
Kazan Federal University, Kazan
bakuninaa@gmail.com, ktsybenko@mail.ru

The article describes how the application development support of management decision making by means of Visual Basic for Application and Microsoft Office Excel, based on the Saati method.

Keywords*: Visual Basic for Application, Microsoft Office Excel, Saati method.*

MODEL OF WORD EMBEDDING ON THE TECHNOLOGY WORD2VEC

F.M. Gafarov, E.I. Shaydullina, V.R. Gafarova
Kazan Federal University, Kazan
fgafarov@yandex.ru, lelechka_29@mail.ru

This paper covers the definition of word embedding – word2vec, analyzes its advantages and disadvantages. Studied and compared the possibilities of the technology in Tatar, English and Russian languages.

SERVICE INTEGRATION NEWS FEEDS ON THE PLATFORM OF CONTROL ELECTRONIC SCIENTIFIC JOURNALS

A.N. Gerasimov
Kazan Federal University
gerasimov.mailstore@gmail.com

On the basis of recommendation systems method of analyzing the content of news feeds scientific portals and forming personalized news set offered. Algorithms harmonize the various formats of news feeds and news filtering through a personal user profile implemented. Consolidation algorithm scientific news content is implemented as a separate module OJS platform.

Keywords: *electronic scientific journal, integration of electronic resources, news feeds*

SEMANTIC ANALYSIS OF LARGE SCIENTIFIC DOCUMENTS COLLECTIONS

A.M. Elizarov[1], E.K. Lipachev[2], S.M. Khaydarov[3]
Kazan (Volga region) Federal University
1 – amelizarov@gmail.com, 2 – elipachev@gmail.com,
3 – 15jkeee@gmail.com

The method of automated processing of large collections of physical and mathematical documents stored in OpenXML format proposed. The method includes the validation of documents; transform them according to the rules of formation of collections, semantic analysis of documents, metadata extraction, and others. Method algorithm is described. An example of successful implementation of the method in the organization of the XI All-Russian Congress on fundamental problems of theoretical and applied mechanics (Kazan, 20–24 August 2015) is shown.

Keywords: *Big Data, semantic analysis of documents, structural analysis of texts, metadata, services, automatic processing of large collections.*

ONTOMATH ECOSYSTEM AND WORLD DIGITAL MATHEMATICAL LIBRARY

A.M. Elizarov[1], N.G. Zhiltsov[2], A.V. Kirillovich[3], E.K. Lipachev[4], O.A.Nevzorova[5]
Kazan Federal University, Kazan
1 – amelizarov@gmail.com, 2 – nikita.zhiltsov@gmail.com,
3 – alik.kirillovich@gmail.com,
4 – elipachev@gmail.com, 5 – onevzoro@gmail.com

Possibilities of use in conducting new studies of all the accumulated body of scientific knowledge are described. Such use requires the widespread introduction of information and communication technologies (ICT) to ensure optimal management of existing knowledge, the organization of effective access to, and sharing and re-use of new types of knowledge structures. The greatest effect of the introduction of modern ICT to further organize the scientific knowledge and enhance their clarity can be expected in the field of mathematics. These expectations were fully confirmed by the project of creation of the World Digital Math Library (WDML). The main directions of the implementation and results of the project WDML to create OntoMath ecosystem as its component represented.

Keywords: *World Digital Library of Mathematical (WDML), ecosystem OntoMath, ontologies, semantic search.*

CREATION OF THE SUBJECT DOMAIN AND CO-AUTHORSHIP NETWORK IN THE FIELD OF LAW ON THE BASIS OF SENSING OF THE GOOGLE SCHOLAR CITATIONS SERVICE

D.V. Lande, V.A. Andrushchenko
Institute for Information Recording of National Academy of Sciences of Ukraine, Kiev
dwlande@gmail.com, valentyna.andrushchenko@gmail.com

The algorithms of creation of the subject domain and co-authorship network in the field of the law regulated by their scientific interests is given in work. The subject domain and network of a co-authorship is formed on the basis of sounding of the Google Scholar Citations service. Subject domain and networks of a co-authorship can be considered as a basis for identification of the concepts and schools of sciences.

Keywords: *subject domain, co-authorship network, legal science, sensing of a network, information network*

USE OF DATA BINDING SYSTEMS FOR FINDING MATCHES BETWEEN BIBLIOGRAPHIC DATA REPOSITORIES

K.S. Nikolaev, O.A. Nevzorova
Kazan Federal University, Kazan
konnikolaeff@yandex.ru

This article describes the technology for presenting data in RDF format, the introduction of the concept of Linked Open Data and software solutions for data binding and experiments with bibliographic data repository.

Keywords: *data binding, RDF, Linked Open Data.*

DEVELOPMENT OF AN ONTOLOGICAL MODEL OF THE OPERATIONS OF COMMERCIAL BANKS

S.A. Pozdeeva
The Moscow State Pedagogical University, Moscow
englishinoz@gmail.com

The banking sector permeates all sectors of the economy, ensuring their monetary stream. We can say that the banks of the country - an indicator of the state of the economy, so it is necessary to structure the knowledge about them. This can be done using ontology.

Ключевые слова: *ontology, bank operation*

CONCEPTUALLY-FIGURATIVE CONTROLLING THE STATEMENTS OF TASKS

P. Sosnin, M. Galochkin, A. Luneckas
Ulyanovsk state technical university, Ulyanovsk
sosnin@ulstu.ru, m.galochkin@ulstu.ru, lunacorp@inbox.ru

The paper provides a means of pseudo code programmable graphics that support designing of automated systems at the conceptual stage. Features of means determine the approach to convert text units to prolog-like structure and semantic graph-scheme in terms of an interaction the designer with the ontology of the project. The ability to reverse the conversion scheme after its correction allows iteratively to bring it and the investigated textual unit till mutually agreed conditions, the checked version of the designer understanding

of the indicated text. The process of such investigation is implemented in the toolkit OwnWIQA.

Keywords: *ontology, understanding, task statement, predicate, semantic graphics*

NATURAL LANGUAGE SEMANTICS MODELLING BASED ON THE MULTIAGENT RESURSIVE COGNITIVE ARCHITECTURE

B.P. Tazhev, D.G. Makoeva, I.A. Pshenokova
Institute of Computer Science and Problems of Regional Management of KBSC of the Russian Academy of Sciences
360000, KBR, Nalchik, 37-a, I.Armand street
E-mail: boristazhevart@mail.ru, d.makoeva@iipru.ru

This article provides a semantic formalization model of the natural language based on the multi-agent recursive cognitive architecture.

Keywords*: multi-agent systems, cognitive architecture, formal semantics, natural language interface.*

ENGLISH ABSTRACTS. SECTION 2

STATISTICAL METHODS IN RESEARCH OF COMPETING VERBS FORMS IN RUSSIAN

T.I. Galeev
Kazan Federal University, Kazan
TIGaleev@kpfu.ru

The article discusses the competition of grammatical synonyms – verbs derived with suffixes –а/я–, and – ива/ыва – (*ozdorovlyat'*/*ozdoravlivat'*). Based on the data obtained using the Google Books, it was possible to describe the basic patterns of change of frequency dynamics of competing forms.

Keywords: *variability, Google Books, language dynamics, verb paradigm, cognitive linguistics.*

FUNCTIONS OF THE SIMILATIVE AFFIX *-DAY* IN TATAR (ON CORPUS DATA)

A.M. Galieva
Research Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan
amgalieva@gmail.com

The paper deals with Tatar word forms, containing affix of the similative -DAy; and the structural and semantic peculiarities of these forms of hybrid nature are examined. The paper is composed of two parts: the first we study affixal chains typical for -DAy forms, then we distinguish syntactic features of -DAy contractions.

Keywords: *affix, simulative, the Tatar language, assimilation meaning.*

GRAMMATICAL PORTRAIT OF THE TATAR TEXT AND ITS STYLISTIC FEATURES

A.M. Galieva, R.R. Gataullin
Research Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan
amgalieva@gmail.com, ramil.gata@gmail.com

This paper studies the issue of correlation of POS characteristics of the text (correlation of grammatical parameters of words within the text) and the style of the text. Preliminary results of development of methodology for automatic classification of Tatar texts from corpus collection according their morphological features are presented.

Keywords: *style, the Tatar language, stylistic characteristics of the text.*

COGNITIVE STRUCTURE OF NEGATION IN TATAR

A.M. Galieva, D.Sh. Suleymanov
Research Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan
amgalieva@gmail.com, dvdt.slt@gmail.com

The paper deals with basic types and means of expressing negation in the Tatar language; a cognitive, structural and semantic complexity of negation in Tatar as linguistic category is shown. The authors distinguish four main ways

of negation expressing, and these ways differently classify and conceptualise situations that fall under negation. None of these means of expressing negation in Tatar is semantically elementary.

Keywords: *negation, the Tatar language, semantics, linguistic categories.*

ABOUT MORPHOLOGICAL TAGGING OF TATAR PARTICIPLE

A.M. Galieva
Research Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan
amgalieva@gmail.com
A.R. Gatiatullin
Research Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan
agat1972@mail.ru

THE ROLE OF ADJECTIVES IN THE SENTIMENT ANALISYS

Banu Yergesh, Altynbek Sharipbay, Gulmira Bekmanova
L.N.Gumilyov Eurasian National University, Astana, Kazakhstan
b.yergesh@gmail.com, sharalt@mail.ru, gulmira-r@yandex.ru

Here we described the linguistic approach for sentiment analysis of texts in the Kazakh language.This approach based on morphological rules.

Keywords: *sentiment, sentiment analysis, Kazakh language, classification, morphological rules.*

THE PHENOMENON OF MEANING IN COGNITIVE LINGUISTICS

G.V. Kolpakova
Kazan Federal University, Kazan
Galina.Kolpakova@kpfu.ru

The article is concerned with the analysis of approaches to the research of meaning in structural and cognitive linguistics. The article deals with problems of structuralizing of a mental vocabulary based on the concept of semantics of modern German researches.

Key words: meaning, semantics, mental lexicon, structural linguistics, cognitive linguistics.

ON THE IMPLEMENTATION OF THE MORPHOLOGICAL TAGGING SYSTEM OF THE CRIMEAN TATAR ELECTRONIC TEXT CORPUS

L.Sh. Kubedinova
Crimean Federal University, Simferopol
kubedinova@gmail.com
A.R. Gatiatullin
Research Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan
agat1972@mail.ru

Nowadays many text electronic corpuses are created for many languages of Turkic group. Such corpuses already exist for Turkish, Tatar, Kazakh, Bashkir, Tuvinian and other Turkic languages. All authors of these corpuses are faced the same problems and start heading the same way of creating their own systems of corpus annotation. Although structure similarity of Turkic languages allows to create a common base of computer and program models for processing texts in Turkic languages.

In this article the system of morphological tagging for the Crimean Tatar electronic corpus and the program of morphological analyses of Crimean Tatar wordforms are considered. This system is developed on the basis of tags which are used for annotation of electronic corpus of Tatar language "Туган тел" ("Mother tongue").

Key words: *morphological annotation, Crimean Tatar electronic corpus.*

ON CERTAIN CHALLENGES OF MORPHOLOGICAL ANNOTATION OF KAZAKH TEXTS

Makazhanov Aibek Omirzhanovich, Sultangazina Aitolkyn Nurlanovna
National Laboratory Astana, Astana, Kazakhstan
aibek.makazhanov@nu.edu.kz, aitolkyn.sultangazina@nu.edu.kz

Morphological annotation with ambiguity resolution is a process of assigning each token (annotation unit) a *single* appropriate morphological parse (a triple consisting of <lemma, part of speech tag, a set of grammemes>) in accordance with a predefined annotation scheme. Tokenization criteria constitute an inseparable part of an annotation scheme, by defining what and when to consider a separate token. Usually one token corresponds to one orthographic word and vice versa, however sometimes this identity can be broken. This paper describes two of such cases present in some Turkic languages, namely: 1) one orthographic word consists of more than one token; 2) one to-

ken consists of more than one orthographic word. We also provide examples of implementation of those tokenization criteria in one of the existing schemes of morpho-syntactic annotation of Kazakh.

Keywords: *morphological annotation, tokenization, Turkic languages, Kazakh language*

USING CORPUS OF TEXTS IN FOREIGN LANGUAGE TEACHING

A.F. Mukhamadyarova
Kazan Federal University, Kazan
liliana_muhamad@mail.ru
B.E. Khakimov
Kazan Federal University, Kazan
"Applied Semiotics" Research Institute of the Tatarstan Academy of Sciences
khakeem@yandex.ru

The article discusses general methodological aspects of using text corpora in foreign language teaching. The main advantages of the corpus-based approach in teaching foreign language to the philology students are defined.

Keywords: *corpus of texts, foreign language, language teaching*

LINGUISTIC ANNOTATION OF THE VERBS VOICE FORMS IN THE SAKHA LANGUAGE

A.N. Nogovitsyna
M.K. Ammosov North-Eastern Federal University, Yakutsk
erkin2007@mail.ru

The article discusses linguistic annotation of the verbs voice forms in the Sakha language. The author offers to annotation voice forms of the verb the following tags: active voice - ACT (Active), causative voice - CAUS (Causative), reflexive voice - REFL (Reflexive), passive - PASS (Passive), together-mutual voice - RECIPR (Reciprocal).

Keywords: *corpus linguistics, the Yakut language, voice forms of the verb, linguistic annotation.*

SEMANTIC PARALLELS IN ALTAIC LANGUAGES

D.B. Ramazanova
IYALI im.G.Ibragimova AN RT
iyali.anrt@mail.ru

The article discusses a parallel semantic model in Tatar (and Turkic), Tungus-Manchurian, and Mongolian languages. The primary lexical base of analysis are thematic groups: somatism, names of clothing, etc. The role of principles of the nomination, and the metaphoric structure of lexemes and development of their semantics is defined.

Key words: *development of semantics, semantic Parallels, the family structure of words, semantic model, the names of the fingers.*

PRINTED AND ELECTRONIC BOOKS ON THE TATAR LANGUAGE

R.K. Sagdieva
Kazan Federal University, Kazan
ramsag777@rambler.ru

The article describes the novelty of new print and electronic textbooks of the Tatar language, methods of working with these benefits, the effective use of new information technologies.

Keywords: *Tatar, electronic textbooks, assignments, learning process.*

MARKING-UP MORPHOLOGICAL CATEGORIES IN NATIONAL CORPUS OF KYRGYZ TEXTS

T. Sadykov
Bishkek Humanities University named after K.Karasaev, Bishkek
tash_sadykov@mail.ru
B. Sharshembayev
"Manas" Kyrgyz-Turkish University, Bishkek
bakyt101@mail.ru

The identification of parts of speech and their distribution in the texts is one of the most complicated problems of Turkology. In contrast of inflexional languages in Turkic languages words and word forms undergo phenomenon of conversion and homonymy. For a marking of morphological categories, and

also for the removal of a text homonymy, the system of a morphological tagging with the harmonized compatible systems for the national corpus of Turkish texts.

BUILDING CORPUS WITH NAMED ENTITIES RECOGNITION IN THE NEWS FOR KAZAKH LANGUAGE

Z.N. Sadykova, V.V. Ivanov
Kazan Federal University, Kazan
Sadykovazn@gmail.com

The article presents the process of building corpus with named entities annotation in the news for Kazakh language. The paper considers basic features of named entities recognition. Also the article describes rules for annotation using online brat rapid annotation tool. At the moment we have named entities recognition in collection of Kazakh news that provided by the staff of Nazarbaev University. However the job of building corpus is not over.

Keywords: corpus, annotation, named entities, named entity recognition

THE PHRASEOLOGICAL UNITS OF TATAR: OPENING HISTORY OF TATAR

F.R. Sibgaeva
Kazan Federal University, Kazan
FiruzaRS@mail.ru

Phraseological units reflect in their semantics a long process of development of culture, they accumulate, fix and transmit from generation to generation cultural attitudes, stereotypes, standards, archetypes and knowledge of the history of the people. In article the semantic analysis of Tatar phraseological units containing cultural information is done.

Keywords: *phraseological units, Tatar language, semantic analysis.*

LEXICAL APPROACH TO DYNAMICS OF THE WELL-BEING LEVEL IN SOCIETY

V.D. Solovyev, V.V. Bochkarev
Kazan federal university, Kazan
maki.solovyev@mail.ru, vbochkarev@mail.ru

The article provides a brief overview of the work devoted to the study of well-being level with the help of lexical methods. These methods are based on the collection of Google Books Ngram and mass surveys of informants and reflect the perception of life of people recorded in millions of published books.

Keywords: *happiness, well-being, income, emotions, lexis*

POKA, KAK AND EXPLETIVE NEGATION

Sergei Tatevosov
Lomonosov Moscow State University, Moscow
Tatevosov@gmail.com

The paper sketches a semantic analysis of the temporal conjunction *poka* in Russian. It argues, in particular, that negation, which the literature treats as expletive, is in effect fully compositional. A ignificant argument supporting this view comes from the interpretation of the Russian counterparts of Enligsh *since*-clauses where negation gets interpreted in much the same way as in *poka*-clauses.

Keywords: *temporal semantics, temporal adjuncts, negation*

SOME RESULTS OF GRAMMATICAL DISAMBIGUATION IN THE CORPUS OF TATAR LANGUAGE

A.R. Fazlyeva
Kazan Federal University, Kazan
afalina.year@gmail.com
B.E. Khakimov
Kazan Federal University, Kazan
"Applied Semiotics" Research Institute of the Tatarstan Academy of Sciences
khakeem@yandex.ru

In the paper results of the context-based study of the selected types of grammatical and functional homonyms are presented. Context features of the

frequent functional homonyms and homo-morphemes are discussed. The results can be used to develop formal disambiguation rules for the corpora of the Tatar language.

Keywords: *homonymy, polysemy, ambiguity, disambiguation, text corpus, grammatical homonym, Tatar language*

ENGLISH ABSTRACTS. SECTION 3

CONTEXTUAL RULE METHOD FOR MORPHOLOGICAL DISAMBIGUATION IN THE TATAR LANGUAGE

Ramil R. Gataullin, Rinat A. Gilmullin
Research Institute of Applied Semiotics
of Tatarstan Academy of Sciences, Kazan
ramil.gata@gmail.com, rinatgilmullin@gmail.com

For last years, scintists from Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences have been developing Tatar language corpus, which contains by these days more than 40 million word usages. Morphological features are automatically annotatted, but the problem of morphological ambiguity has not been solved yet. Since Tatar language is one of agglutinative languages, types of morphological ambiguities are theoretically infinite. It means that machine learning algorithms will not cover all cases of them, due to data sparseness. So it is necessary to combine them with rule base methods for these cases. And this work introduces such tool for Tatar language.

Actually, the rule development tool was constructed as web service. It is available at http://tatcorp.antat.ru. To get more efficiency, “crowdsource” approach is used. It means, that rules are created with help of many users of systems. Of course, as production rules will be used only successfully tested rules.

For now the tool is in testing phase. By January 2016, nine testing rules were created and tested on corpus data. Next step will be the development of machine learning algorithms, which will be combined with this tool.

Keywords: *morphological disambiguation, contextual method, Tatar language*

MULTIFUNCTIONAL MODEL OF TURKIC MORPHEME: CERTAIN ASPECTS

Dzh. Sh. Suleymanov,
Scientific Research Institute of Applied Semiotics, Tatarstan Academy of Sciences
dvdt.slt@gmail.com
A.R. Gatiatullin,
Scientific Research Institute of Applied Semiotics, Tatarstan Academy of Sciences
agat1972@mail.ru
A.B. Almenova,
Scientific Research Institute of Applied Semiotics, Tatarstan Academy of Sciences
almen_akmaral-baijan@mail.ru
A.M. Bashirov
LTD "TemirTech", Vice-Director for Technical Development
a.basheerov@gmail.com.

This article is devoted to description of a multi-functional model of the Turkic morpheme; and this model, when filled the corresponding content, may have different practical applications, primarily as a resource base for the software, carrying out the computer processing of the Turkic languages.

The main component of the multifunctional model is relational-situational model used to describe a variety of aspects of the model.

This model may serve as an effective tool in comparative studies of turkologists, in particular, for the comparative analysis of Turkic language units.

Keywords: *Turkic morpheme, multi-functional model of Turkic morpheme, relational-situational model, linguistic processors.*

APPLICATION OF NEURAL NETWORKS FOR FINANCIAL MARKETS PREDICTION

F.M. Gafarov, Z.T. Galimhanova
Kazan Federal University, Kazan
fgafarov@yandex.ru, zuhra-1996@mail.ru

Forecasting financial time series - is an important element of any investment. For a successful game in the financial market is necessary to develop a system that on the past behavior will predict the time series

dynamics in subsequent times. In this work we investigated the possibility of using feedforward neural networks to predict the dynamics of exchange rates against the US dollar.

Keywords: *financial markets, time series forecasting, the currency exchange market, the exchange rate, artificial intelligence, neural networks.*

MORPHOLOGICAL ANALYZER AS A DLL IN THE SYSTEM OF CROSS-PLATFORM JAVA

Zheltov V.P., Gubanov A.R.
Federal state budget educational institution of higher professional education "Chuvash State University named after I.N. Ulyanov", Cheboksary
chnk@mail.ru, AlexGubM@gmail.com

Modern search engines morphological analysis module is used for text indexing allows you to generate a search query in natural language. As an apparatus for building language models for simulating the interaction of the system components, i.e. as a meta-machine in the morphological analyzer Petri nets are used.

Keywords: *morphological analyzer, DLL, cross-platform system Java.*

DESIGN OF VIRTUAL KEYBOARD FOR TATAR-SPEAKING USERS ON THE BASIS OF THE MOBILE OPERATING SYSTEM ANDROID

A.V. Danilov, T.A. Ilyasov
Kazan Federal University, Kazan
tukai@yandex.ru

This paper presents the elaboration of a virtual keyboard for the Tatar language based on the Android mobile operating system. The authors analyzes the features that influenced the development. They describe the main components of the keyboard and their functions. In particular, authors describe the process of developing a predictive text input system. This development is positioned as an example of preservation and development of the Tatar language in social - humanitarian sphere using the modern ICT.

Keywords: *Tatar language, localization, Android, keyboard, predictive dictionary, mobile communication.*

DEPENDENCE OF THE ENERGY SEGMENT WAVELET TRANSFORM SPEECH FROM THE SCALE FACTOR

[1]P.V. Zheltov., [2].P. Zheltov , [3]I. Semenov , [4]A.K. Shurbin
Federal state budget educational institution of higher professional education "Chuvash State University named after I.N. Ulyanov", Cheboksary
[1]chnk@mail.ru, [2]zheltov42@mail.ru,
[3]syundyukovo@yandex.ru, [4]shurti@mail.ru

location Pictures of phonemes in words and sentences can be set by examining the dependence of the energy spectrum of the wavelet-segments of the scale factor. For research use MHAT-wavelet. speech signal wavelet analysis shows that the vowel phonemes have maximum energy at average values.

Keywords: *wavelet transform, the scale factor, a segment of the Fourier transform, the speech signal.*

KARAKALPAK-UZBEK TEXT TRANSLATOR BASED ON RECURRENT NEURAL NETWORK

A.A. Kadirov
Nukus branch of the Tashkent University of Information Technologies
censor2005@mail.ru

The article discusses the application of neural networks for the development of Karakalpak-Uzbek text translator. Due to the close relation of these languages, this approach can simplify the learning of the neural network. We give a preliminary bound for the complexity of development.

Keywords: *translation neural network, karakalpak, uzbek*

E-CATALOGUE OF THE VIRTUAL MAKHMUTOV MUSEUM-LIBRARY: DOCUMENT REPRESENTATION AND SEARCH

M.I. Kurmanbakiev. O.A. Nevzorova, D.Sh. Suleymanov, D.M. Shakirova
Research Institute of Applied Semiotics, TAS, Kazan
write@marat.link, onevzoro@gmail.com, dvdt.slt@gmail.com,
shdilyara_m@mail.ru

This article presents the model of e-catalogue of the Makhmutov virtual museum-library project. Authors describe the structure of e-catalogue, the model of document metadata and the features of document searching.

Keywords: *virtual museum-library, information system, e-catalogue, search for information, metadata model*

SPELLING CORRECTION USING CHUNKING IN THE MODEL OF DEPENDENCY TREES IN RUSSIAN AND ENGLISH

Anisimov Ivan Sergeevich
LLC. Moscow, Russia
ivananisimov2010@gmail.com Yandex,
Makarova Elena Andreevna
Institute of Linguistics RAS. Moscow, Russia
antaresselen@mail.ru

Polyakov Vladimir Nikolaevich
Institute of Linguistics RAS;
NUST "MIS&S" Moscow, Russia
pvn-65@mail.ru

Abstract

The article describes a method of spelling correction using chunking in the model of dependency trees. This method can eliminate the possible alternative correction generations of erroneous words according to the morphological dictionary. The method is based on correction generation with help of Levenshtein method, of building a graph of all possible chunks and further transformation of the grapgh into a set of trees. Then the program choses a tree that consists of the biggest number of words. The method can be applied in a batch mode.

Keywords*: spelling correction, chunking, syntax, dependency model, Russian, English.*

DEVELOPMENT OF SEMANTIC MODEL OF THE MACHINE TRANSLATION SYSTEM FOR THE RUSSIAN-KAZAKH LANGUAGE PAIR

D.R. Rakhimova
Al-Farabi Kazakh National University, Almaty, Kazakhstan
diana.rakhimova@kaznu.kz

The article presents the semantic model of the machine translation system for the Russian-Kazakh language pair on the basis of the offered expanded attribute grammar is described. This method has been developed taking into account specifics of two various languages and their comparison on the example of simple sentences.

Keywords: *semantics, machine translation, Russian and Kazakh.*

ABOUT SOME APPROACHES TO THE PROBLEM OF AUTOMATIC SPEECH RECOGNITION

Tazhev Boris Petrovch
Institute of Computer Science and Problems of Regional Management
Kabardino-Balkarian Research Center of RAS
boristazhevart@mail.ru

Gurtueva Irina Aslanbekovna
Institute of Computer Science and Problems of Regional Management
Kabardino-Balkarian Research Center of RAS
gurtueva-i@yandex.ru

There is concise review of current approaches to solve the problem of automatic speech recognition in the given article. Acoustic phonetic, pattern recognition, artificial intelligence approaches are considered.

Key words: *speech recognition, hidden Markov model*

AUTOMATIC MORPHEMIC ANALYSIS OF VERBS IN AGGLUTINATIVE LANGUAGES

B.P. Tazhev, I.P. Tazhev, A.M. Ksalov
Institute of Computer Science and Problems of Regional Management of KBSC of the Russian Academy of Sciences, Nalchik
Kabardino-Balkarian State Agrarian University named after V.M. Kokov, Nalchik
boristazhevart@mail.ru, itazhev@yandex.ru, arsenksal@gmail.com

This article considers potentials of morphemic analysis of verbs in agglutinative languages in the context of Kabardian-Circassian language, and its use in different systems.

Keywords: *agglutinative language, morphemic analysis, context-free grammar*

MORPHOLOGICAL ANALYSIS OF KAZAKH BASED ON COMPLETE SYSTEM OF AFFIXES

U.A.Tukeev, A.Turgynova

Al-Farabi Kazakh National University, Almaty, Kazakhstan
ualsher.tukeyev@gmail.com, turgynovaa@gmail.com

This paper describes an approach to morphological analysis of Kazakh based on complete system of affixes that secure correct analysis of any word form. We propose the following algorithm of morphological analysis: 1) to select a word stem and affixes, and then to fix grammatical characteristics of a word form; 2) to build transducer (Mealy machine) with one state. The experiments have shown the high efficiency of the proposed approach to morphological analysis.

Keywords: *transducer, morphological analysis, the Kazakh language.*

LANGUAGE MODELS COMPARISION FOR TATAR SPEECH RECOGNITIONS SYSTEM

A.F. Khusainov
Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan
Kazan Federal University, Kazan
Khusainov.aidar@gmail.com

The research is focused on building different types of language models for the Tatar language. We are planning to use built language model as the part of our speech recognition system, but it can also be used in variety of tasks, e.g. machine translation, spell-checking.

Keywords: *language modelling, the Tatar language, speech recognition.*

ONTOLOGICAL MODELING MORPHOLOGICAL RULES OF ADJECTIVES IN KAZAKH AND TURKISH

A. Sharipbay, G. Bekmanova, L. Zhetkenbay
Eurasian National Universitynamed after L.N.Gumilev, Astana
sharalt@mail.ru, gulmira-r@yandex.ru, jetlen_7@mail.ru

This work is devoted to ontological modelling of morphological rules of functioning of adjectives in Kazakh and Turkish. The study emphasize similarities and differences between these languages. The results may be used to create semantic translation systems from Kazakh to Turkish, and vice versa, and for learning these languages via computers or the Internet.

Key words: *Natural language processing, adjective, morphological rules, ontological modeling, machine translate.*

ENGLISH ABSTRACTS. SECTION 4

THE ISSUE OF COMPILING AND DEVELOPING EVENK-ENGLISH DICTIONARY

A.B. Anisimov
North-Eastern Federal University, Yakutsk
anis_and@mail.ru

This article deals with the issue of compiling and developing Evenk-English dictionary. The dictionary like that has not yet been compiled in modern Russian lexicography. The article discusses two basic work stages on the Evenk-English dictionary.

Keywords: *the Evenks, the Evenk language, dictionary*

FORMATION OF TERMINOLOGICAL SYSTEM OF THE TATAR AND CHUVASH LANGUAGES IN THE FIELD OF INFORMATION AND COMMUNICATION TECHNOLOGIES AND ITS IMPLEMENTATION ON THE EXAMPLE CHETYREHYAZYCHNY DICTIONARY

D. SH. Sulejmanov, A. F. Galimjanov
Tatarstan Academy of Sciences, Research Institute of "Applied Semiotics"
dvdt.slt@gmail.com, anis_59@mail.ru

The article deals with the problems of terminology on computer science and information technology, and the formation of new terms for the Tatar and Chuvash languages.

Keywords: *computer science and information technology, the formation of new terms*

LINGUISTIC WAYS OF CREATING IMAGERY IN WORKS OF A.YENIKY AND THEIR TRANSLATION INTO RUSSIAN

E. Denmukhametova, R. Garayeva
Kazan Federal University, Kazan
denmukhametova@gmail.com, rezeda-92g@mail.ru

This article discusses linguistic ways of creating imagery in works the Tatar writer A. Yeniki and ways of translation of them into Russian; the authors analyse the felicitous findings of translators in using translation transformations and describe some inaccuracies in translated texts that lead to distortion of the literary images created by the Tatar writer.

Keywords: *the Tatar language, translation, literary image, transformations,literary text.*

PRESENTATION OF SPATIAL-TEMPORAL RELATIONS IN KYRGYZ LANGUAGE

Sonunbubu Karabaeva
Bishkek (Kyrgyzstan)
sonun2008@mail.ru, k.sonun@gmail.com

The article is devoted to the peculiarities of concepts perception such as “space” and “time” in the Kyrgyz language. We investigated spatial relations in Kyrgyz language [3], [4]. Here we consider relations between space and time.

Key words: *time and space; culture and language; spatio-temporal relations in the Kyrgyz language.*

CREATING A DATA BASE OF THE WORD-STOCK OF THE TUVAN LANGUAGE

B. Oorzhak, A. Khertek, M. Kuzhuget, A. Salchak, V. Ondar, E. Chamzyryn

Tuvan State University
oorzhak.baylak@mail.ru

The article presents the development of data base of the word-stock of the Tuvan language for the electronic corpus of Tuvan. The database will be

implemented as a reference search engine that will enable search for text fragments corresponding to purposes of the user.

Keywords*: the Tuvan language, database, word-stock, semantic classes, =semantic subclasses, lexical compatibility, automated search engine.*

ELECTRONIC DATA BASE ATLAS RUSSIAN DIALECTS

A.G. Pilugin, F.I. Salimov, V.D. Solovyev
Kazan Federal University, Kazan
pag@kcn.ru, Farid.Salimov@kpfu.ru, maki.solovyev@mail.ru

The article deals with issues related to the creation of an electronic database dialect atlas of Russian dialects

Keywords: *dialects of the Russian language, database*

ETHNOLINGUISTIC ELECTRONIC DICTIONARY TERMS TATAR LANGUAGE

F.I. Salimov, R.F.Salimov
Kazan Federal University, Kazan
Farid.Salimov@kpfu.ru, Rust1k@gmail.com

The article describes the experience of creating an electronic dictionary of ethno-linguistic, built on the basis of materials collected by scientists of the Institute of language, literature and art of Tatarstan Academy of Sciences during field expeditions.

Keywords: *dialects of the Russian language, database*

FOLK TRADITIONS THROUGH THE PRISM OF LANGUAGE

F.S. Bajazitova, F.I. Salimov, L.G. Habibullina
The Institute of language, literature and art of Tatarstan Academy of Sciences, Kazan Federal University, Kazan
fbajazit@gmail.com, farid.salimov@kpfu.ru, valievalg@mail.ru

In the scientific circulation, introduced abundant new material, "the monuments" of popular culture – wedding ritual texts. We are talking about dialects of the language and culture, in particular about wedding ritual

terminology and dialect-folklore texts. The material collected and systematized in all regions densely populated by Tatars.

***Keywords**: ethnolinguistics, ethnography, dialects, Tatar language*

ENGLISH ABSTRACTS. SECTION 5

EFFICIENCY OF THE INTERNET PLATFORM DUOLINGO WHEN TEACHING FOREIGN LANGUAGES

A.I. Abdullin
Kazan Federal University, Kazan
Aidar-abd@mail.ru

The article describes the experience of using the Internet platform duolingo in primary and secondary schools as well as expands the introduction of a mechanism in the educational process. This experiment clearly demonstrates the effectiveness of the use of the Internet platform when teaching English

***Keywords:** Duolingo, English, Internet, platform,* efficiency

USE OF ICT IN THE CONTENT AND LANGUAGE INTEGRATING TEACHING

M.A. Romanova, R.R. Zaripova, L.L. Salekhova
Kazan Federal University, Kazan
romanova.maria.rus@yandex.ru, rinata-z@yandex.ru,
salekhova2009@gmail.com

This paper surveys the issues connected with application of ICT in Content and Language Integrated Learning (CLIL). A characteristic of CLIL is given together with linguo-information competence formed as a result of ICT introduction in teaching process.

***Keywords:** Content and Language Integrated Learning, Linguo-Information Competence, Information and Communication Technologies, Foreign Language.*

THE DEVELOPMENT OF INFORMATION COMPETENCE OF STUDENTS OF HUMANITIES (Evidence from the Information Technology course)

M.A. Lukoyanova
Kazan Federal University, Kazan
marina-lkn@yandex.ru
R.R. Ibragimova
Kazan Federal University, Kazan
ibragimova1492@gmail.com

The article describes the solution to the problem of the development of information competence of students of humanities able to use information technology in professional activities.

Keywords*: information competence, Information Technology, students of humanities.*

INTELLECTUALIZATION OF TESTING PROCESSES IN FOREIGN LANGUAGE TUTORING EXPERT SYSTEM

Mamedova M.H.
Institute of Information Technologies, ANAS, Baku, Azerbaijan
depart15@iit.ab.az
Guliyeva Z.Y.
Institute of Information Technologies, ANAS, Baku, Azerbaijan
guliyeva_z_y@hotmail.com

The article considers problems of diagnostic test-block (DTB) designing within the framework of suggested conceptual approach to the expert tutoring system assigned for English language teaching. By the example of grammatical module of DTB the architecture, functioning principles and structural components of knowledge representation model, and expert requirements to new test compiling incorporated into educational content of the latter are illustrated.

Keywords: *the expert tutoring system, diagnostic test-block, grammar module, knowledge base, linguistic variable*

TOWARDS USING CORPUS ORIENTED APPROACH IN TEACHING OF TATAR LANGUAGE

A.A. Mubarakshina
Kazan Federal University, Kazan
mubarakshinaaa@poelidovolen.ru
B.E. Khakimov
Kazan Federal University, Kazan
"Applied Semiotics" Research Institute of the Tatarstan Academy of Sciences
khakeem@yandex.ru

The article discusses the problem of teaching national languages based on the corpus-oriented approach on the example of the Tatar language. General methodological aspects of using corpora in language teaching and the opportunities of using text corpora in Tatar language teaching are investigated. The examples of student corpus-based research tasks are given.

Keywords: *corpus of texts, Tatar language, language teaching, national education*

USE OF ICT IN THE CONTENT AND LANGUAGE INTEGRATING TEACHING

M.A. Romanova, R.R. Zaripova, L.L. Salekhova
Kazan Federal University, Kazan
romanova.maria.rus@yandex.ru, rinata-z@yandex.ru,
salekhova2009@gmail.com

This paper surveys the issues connected with application of ICT in Content and Language Integrated Learning (CLIL). A characteristic of CLIL is given together with linguo-information competence formed as a result of ICT introduction in teaching process.

Keywords: *Content and Language Integrated Learning, Linguo-Information Competence, Information and Communication Technologies, Foreign Language.*

THE USE OF INFORMATION AND COMMUNICATION TECHNOLOGIES IN SPECIAL PRESCHOOL EDUCATION

M.M. Romanenko, R.R. Zaripova
Kazan Federal University, Kazan
MMRomanenko@stud.kpfu.ru**,** rinata-z@yandex.ru

The article discusses the various ICT developed by domestic and foreign scientists, designed for preschool children with various developmental disorders (hearing impairment, visual impairment, children with attention deficit hyperactivity disorder, and others.).

Keywords*: ICT, pre-school organizations, special education, speech therapists*

WEB 2.0 IN TEACHING FOREIGN LANGUAGES

L.L. Salekhova
Kazan Federal University
salekhova2009@gmail.com

K.S. Grigorieva
Kazan National Research Technical University named after A.N. Tupolev
grigks@yandex.ru

The article describes possible usage of web 2.0 services in teaching foreign languages in a technical university. Classification and characteristics of web 2.0 are presented. The authors share their experience of implementing web 2.0 in teaching ESL.

Key words: *ESL, web 2.0, social services, teaching, CLIL.*

THE IMPLEMENTATION OF WEB 2.0 TECHNOLOGIES IN THE ORGANIZATION OF THE INDEPENDENT WORK OF SECONDARY SCHOOL STUDENTS ON THE ENGLISH LANGUAGE

A.N. Ulyanova
Kazan Federal University, Kazan
anna.ulyanova.92@mail.ru

The article examines Web 2.0-technologies, namely Web 2.0 platform Nearpod in the process of organization of independent work of high school students at the English language lessons. We describe the advantages of this type of independent work, as well as the results of the done experimental work.

Keywords: *Web 2.0 technologies, Nearpod. independent work of students*

DEVELOPMENT OF WEB-APPLICATIONS FOR REMOTE SOLVE MATH PROBLEMS

L.E. Khairullina, M.M. Zagidullin
Kazan Federal University, Kazan
liliya-v1@yandex.ru, Zagmysa@mail.ru

This article describes a way to develop a Web application for a remote solve mathematical problems by means of MATLAB and Asp.net.

Keywords*: Web application, MATLAB, Asp.net*

CLOUD TECHNOLOGIES AS THE LEADING SMART - LEARNING TOOLS

A. Khusainova
Kazan Federal University, Kazan
alfirahamzovna @gmail.com

The article describes the procedure of creating the educational environment of discipline on the basis of Smart Education - modern teaching methods, based on cloud technologies and providing interactivity of the educational process, free access to many information sources, the opportunity providing of the conditions for building an individual educational path, . developing the skills of communication, collaboration, and creative approach to problem solving.

Keywords: *Smart Education, cloud technologies, the educational environment of discipline.*

MULTIMODALITY IN EDUCATION

A.F. Khusainov, D.D. Yakubova
Kazan Federal University, Research Institute of Applied Semiotics, Kazan
khusainov.aidar@gmail.com, suleymanovad@gmail.com

Zh.E. Vavilova
Kazan State Power Engineering University, Kazan
zhannavavilova@mail.ru

A.V. Parkalov
Belarusian State University of Informatics and Radioelectronics, Minsk
a.parkalov@gmail.com

This paper describes the implementation of the multimodal approach in education. The authors define the concepts of multimodal literacy and multimodal text, as well as characterize a learning software system that adapts to the possibilities of the student.

Keywords: *multimodality, multimodal literacy, multimodal text, computer training system*

IMPLEMENTATION OF E-EDUCATIONAL RESOURCES IN SCIENTIFIC RESEARCH AND PROJECT ACTIVITIES OF STUDENTS

Iskander E. Yarmakeev
Kazan (Volga region) Federal University, RUSSIA
ermakeev@mail.ru
Albina R. Abdrafikova
Kazan (Volga region) Federal University, RUSSIA
albina112000@mail.ru
Tatiana S. Pimenova
Kazan (Volga region) Federal University, RUSSIA
pimenova_ts@mail.ru

ABSTRACT

The article deals with the main types, stages and ways of organizing research and project activities of students with the use of electronic educational resources (E-resources). General and specific requirements for the

content and design of research work, its structure and main elements were determined. The evaluative criteria were outlined. The authorial E-resource "Scientific research and project activities of students" contains samples of a scientific article, a scientific thesis and a scientific project with full description and orients students how to define the relevance of the research topic, to set objectives, to distinguish between the subject and the object of study; to determine the hypothesis, to explain theoretical and practical significance; to choose research methods. Research areas, methodical recommendations for faculty and students, lists of references are suggested. The article presents the results of an experiment that allows the authors to state that the use of E-resources contributes to the scientific culture of students while conducting scientific research and project activities. The E-educational resource "Scientific research and project activities of students" was created for students who do a bachelor's degree with the aim in view to help students prepare for conducting and writing research papers on methodical issues in teaching foreign languages.

Keywords: *scientific research, project activity, scientific article, assessment criteria, theoretical and practical significance, scientific rhetoric.*

С О Д Е Р Ж А Н И Е

Раздел I. **Семантические технологии**

Автоматизированная обработка архивной коллекции научного журнала «Электронные библиотеки» *Д.Ю. Ахметов* *4*
Разработка приложения поддержки принятия управленческого решения *А.Ф.Бакунина, К.С.Цыбенко* *8*
Модели векторного представления слов на основе технологии word2vec *Ф.М. Гафаров, Э. И. Шайдуллина, В.Р. Гафарова* *13*
Сервис интеграции новостных лент на платформе управления электронными научными журналами *А.Н. Герасимов* *19*
Семантический анализ больших коллекций научных документов *А.М. Елизаров, Е.К. Липачёв, Ш.М. Хайдаров* *21*
Экосистема Ontomath и проект всемирной цифровой математической библиотеки *А.М. Елизаров, Н.Г. Жильцов, А.В. Кириллович, Е.К. Липачёв, О.А. Невзорова* *25*
Построение модели предметной области и сети соавторства в области юриспруденции на основе зондирования сервиса Google Scholar Citations *Д.В. Ландэ, В.Б. Андрущенко* *29*
Использование систем связывания данных для установления соответствий между хранилищами библиографических данных *К.С. Николаев, О.А. Невзорова* *34*
Разработка онтологической модели операции коммерческих банков *С.А. Поздеева* *45*
Предикативно-образный контроль постановок задач *П.И. Соснин, М.В. Галочкин, А.А. Лунецкас* *49*
Моделирование семантики естественного языка на основе мультиагентной рекурсивной когнитивной архитектуры *Б.П. Тажев, Д.Г. Макоева, И.А. Пшенокова* *57*

Раздел II. **Корпусная грамматика и системы аннотирования (UniTurk)**

Статистические методы исследования конкурирующих глагольных форм в русском языке *Т.И. Галеев* *62*
Функционирование аффикса симилятива -ДАй в татарском языке (на корпусных данных) *А.М. Галиева* *68*
Грамматический портрет татарского текста и его стилевая принадлежность *А.М. Галиева, Р.Р. Гатауллин* *72*
Когнитивная структура отрицания в татарском языке *А.М. Галиева, Д.Ш. Сулейманов* *78*
О морфологической разметке татарских причастий *А.М. Галиева, А.Р. Гатиатуллин* *81*

Роль имен прилагательных в определении тональности текста
Б.Ж. Ергеш, А.А. Шарипбай, Г.Т. Бекманова *85*
Феномен значения в когнитивистике *Г.В. Колпакова* *89*
О реализации системы морфологической разметки крымскотатарского электронного корпуса *Л.Ш. Кубединова, А.Р. Гатиатуллин* *94*
О некоторых сложностях морфологической разметки казахских текстов
А.О. Макажанов, А.Н. Султангазина *98*
Использование корпуса текстов при обучении иностранному языку
А.Ф. Мухамадьярова, Б.Э. Хакимов *104*
Лингвистическое аннотирование залоговых форм глагола языка саха
А.Н. Ноговицына *111*
Семантические параллели в алтайских языках *Д.Б.Рамазанова* *116*
Татар теленнән басма һәм электрон дәреслекләр *Р.К. Сәгъдиева* *121*
Разметка морфологических категорий в национальном корпусе кыргызских текстов *Ташполот Садыков, Бакыт Шаршембаев* *126*
Формирование корпуса с разметкой сущностей в новостных медиа ресурсах для казахского языка *З.Н. Садыкова, В.В. Иванов* *137*
Тарихи эчтәлекле фразеологик берәмлекләр (Н. Исәнбәтнең «Татар теленең фразеологик сүзлеге» материалында) *Ф.Р. Сибгаева* *141*
Лексический подход к оценке динамики уровня благополучия в обществе *В.Д.Соловьев, В.В.Бочкарев* *146*
Пока, Как и эксплетивное отрицание *С.Г. Татевосов* *149*
Татар теле корпусында грамматик омонимияне чишүнең кайбер нәтиҗәләре *А.Р. Фазлыева, Б.Э. Хәкимов* *154*

Раздел III. **Модели и технологии для лингвистических приложений**

Разработка контекстных правил для разрешения морфологической многозначности в корпусе татарского языка
Р.Р. Гатауллин, Р.А. Гильмуллин *162*
Многофункциональная модель тюркской морфемы: отдельные аспекты
Д.Ш. Сулейманов, А.Р. Гатиатуллин, А.Б. Альменова, А.М. Баширов *168*
Применение нейронных сетей для прогнозирования финансовых рынков
Ф.М. Гафаров, З.Т. Галимханова *172*
Морфологический анализатор как dll в кроссплатформенной системе Java
В.П. Желтов, А.Р. Губанов *177*
Разработка виртуальной клавиатуры для татароязычных пользователей на базе мобильной операционной системы Android *А.В. Данилов, Т.А. Ильясов* *183*
Зависимость энергии сегментов вейвлет-преобразования речевого сигнала от значения масштабного коэффициента *П.В. Желтов, В.П. Желтов, В.И. Семенов, А.К. Шурбин* *193*
Каракалпакско-узбекский перевод текстов на основе рекуррентной нейронной сети *А.А.Кадыров* *196*

Электронный каталог виртуального музея-библиотеки М.И. Махмутова: представление документов и поиск *М.И. Курманбакиев, О.А. Невзорова, Д.Ш. Сулейманов, Д.М. Шакирова* *199*

Коррекция правописания с поддержкой чанкинга в модели деревьев зависимостей в русском и английском языках *И.С. Анисимов, Е.А. Макарова, В.Н. Поляков* *206*

Разработка семантической модели системы машинного перевода для русско-казахской языковой пары *Д.Р. Рахимова* *211*

О некоторых подходах к решению задачи автоматического распознавания речи *Б.П. Тажев, И.А. Гуртуева* *217*

Автоматический морфемный разбор глаголов агглютинативного языка *Б.П. Тажев, И.П. Тажев, А.М. Ксалов* *221*

Морфологический анализ казахского языка на основе полной системы окончаний *У.А.Тукеев, А.Тургынова* *225*

Сравнение языковых моделей для системы распознавания татарской речи *А.Ф. Хусаинов* *231*

Онтологическое моделирование морфологических правил прилагательных казахского и турецкого языков *А.Шарипбай, Г.Бекманова, Л.Жеткенбай* *239*

Морфологический анализатор казахского языка на основе онтологического моделирования морфологических правил *А. Шарипбай, Б. Разахова, А. Зулхажав* *248*

Раздел IV. **Лексикографические базы данных**

К вопросу о составлении и разработке эвенкийско-английского словаря *А.Б. Анисимов* *253*

Формирование терминологической системы татарского и чувашского языков в области инфокоммуникационных технологий и его реализация на примере четырехъязычного словаря *Д.Ш. Сулейманов Д.Ш., А.Ф. Галимянов* *259*

Ә.Еники әсәрләре һәм аларның рус теленә тәржемәләрендә образ тудыруда тел чаралары *Э.Н.Денмухаметова, Р.Г.Гараева* *269*

Presentation of Spatial-Temporal Relations in Kyrgyz Language *Sonunbubu Karabaeva* *274*

Создание базы данных лексический фонда тувинского языка *Б.Ч. Ооржак, А.Б. Хертек, М.А. Кужугет, А.Я. Салчак, В.С. Ондар, Е.Т. Чамзырын* *278*

Электронная база данных атласа русских говоров *А.Г. Пилюгин, Ф.И Салимов., В.Д. Соловьев* *282*

Этнолингвистический электронный словарь терминов татарского языка *Ф.И Салимов., Р.Ф. Салимов* *287*

Народные традиции сквозь призму языка *Ф.С. Баязитова, Ф.И. Салимов, Л.Г.Хабибуллина* *293*

Раздел V. **E-learning**

Эффективность использования интернет платформы дуолинго при обучении иностранным языкам *А.И. Абдуллин* *298*
Применение икт в процессе обучения дисциплин на иностранном языке в вузе *М.А. Романова, Р.Р. Зарипова, Л.Л. Салехова* *301*
Формирование информационной компетенции студентов гуманитарных специальностей *М.А. Лукоянова, Р.Р. Ибрагимова* *307*
Интеллектуализация процессов тестирования в экспертной системе обучения иностранному языку *М.Г. Мамедова, З.Ю. Кулиева* *312*
К вопросу об использовании корпусно-ориентированного подхода в преподавании татарского языка А.*А. Мубаракшина, Б.Э. Хакимов*..... *320*
Использование информационно-коммуникационных технологий в специальном дошкольном образовании *М.М. Романенко, Р.Р. Зарипова, Л.Л. Салехова* *324*
Сервисы web 2.0 в обучении иностранным языкам *Л.Л. Салехова* *328*
Внедрение Web 2.0-технологий на базе программы Nearpod в организацию самостоятельной работы учащихся средней школы по английскому языку *А.Н. Ульянова* *333*
Разработка Web-приложения для удаленного решения математических задач *Л.Э. Хайруллина, М.М. Загидуллин* *338*
Облачные технологии как ведущий инструмент Smart- обучения *А.Х. Хусаинова* *342*
Мультимодальность в образовании *А.Ф. Хусаинов, Д.Д. Якубова, Ж.Е. Вавилова, А.В. Паркалов* *348*
Электронно-образовательные ресурсы в научно-исследовательской и проектной деятельности студентов *И.Э. Ярмакеев, А.Р. Абдрафикова, Т.С. Пименова* *351*

ТРУДЫ
МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
ПО КОМПЬЮТЕРНОЙ
И КОГНИТИВНОЙ ЛИНГВИСТИКЕ TEL-2016

Сборник материалов
Выпуск 17

Компьютерная верстка
Р.М. Абдрахмановой

Дизайн обложки
Р.М. Абдрахмановой

Подписано в печать 10.06.2016.
Бумага офсетная. Печать цифровая.
Формат 60x84 1/16. Гарнитура «Times New Roman». Усл. печ. л. 8,37.
Уч.-изд. л. 7,30. Тираж 100 экз. Заказ 292/5.

Отпечатано в типографии
Издательства Казанского университета

420008, г. Казань, ул. Профессора Нужина, 1/37
тел. (843) 233-73-59, 233-73-28

ISBN 978-5-00019-650-2