Оглавление

ОТ СОСТАВИТЕЛЯ

Настоящий выпуск журнала «Электронные библиотеки» представляет собой тематический сборник статей, подготовленных сотрудниками Института прикладной семиотики Академии наук Республики Татарстан (www.ips.antat.ru). Институт прикладной семиотики выполняет фундаментальные и прикладные исследования в области прикладной семиотики, компьютерной и когнитивной лингвистики, интеллектуальных информационных технологий. Одним из ключевых проектов Института в настоящее время является разработка Национального корпуса татарского языка «Туган тел» (www.corpus.antat.ru). Этот проект выполняется в рамках мероприятий по государственной программе «Сохранение, изучение и развитие государственных языков Республики Татарстан и других языков в Республике Татарстан на 2014 – 2020 годы» и нацелен на разработку аннотированного корпуса татарского языка для гуманитарных и образовательных приложений. Корпусное исследование языка дает богатейший материал для построения лингвистических моделей и ресурсов, применимых в задачах перевода, семиотических и когнитивных исследованиях, образовательных программах.

Обзорные статьи А.М. Галиевой и Р.Р. Гатауллина, представленные в настоящем выпуске, посвящены актуальным проблемам корпусной лингвистики – разрешению грамматической многозначности корпусных данных и подготовке лингвистических ресурсов, которые могут быть использованы при семантическом аннотировании корпусных данных. Для татарского языка в настоящее время отсутствуют большие коллекции данных со снятой многозначностью, что затрудняет применение методов машинного обучения в лингвистических приложениях. Тем не менее, авторы рассматривают машинное обучение как наиболее перспективное направление для снятия многозначности и показывают сравнительные оценки методов применительно к различным языкам. В статье А.М. Галиевой дан обзор основных англоязычных электронных лексикографических ресурсов, разработанных для представления семантики глагола. Данное направление имеет важнейшие применения в задачах обработки естественного языка, поэтому статья является весьма своевременной и актуальной.

Статья А.Ф. Хусаинова, А.Х. Хусаиновой и Р.А. Гильмуллина посвящена еще

одному направлению деятельности Института прикладной семиотики, связанному с разработкой интеллектуальных обучающих систем и технологий. Эти исследования ориентированы на разработку семиотических моделей в обучении, их реализацию в прикладных системах и образовательных ресурсах, формирование концепции и методик обучения в условиях инфокоммуникационной образовательной среды. В названной статье описана технология создания электронных версий обучающих материалов в образовательной среде, построенной на основе принципов Smart Education – современного метода обучения, базирующегося на облачных технологиях и обеспечивающего интерактивность учебного процесса.

Публикуемые материалы отражают круг актуальных проблем и задач, которые были представлены для обсуждения на Международной конференции по компьютерной и когнитивной лингвистике TEL-2016, прошедшей в Казани 21–24 апреля 2016 года.

Составитель тематического выпуска

О.А. Невзорова

УДК 81'37

ГЛАГОЛЬНАЯ ЛЕКСИКА В ЛЕКСИКОГРАФИЧЕСКИХ БАЗАХ ДАННЫХ: ОБЗОР ОСНОВНЫХ РЕСУРСОВ

А.М. Галиева

НИИ «Прикладная семиотика» Академии наук Республики Татарстан
amgalieva@gmail.ru

Аннотация

Дан краткий обзор электронных лексикографических ресурсов и баз данных, представляющих семантику глагола. Глагол как одна из самых сложных, семантически ёмких и грамматически содержательных частей речи в любом языке характеризуется разветвлённой системой значений и грамматических форм. Семантическая структура глагола – это комплекс онтологических и реляционных семантических компонентов, которые могут получать свое формальное выражение на разных уровнях языковой структуры. При фиксации глаголов в электронных лексикографических ресурсах разработчики исходят из различных методологических установок и отдают приоритет различным аспектам семантической организации глагольной лексики.

***Ключевые слова**: лексикографические ресурсы и базы данных, семантика, глагол, семантические классы слов.*

ВВЕДЕНИЕ

В последние десятилетия появление поисковых систем и аннотированных лингвистических корпусов значительно расширило арсенал лингвистических исследований. Созданы и продолжают разрабатываться специальные лексикографические ресурсы, позволяющие получить актуальную информацию о семантике и распределении слов отдельных частей речи, о моделях управления лексемы, вариативности лексических единиц разных классов, посмотреть иллюстративный материал на примерах из реальных текстов и т. п.

Компьютерные словари глаголов являются ключевыми в процессе обработки естественного языка, нацеленного на интерпретацию данных. Данная статья представляет собой краткий обзор основных англоязычных электронных лексикографических ресурсов, разработанных для представления семантики глагола.

Идеографический словарь является специфическим объектом, который дает возможность исследовать системные свойства лексики языка, разнообразные сигнификативные и логико-семантические отношения, проявления реальных связей данной лексемы с другими, разнообразные реляционные свойства.

СЕМАНТИКА ГЛАГОЛА: КЛЮЧЕВЫЕ ОСОБЕННОСТИ

Для различных типов слов в лингвистической литературе предлагаются различные форматы представления значения. Глагол – одна из самых сложных, семантически ёмких и грамматически содержательных частей речи с чрезвычайно разветвлённой системой значений и грамматических форм. Семантическая структура глагола – это комплекс семантических компонентов, которые могут получать свое формальное выражение на разных уровнях языковой структуры. Подразделение всей глагольной лексики конкретного языка на семантические классы и семантические подклассы, определение типов отношений между лексическими единицами имеет важное значение не только в теоретическом, но и практическом отношении. Полное семантическое описание глагола должно, с одной стороны, основываться на определении места лексемы в системе языка, представлении его парадигматических отношений с другими словами, с другой стороны, учитывать синтаксический потенциал и типичные коллокации глагольной единицы.

Значение слов с предметным значением кардинально отличается от манифестации содержательной составляющей признаковых – глагольных и адъективных – лексем. Если базовое общекатегориальное значение предметности у существительных проистекают из эссенциалистских установок (философская категория сущности), то в глагольных лексемах отражаются процессуальные признаки, действия и состояния.

Семантика существительных основывается на общекатегориальном значении предметности и «носит абсолютный, самодостаточный характер», в то время как семантика глагольных и адъективных лексем «носит реляционный характер»

в силу того, что в ее основе лежит понятие признака, который «имплицирует некую субстанцию, предмет, которым он должен и может быть придан» [1]. Глагол задает минимальные логико-семантические модели: субъект и его действие или состояние; действие и объект, над которым осуществляется это действие и т. п. Сами эти модели вытекают из семантической валентности — детерминированной лексическим значением необходимости сочетания глагола с другими словами. Следствием этого является то, что семантика глагольного слова — «не элементарна, а комплексна в том смысле, что она отображает не законченное, полное понятие о классе предметов, как это имеет место в предметных именах, а минимальные дискретные «кусочки действительности», приближающиеся к элементарным ситуациям и событиям» [1]. Поэтому «универсальным свойством лексического значения глагольных лексем является то, что каждый полнозначный глагол представляет собой потенциальную синтагму; в содержательном плане признаковое имя формирует свое значение: 1) в акте знакообразования, в номинации с учетом носителя (субъекта, объекта) данного признака; 2) при функционировании в речи, где оно дополнительно уточняется, конкретизируется и формирует тем самым круг сочетающихся с ним предметных имен» [1].

Поскольку семантика глагола имеет множество измерений, классификационные схемы, предложенные разными исследователями, могут в корне отличаться. Имеются разнообразные семантические классификации глаголов, выполненные на материале различных языков. Наиболее часто в литературе упоминается классификация B. Levin, которая классифицирует английские глаголы (около 3100 единиц); основываясь на сходстве их значений и синтаксических свойств; исследователь выделяет семантические классы, принимая в расчет широкий спектр возможных синтаксических преобразований лексем (преимущественно мены диатезы, локативности и пр.), которые отражают значение глагола [2]. Работа B. Levin, основанная на идеях генеративной лингвистики, показывает, что семантические признаки глагола имеют явно выраженную корреляцию с его синтаксическими свойствами и интерпретацией его аргументов. Несмотря на то, что полнота и пригодность классификационной сетки, предложенной B. Levin, для описания реальных языковых фактов вызывает некоторые вопросы (см. об этом ниже), семантические классы, выделенные B. Levin, стали основой для многих других

классификаций и лексикографических баз данных.

ПРОЕКТ FRAMENET

Проект FrameNet (https://framenet.icsi.berkeley.edu) представляет собой практическую реализацию семантики фреймов (Frame Semantics) – лингвистической концепции, предложенной Ч. Филлмором [3–6]. Семантика фреймов является своего рода продолжением падежной грамматики и представляет собой подход к формализованному описанию деятельности человека в контексте ситуации. Основную идею семантики фреймов можно определить следующим образом: без подключения энциклопедических знаний о деятельности человека нельзя понять и описать значение слова. Слово активирует рамки семантических знаний, относящихся к конкретному понятию, с которым оно соотносится. Таким образом, в проекте FrameNet слова группируются согласно концептуальным структурам (фреймам), лежащим в их основе, и их сочетаемость выводится индуктивным путем на основе корпусных данных.

Работа над проектом FrameNet была начата в 1997 году в Международном институте информатики (International Computer Science Institute) в Беркли, Калифорния. FrameNet основан на таких терминах семантики фреймов, как семантический фрейм, элементы фрейма, взаимоотношения между фреймами.

Основание для сравнения	**FrameNet**	**Levin 1993**
Группировки	230 семантических фреймов	193 глагольных класса
Основание	лексическая семантика	синтаксис аргументов
Источники данных	корпусы	литература по лингвистике
Покрытие	2100 сущ., 1700 глаголов (включая аналитические конструкции), 460 прил.	3100 глаголов
Результаты	описания фреймов и аннотированные примеры	глагольные классы и чередования (преимущественно с описанием)

Таблица 1. Сопоставление проекта FrameNet и классификации глаголов B. Levin

Основными особенностями данного проекта являются опора на корпусные данные при выведении семантических и синтаксических обобщений, а также представление валентностей целевых слов, в которых смысловая составляющая выводится через семантику фреймов [4].

В статье [7] представлено сопоставление классификации глаголов B. Levin с представлением глаголов в проекте FrameNet. Для наглядности воспользуемся таблицей из данной работы (количественные данные по FrameNet отражают состояние проекта на 2002 г., когда была опубликована эта статья).

Так как количество категорий в обоих ресурсах сопоставимо (230 фреймов против 193 глагольных классов), можно предположить, что значительная часть фреймов будет так или иначе соотноситься с глагольными классами, выделенными B. Levin. В действительности между анализируемыми группировками имеются значительные расхождения. В проекте FrameNet отнесение предикатов к фрейму обусловлено лишь общностью их семантики, и сходства в синтаксическом поведении лексемы не требуется, соответственно, в один фрейм могут быть объединены глаголы с разными типами мены диатезы и другими альтернациями. В то время как модели альтернаций для B. Levin — решающий критерий, лексемы с разным типом синтаксического поведения в один класс объединены быть не могут. Классификация FrameNet строится на реальных корпусных данных, а классификация B. Levin во многом носит умозрительный характер; авторы статьи [7] отмечают, что во многих классах у B. Levin имеются примеры гипотетических конструкций, которые не подкрепляются корпусными свидетельствами. Кроме того, основной тезис B. Levin о том, что группировка слов согласно их синтаксическому поведению дает семантически релевантные классы, не всегда подтверждается, так как слова, близкие по значению, могут попасть в разные классы и, наоборот, слова, довольно далекие по значению, могут быть объединены в один класс. Кроме того, семантические фреймы, выделенные в проекте FrameNet, позволяют описывать не только глаголы, но и существительные и прилагательные [7].

На рис. 1 представлены результаты на запрос по глаголу *to see* 'видеть' в проекте FrameNet.

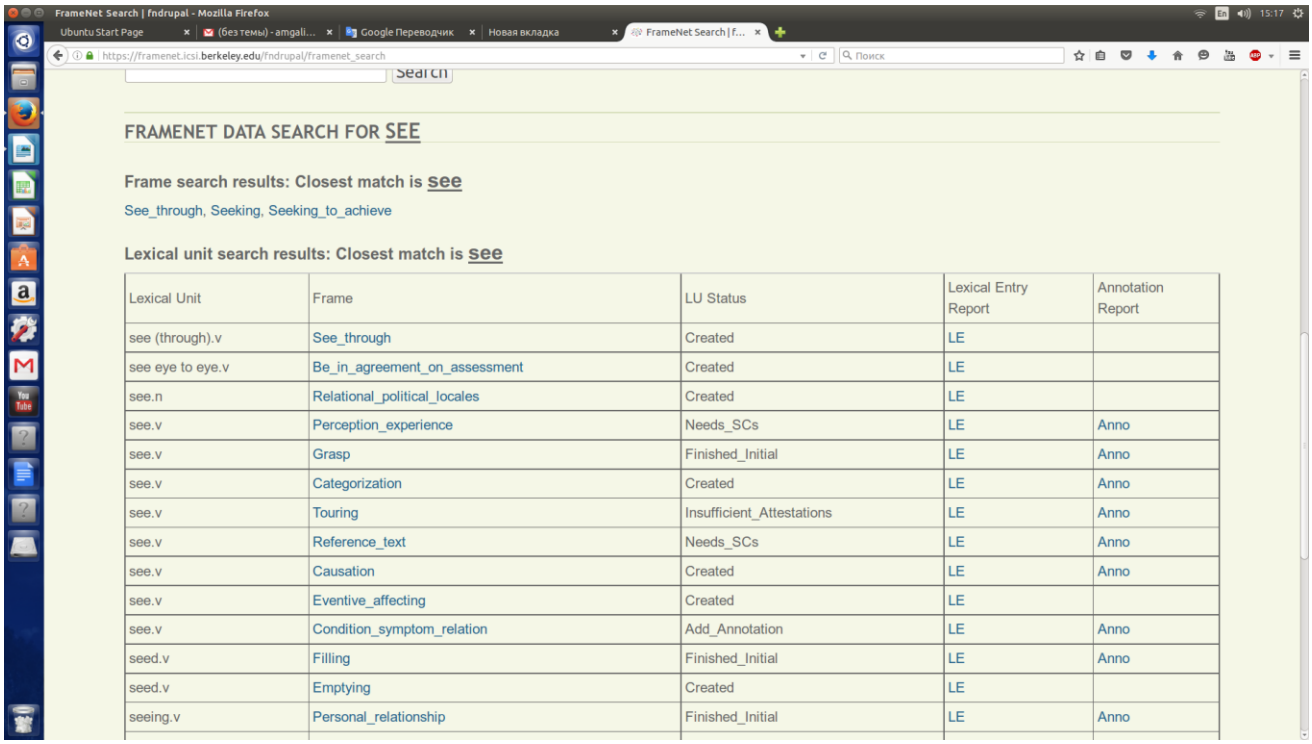


Рис. 1. Фреймы с глаголом *to see* в проекте FrameNet

Основополагающие принципы англоязычного FrameNet'а успешно используются для анализа и описания типологически разных языков: испанского, немецкого, шведского, корейского, японского, китайского и других [8].

В настоящее время в России разрабатывается проект общедоступного электронного словаря глагольных конструкций FrameBank (http://framebank.ru). Разработчики позиционируют свой проект как «создание русского фреймнет-ориентированного ресурса, спроектированного с учетом традиций отечественной лексической семантики и специфики русского языка, где информация о предложно-падежной реализации управления предикатов и поверхностно-синтаксических свойствах других конструкций имеет особую ценность» [9].

FrameBank не является русскоязычной копией FrameNet'а, а имеет свою специфику. Если центральным элементом FrameNet'а являются фреймы — типовые ситуации с известным набором участников и расписанными ролями, то русский FrameBank строится вокруг конструкций конкретных лексем [10].

FrameBank описывает:

- русскую лексическую систему, структуру лексико-семантических групп и

полисемии в русском языке (разработчики англоязычного проекта FrameNet исходят из идеи, что сеть фреймов универсальна для всех языках);

• парадигматические отношения между значениями многозначных слов (как они отражаются в системе связанных с этими значениями лексических конструкций);

• лексико-семантические ограничения на слоты конструкций;

• грамматические особенности русского языка (порядок слов, особенности использования падежей, согласования и т. п.) (10).

Ядро системы FrameBank составляют 2200 ключевых русских глаголов и ассоциированных с ними конструкций и корпусных примеров, для каждого глагола указывается семантический класс. Каждая конструкция представлена в виде шаблона, в котором указаны: а) морфосинтаксические характеристики элементов конструкции; б) синтаксический ранг участника; в) экспликация (роль) участника; г) основные семантические ограничения на заполнение слота [10]. На рис. 2 представлены конструкции с глаголом *видеть* из списка конструкций FrameBank.

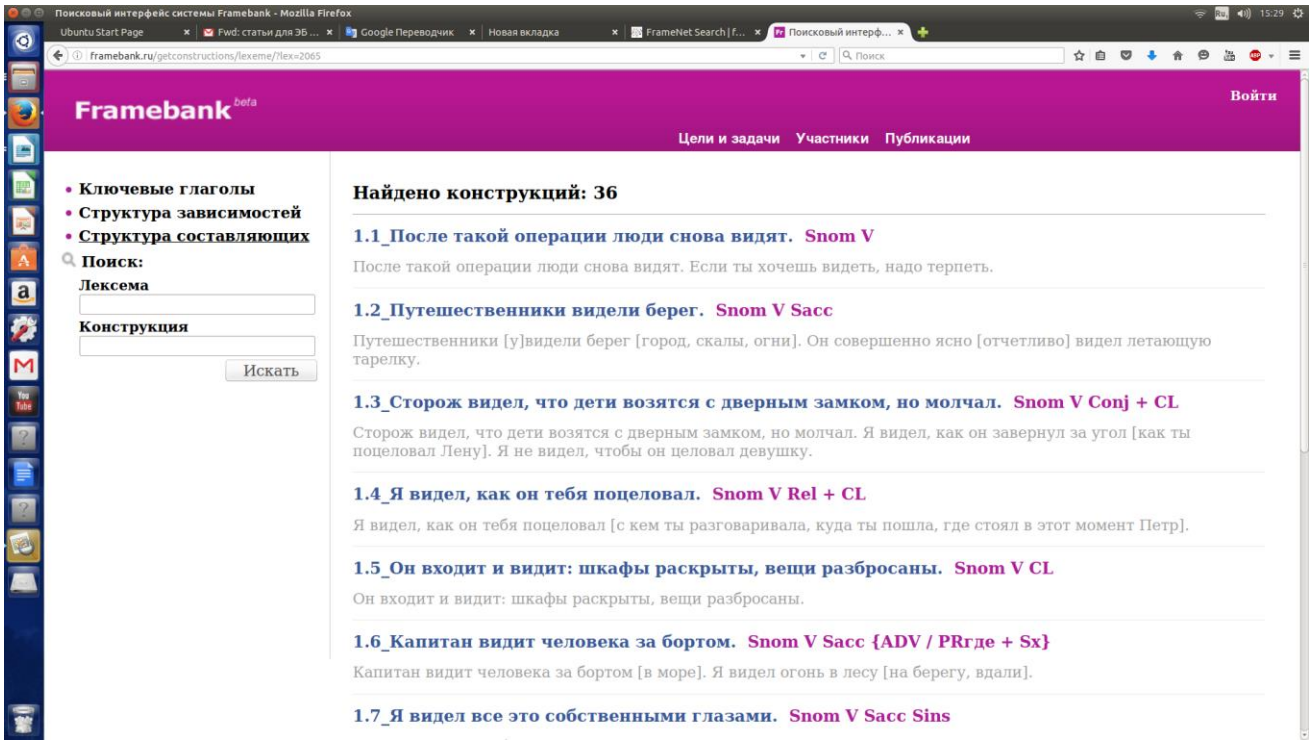


Рис. 2. Конструкции, содержащие глагол *видеть*, в проекте FrameBank

ПРОЕКТ VERBNET

Как уже отмечалось, семантические классы, выделенные B. Levin, стали основой для многих других классификаций и лексикографических баз данных, в частности, проекта VerbNet [11], который, по мнению разработчиков, в настоящее время является самым большим глагольным компьютерным лексикографическим ресурсом. В данном тезаурусе представлен более расширенный, по сравнению с классификацией B. Levin, вариант семантических классов; для каждого класса описаны тематические роли, представлены ограничения на отбор аргументов, а также синтаксические фреймы. VerbNet представляет собой иерархический тезаурус, не привязанный к конкретной предметной области. В VerbNet каждый класс имеет синтаксическое описание, отображающее возможные поверхностные реализации структуры аргумента в типах конструкций, содержащих переходные и непереходные глаголы, сочетания глагола с предлогами, а также большой набор залоговых чередований. Семантические ограничения (например, живое существо, человек, организация) использованы для того, чтобы отобразить типы семантических ролей. Первоначальная версия VerbNet, как уже отмечалось, была основана на классах, выделенных B. Levin, и применение данного ресурса для обработки естественного языка сталкивалось со значительными трудностями. Поэтому исследователи предлагают способы автоматического расширения данного ресурса за счет корпусных данных [11, 12]. VerbNet предполагает отображение связей между глаголами в данном ресурсе и отдельными значениями глаголов, представленными в WordNet, а также фреймами FrameNet.

Рис. 3 представляет фреймы с глаголом *to see* в проекте VerbNet.

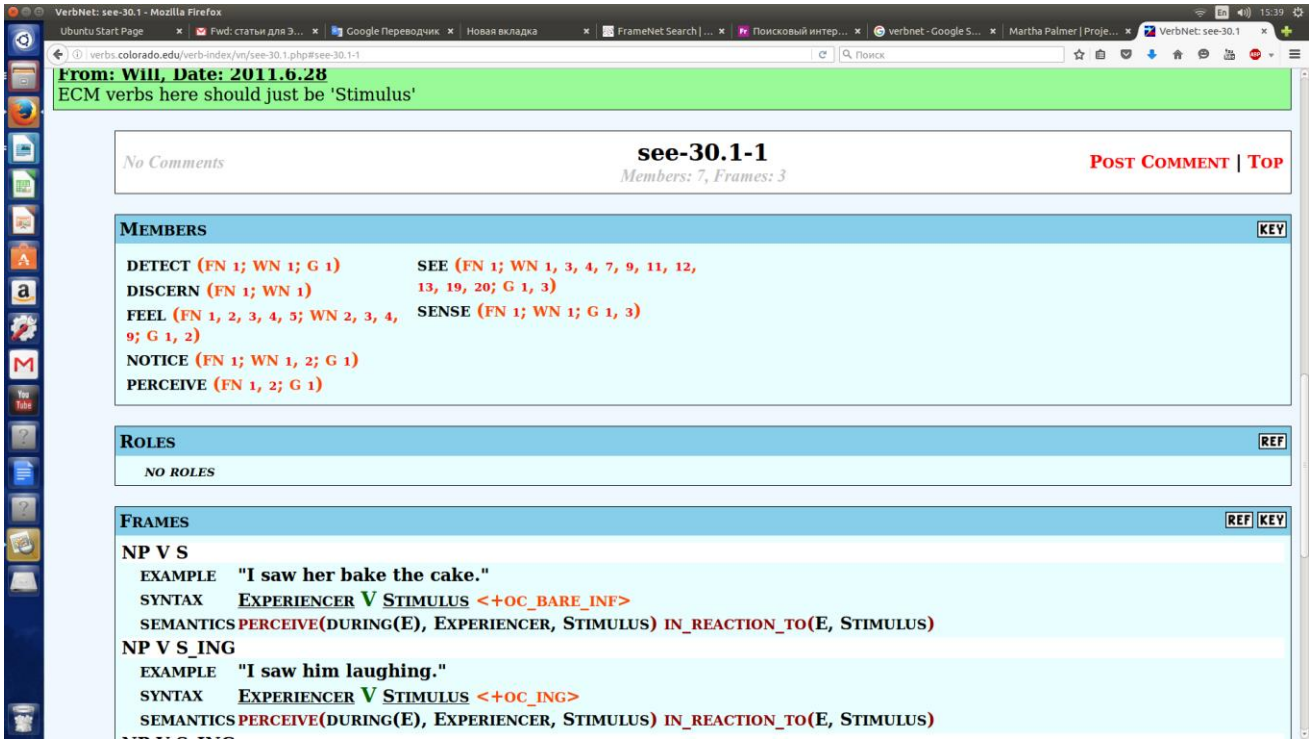


Рис. 3. Фреймы, содержащие глагол *see*, в проекте VerbNet

ПРОЕКТ PROPBANK

Еще одним англоязычным ресурсом, нацеленным на представление глагольной семантики, является PropBank (http://propbank.github.io) который представляет «банк пропозиций» английских глаголов [13]. Изначально проект разрабатывался как корпус текстов с аннотацией важнейших семантических пропозиций. Отношения «предикат – аргументы» были добавлены к синтаксическим деревьям проекта Penn Treebank (http://www.cis.upenn.edu/~treebank). Penn Treebank позволяет приписывать некоторые семантические тэги (например, время или локативность) для ряда конструкций, но не определяет семантические роли грамматического субъекта или объекта. Так как один и тот же глагол в сходном синтаксическом окружении может иметь аргументы с разными семантическими ролями, данные роли не могут правильно определяться автоматически. PropBank нацелен на создание независящего от предметной области корпуса с ручной разметкой семантических ролей. Для каждого глагола определяется набор семантических ролей его аргументов, и таким образом аннотируется каждый пример из Penn Treebank [13].

Рис. 4 показывает представление предиката to see в проекте PropBank.

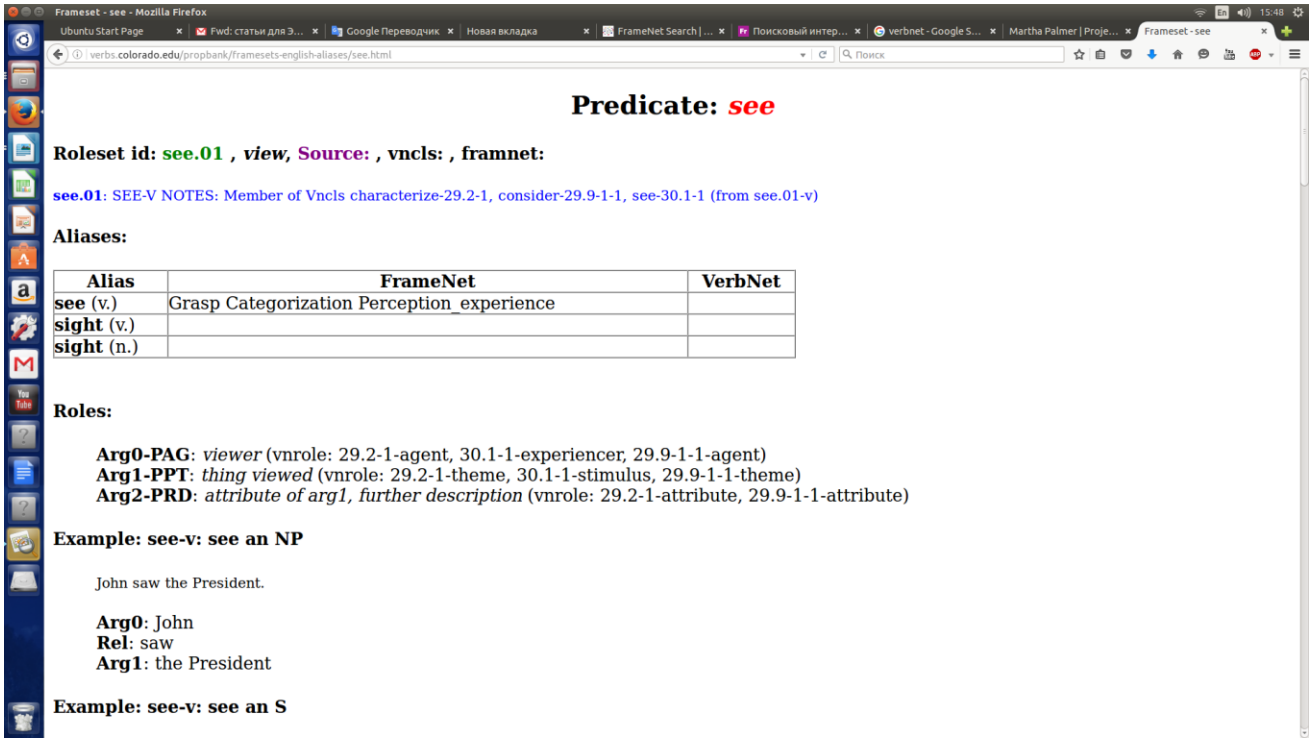


Рис. 4. Глагол *to see* в проекте PropBank

Отметим основные отличия PropBank от FrameNet. PropBank представляет собой специальный ресурс, разработанный для глагольной лексики, в то время как FrameNet описывает значение глаголов (как и других частей речи) в рамках более общей семантики фреймов. PropBank представляет глаголы из конкретного корпуса, а FrameNet выбирает наборы примеров предложений из большого корпуса, и только в некоторых случаях имеются аннотированные отрезки непрерывного текста. Формат аннотирования в PropBank тесно связан с синтаксическим уровнем (требуется, чтобы все аргументы глагола имели синтаксическое выражение; считается, что значения слов отличаются только в том случае, если имеются различия в аргументах), а аннотация в FrameNet в большей степени обусловлена семантикой.

В настоящее время остро встает задача установления соответствий между разными ресурсами, описывающими глагольную лексику, в частности, важное значение имеет связывание проектов VerbNet и PropBank [14]

ПРЕДСТАВЛЕНИЕ ГЛАГОЛОВ В WORDNET

Тезаурус WordNet (https://wordnet.princeton.edu) является одним из наиболее известных лексикографических ресурсов в области компьютерной лингвистики и автоматической обработки текстов. Он был разработан в 1995 году в Принстонском университете. WordNet представляет собой иерархическую сеть лексикализованных понятий (синсетов). Основными единицами структуры wordnet являются синонимические ряды - синсеты, которые связаны между собой различными семантическими отношениями [15, 16]. Синонимические отношения в тезаурусе определяются не между словами, а между отдельными значениями слов.

Несмотря на то, что понятие синонимии является общепринятым, точные критерии синонимичности до сих пор являются предметом дискуссий. В рамках проектов Принстонского WordNet'а и EuroWordNet'а синонимия определяется через понятие взаимозаменяемости: так, в проекте EuroWordNet слова считаются семантически эквивалентными, когда они обозначают один и тот же ряд сущностей, независимо от морфолого-синтаксических, стилистических, диалектных различий, а также различий в прагматическом использовании слова. Кроме того, синонимы не могут быть связаны между собой другими типами семантических отношений [17].

Разработчики wordnet-тезауруса RussNet (http://project.phil.spbu.ru/RussNet/index_ru.shtml) для русского языка критерий взаимозаменяемости рассматривают как дополнительный по отношению к критерию семантической близости. Последний выявляется при дефиниционном анализе, для которого требуется установление идентичности словарных определений или взаимная отсылка в синонимических определениях [18].

В WordNet для разных частей речи используются различные типы отношений. Семантические отношения между синсетами для глаголов описываются следующим образом:

a) отношения следования (Entailment): *идти – шагать*;

b) отношения тропонимии: *сказать – шептать;*

c) отношения каузативности: *есть — кормить.*

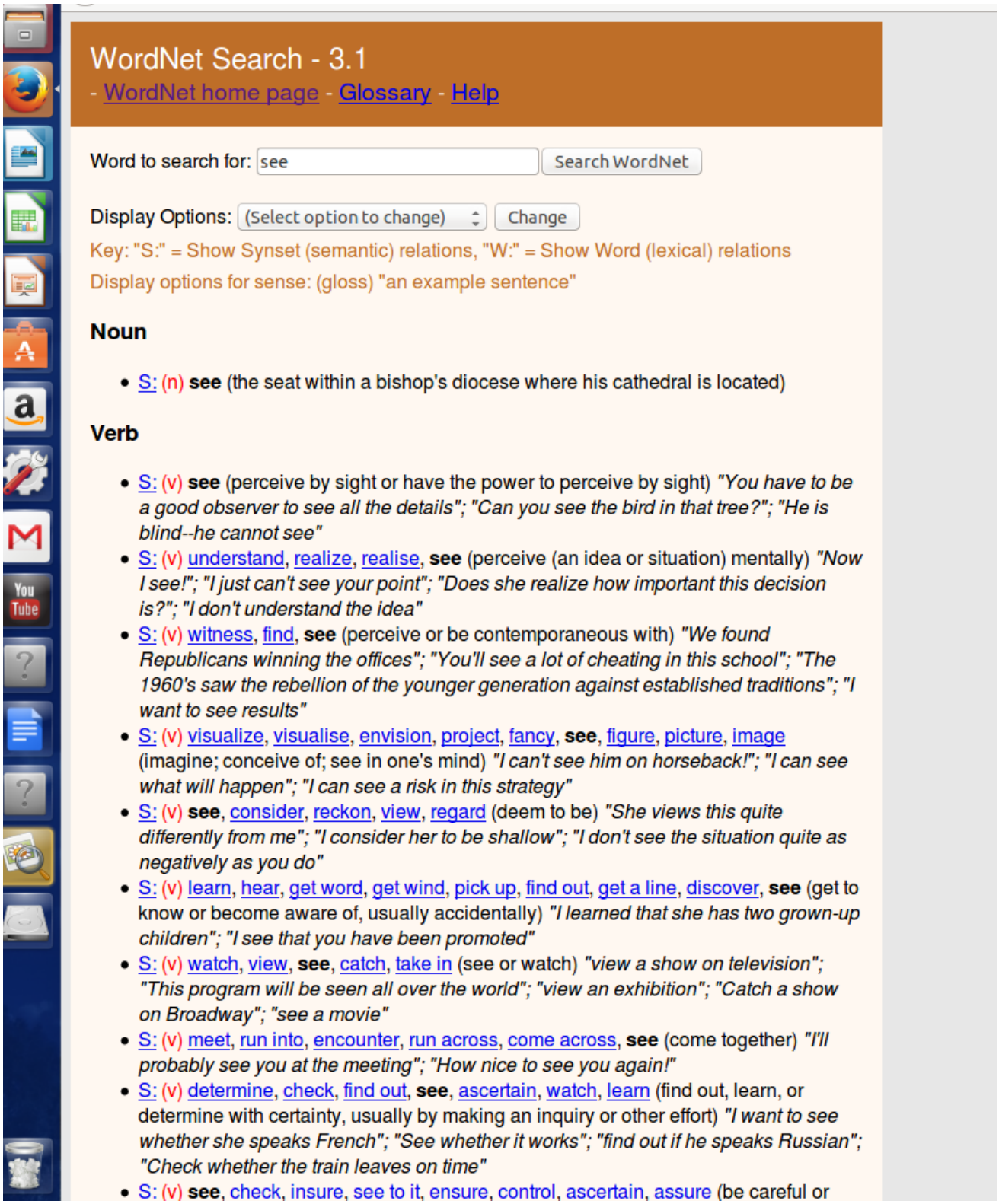


Рис. 5. Глагол to see в проекте WordNet

Отношение тропонимии – это особый случай отношения следования. Глагольные иерархии, представляющие отношение тропонимии, обладают более узкой, но в то же время более кустистой структурой по сравнению с существительными, число уровней в иерархии при этом обычно не превышает четырех [19].

Рис. 5 дает представление о том, как отображается глагол to see в принстонском проекте WordNet.

В тезаурусах типа WordNet фиксируются семантические отношения между глаголами различных классов, при этом валентность глагола и особенности его аргументов, а также типы фреймов, в которых он используется, не принимаются в расчет.

Тезаурусы типа Wordnet созданы во многих странах и успешно используются при обработке естественного языка – при информационном поиске, разрешении многозначности, анализе тональности и др. Информация о существующих и разрабатываемых в настоящее время wordnet'ах представлена на сайте Всемирной WordNet-ассоциации [20].

ЗАКЛЮЧЕНИЕ

К настоящему время разработано значительное число специальных ресурсов для автоматической обработки текстов, большая часть из них создана для английского языка. Особой ролью глагольной лексики в языках можно объяснить то, что многие ресурсы создаются специально для представления значений глаголов. Глагол является синтаксическим ядром предложения, его значение не только номинативно, но и реляционно: именно глагол задает отношения между участниками ситуации и кодирует их в синтаксических структурах.

Многомерность глагольной семантики обусловливает разнообразие подходов к его представлению в лексикографических ресурсах, каждый из которых создавался в рамках конкретных теоретических и методологических установок и ввиду конкретных целей, стоящих перед разработчиками. Значение глагола как синтаксического ядра предложения рассматривается в проектах типа VerbNet и PropBank. FrameNet описывает глаголы в рамках семантики фреймов, а WordNet устанавливает иерархические семантические отношения между отдельными глаголами, не принимая в расчет аргументную структуру глагола.

СПИСОК ЛИТЕРАТУРЫ

1. *Уфимцева А.А.* Лексическое значение: принцип семиологического описания лексики. М.: Наука, 1986. 239 с.

2. *Levin B*. English Verb Classes and Alternations: A Preliminary Investigation. Chicago, University of Chicago, 1993. 348 p.

3. *Fillmore C.* Frame Semantics and the Nature of Language // Annals of the New York Academy of Sciences. 1976. 280 (1). P. 20–32.

4. *Baker C.F., Fillmore C.J., Lowe J.B*. The Berkeley FrameNet Project // Proceedings of the 17th International Conference on Computational linguistics. 1998, August. V. 1. Association for Computational Linguistics. P. 86–90.

5. *Fillmore C.J., Baker C.F.* Frame semantics for text understanding // Proceedings of WordNet and Other Lexical Resources Workshop. Pittsburgh, 2001. URL: http://www.ccs.neu.edu/course/csg224/resources/framenet/framenet.pdf.

6. *Fillmore C.J., Baker C.F., Sato H.* FrameNet as a “Net” // Proceedings of LREC, Lisbon, 2004. V. 4. P. 1091–1094.

7. *Baker C.F., Ruppenhofer J*. FrameNet's Frames vs. Levin's verb classes // Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society. 2002. P. 27–38.

8. *Boas H.C.* (Ed.). Multilingual framenets in computational lexicography: Methods and Applications. 2009. V. 200. Walter de Gruyter. 352 p.

9. FrameBank. URL: http://framebank.ru.

10. *Кашкин Е.В., Ляшевская О.Н*. Семантические роли и сеть конструкций в системе FrameBank // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог 2013. Т. 1. С. 297–311.

11. *Kipper K., Korhonen A., Ryant N., Palmer M*. Extending VerbNet with novel verb classes // Proceedings of the Fifth International Conference on Language Resources and Evaluation – LREC'06. May, 2006, Genoa, Italy: 2006. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1.380.5541&rep=rep1&type=pdf.

12. *Kipper K., Korhonen A., Ryant N., Palmer M*. A large-scale classification of English verbs// Language Resources and Evaluation. 2008. V. 42 (1). P. 21–40.

13. *Palmer M., Kingsbury P., Gildea D.* The proposition bank: an annotated

corpus of semantic role // Computational Linguistics. 2005. V. 31 (1). P. 71–106.

14. *Loper E., Yi S.T., Palmer M.* Combining lexical resources: mapping between PropBank and VerbNet // Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg, the Netherlands. 2007. URL: http://verbs.colorado.edu/~kipper/Papers/semlink_iwcs7.pdf.

15. *Fellbaum C*. *WordNet. An Electronic Lexical Database*. Cambridge, Mass: MIT Press, 1998. 423 p.

16. *Mille G.A*. WordNet: A lexical database for English // Communications of the ACM. 1995. V. 38, No 11. P. 39–41.

17. *Vossen P.* (Ed.) *EuroWordNet General Document. Version 3.* URL: http://vossen.info/docs/2002/EWNGeneral.pdf.

18. *Азарова И.В., Митрофанова О.А., Синопальникова А.А.* Компьютерный тезаурус русского языка типа wordnet. URL: http://www.dialog-21.ru/Archive/2003/Azarova.htm.

19. *Лукашевич Н.В.* Тезаурусы в задачах информационного поиска. М.: Изд-во Московского ун-та, 2011. 511 с.

20. The Global WordNet Associaton. URL: http://globalwordnet.org.

VERBAL VOCABULARY IN LEXICOGRAPHICAL DATA BASES: REVIEWING MAIN RESOURCES

A.M. Galieva

Research Institute of Applied Semiotics of Tatarstan Academy of Sciences

amgalieva@gmail.com

Abstract

This paper gives a brief review of available lexicographical resources and databases representing verbal vocabulary. The verb, being one of the most complicated, semantically intricate and grammatically sophisticated parts of speech, is characterized by multiplicity of senses and forms. Semantic structure of any verb is a complex of ontological and relational meaning components that may find a formal expression on dif-

ferent levels of the linguistic structure. Fixing verbs in electronic lexicographical resources, researchers come from different methodological orientations, and prioritize different aspects of the organization of the semantics of verbal vocabulary.

***Keywords**: lexicographical resources and databases, semantics, verb, semantic classes of words.*

REFERENCES

1. *Ufimtseva A.A.* Leksicheskoye znacheniye: printsip semiologicheskogo opisaniya leksiki M.: Nauka. 1986. 239 s.

2. *Levin B*. English Verb Classes and Alternations: A Preliminary Investigation. Chicago, University of Chicago, 1993. 348 p.

3. *Fillmore C.J*. Frame Semantics and the nature of language // Annals of the New York Academy of Sciences. 1976. 280 (1). P. 20–32.

4. *Baker C.F., Fillmore C.J., Lowe J.B*. The Berkeley FrameNet project // Proceedings of the 17th International Conference on Computational linguistics. 1998, August. V. 1. Association for Computational Linguistics. P. 86–90.

5. *Fillmore C.J., Baker C.F.* Frame semantics for text understanding // Proceedings of WordNet and Other Lexical Resources Workshop. Pittsburgh, 2001. URL: http://www.ccs.neu.edu/course/csg224/resources/framenet/framenet.pdf.

6. *Fillmore C.J., Baker C.F., Sato H.* FrameNet as a “Net” // Proceedings of LREC, Lisbon, 2004. V. 4. P. 1091–1094.

7. *Baker C.F., Ruppenhofer J*. FrameNet's frames vs. Levin's verb classes // Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society. 2002. P. 27–38.

8. *Boas H.C.* (Ed.). Multilingual FrameNets in Computational Lexicography: Methods and Applications. 2009. V. 200. Walter de Gruyter. 352 p.

9. FrameBank. URL: http://framebank.ru.

10. *Kashkin E.V. Lyashevskaya O.N*. Semanticheskiye roli i set konstruktsiy v sisteme FrameBank // Kompyuternaya lingvistika i intellektualnyye tekhnologii. Po materialam ezhegodnoy mezhdunarodnoy konferentsii “Dialog” 2013. T. 1. S. 297–311.

11. *Kipper K., Korhonen A., Ryant N., Palmer M*. Extending VerbNet with novel

verb classes // Proceedings of the Fifth International Conference on Language Resources and Evaluation – LREC'06. May, 2006, Genoa, Italy: 2006. URL: http:// citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.380.5541&rep=rep1&type=pdf.

12. *Kipper K., Korhonen A., Ryant N., Palmer M*. A large-scale classification of English verbs// Language Resources and Evaluation. 2008. V. 42 (1). P. 21–40.

13. *Palmer M., Kingsbury P., Gildea D.* The proposition bank: an annotated corpus of semantic role // Computational Linguistics. 2005. V. 31 (1). P. 71–106.

14. *Loper E., Yi S.T., Palmer M.* Combining lexical resources: mapping between PropBank and VerbNet // Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg, the Netherlands. 2007. URL: http://verbs.colorado.edu/~kipper/Papers/semlink_iwcs7.pdf.

15. *Fellbaum C*. WordNet. An electronic lexical database. Cambridge, Mass: MIT Press, 1998. 423 p.

16. *Mille G.A*. WordNet: A lexical database for English // Communications of the ACM. 1995. V. 38, No 11. P. 39–41.

17. *Vossen P.* (Ed.) EuroWordNet general document. Version 3. URL: http://vossen.info/docs/2002/EWNGeneral.pdf.

18. *Azarova I.V., Mitrofanova O.A., Sinopalnikova A.A.* Kompyuternyy tezaurus russkogo yazyka tipa wordnet. URL: http://www.dialog-21.ru/ Archive/2003/ Azarova.htm.

19. *Lukashevich N.V.* Tezaurusy v zadachakh informatsionnogo poiska. M.: Izd-vo Mosk. un-ta, 2011. 511 s.

20. The Global WordNet Associaton. URL: http://globalwordnet.org.

СВЕДЕНИЯ ОБ АВТОРЕ

ГАЛИЕВА Альфия Макаримовна – ведущий научный сотрудник Научно-исследовательского института «Прикладная семиотика» Академии наук Республики Татарстан. Сфера научных интересов: семантика, грамматика, философия языка.

email: amgalieva@gmail.com

Alfiia GALIEVA – Senior Researcher of Research Institute of Applied Semiotics of the Academy of Sciences of Republic of Tatarstan. Research interests: semantics, grammar, philosophy of language.

email: amgalieva@gmail.com

Материал поступил в редакцию 15 марта 2016 года

УДК 81'322.2 + 81'322.3

АНАЛИТИЧЕСКИЙ ОБЗОР МЕТОДОВ РАЗРЕШЕНИЯ МОРФОЛОГИЧЕСКОЙ МНОГОЗНАЧНОСТИ

Р.Р. Гатауллин

Институт вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета
Институт прикладной семиотики Академии наук Республики Татарстан
ramil.gata@gmail.com

Аннотация

Проанализированы основные методы разрешения морфологической многозначности применительно к татарскому языку. Описано текущее состояние работ и приведены основные результаты по данному направлению, сделаны выводы о применимости методов разрешения с оценкой их точности.

***Ключевые слова**: разрешение морфологической многозначности, контекстные методы, статистико-вероятностные методы, татарский язык.*

ВВЕДЕНИЕ

Многозначность языковых форм – одна из природных особенностей естественного языка, способствующая качественному развитию словарного запаса, тем самым «экономящая» словесный материал [2]. Разрешение многозначности (т. н. дизамбигуация) является одной из важнейших задач автоматической обработки естественного языка. Результаты разрешения используются для повышения точности методов классификации и кластеризации текстов, улучшения качества машинного перевода, информационного поиска и других приложений [2].

Исследователи выделяют несколько типов многозначности естественного языка: *морфологическую, синтаксическую и лексико-семантическую многозначности*. Иногда к ним добавляют прагматическую многозначность. *Для* работы с каждым из этих типов существуют собственные методы [2].

Задача разрешения *морфологической многозначности* заключается в определении для слова части речи и грамматических признаков, соответствующих контексту. Морфологическая многозначность, в основном, представлена грамматической омонимией, т. е. совпадением слов в отдельных грамматических формах. Например, слово «стекло» в зависимости от контекста может быть либо существительным, обозначающим материал («смотреть через стекло»), либо глаголом в прошедшем времени 3-го лица единственного числа («масло стекло»).

Задача разрешения синтаксической многозначности (*многозначность синтаксических структур)* заключается в правильном определении функций синтаксических единиц предложения. Примером такой неоднозначности является предложение «мужу изменять нельзя» (словоформа *мужу* – субъект или объект предложения?) [2].

Значения слов могут относиться к одной части речи, но различаться по смыслу, например, «*platform*» – железнодорожная или компьютерная платформа. В этом случае речь идет о *полисемии*, когда у одного слова имеются два или более значения, взаимосвязанных по смыслу и происхождению. Полисемия относится к *лексической многозначности*. Сюда же следует относить и *лексическую омонимию (*слова совпадают в звучании и написании, но имеют разные значения). Такими омонимами являются слова *лук* («оружие») и *лук (*«растение»). Задача разрешения такой неоднозначности состоит в установлении значений слов или составных терминов в соответствии с контекстом, в котором они использовались [2].

Еще один тип неоднозначности возникает в результате употребления местоимений или специальных существительных типа *one, another* (еще один). Так, в предложении «*Она уронила карандаш на стол и сломала его*» невозможно однозначно определить, что именно было сломано – *карандаш* или *стол* (нельзя однозначно разрешить референцию местоимения *его*) [2]. В этом случае говорят о *прагматической неоднозначности*.

Сложность и особенности разрешения многозначности для каждого конкретного языка проявляются по-разному. Например, для английского языка с бедной морфологией и жестким порядком слов в предложении разрешение морфологической многозначности, как правило, сводится к задаче POS-теггинга (от

англ., part of speech – определение части речи слова) и решается применением достаточно простых методов. Для русского языка морфологическая многозначность не столь характерна, как для английского и татарского, но, тем не менее, присуща. Дополнительную сложность добавляет свободный порядок слов в русском языке. В татарском языке, как и в других агглютинативных языках, таких, как турецкий и венгерский, морфемы несут как семантическую, так и синтаксическую информацию. Имея теоретически неограниченное количество присоединяемых к основе морфем, морфологическая многозначность приобретает разнообразные формы, что значительно усложняет задачу разрешения.

КОНТЕКСТНЫЕ МЕТОДЫ

Эти задачи были поставлены еще в 1950–1960-х годах, и теоретические исследования имеют многолетнюю историю. Еще в конце 1950-х годов в работах K.E. Harper [5], A. Caplan [6] основным способом снятия омонимии признавались изучение и описание тех контекстных условий, в которых реализуется то или иное значение слова. При этом под контекстом понималось окружение слова в тексте, т. е. слова, с которыми данное слово употребляется.

Актуальным для исследуемой задачи также являлся вопрос о минимальном разрешающем контексте. В этой связи заслуживают внимания результаты, полученные A. Caplan [6] по исследованию минимального разрешающего контекста. В работе анализировались 140 многозначных употребительных английских слов (в основном, лексических омонимов), находившихся в различных контекстных условиях. Автором выделены следующие виды контекстов:

- сочетание с предшествующим словом – P1;
- сочетание с последующим словом – F1;
- сочетание с предшествующим и последующим *словами* – B1 (both);
- сочетание с двумя предшествующими словами – P2;
- сочетание с двумя последующими словами – F2;
- сочетание с двумя предшествующими и двумя последующими словами – B2;
- все предложение в целом – S (sentence).

Основной вывод заключался в том, что цепочка B1 по эффекту редуцирования многозначности (отношение количества значений слова в конкретном контексте к их количеству в нулевом контексте) более продуктивна, чем контекст, состоящий из двух предшествующих или двух последующих слов (P2 и F2), и приближается к эффекту, даваемому целым предложением (S) [6].

В другом выводе подчеркивается важное значение материального типа контекста, т. е. входят ли в непосредственное окружение знаменательные слова, или слова, называемые автором «particles» (предлоги, союзы, глаголы типа will или do, артикли, местоимения и наречия типа there и др.). Первый тип контекста дает значительно большую редукцию многозначности, чем контекст, содержащий слова без конкретного лексического наполнения [6, 7].

Общие выводы A. Caplan сводятся к тому, что наиболее практичным является контекст, состоящий из одного слова слева и одного слова справа от анализируемой многозначной лексемы. Если же одно из слов окружения – «particle», то следует «усилить» контекст до двух слов с обеих сторон [6, 7].

Исследования такого подхода для русского языка [7] показали, что его применимость в реальных контекстах вряд ли возможна. Реальная ситуация с разрешением омонимии в русском языке значительно сложнее и не может быть разрешена на основе упрощенных схем. В отличие от английского, в русском языке порядок слов свободный, предполагается, что количество возможных контекстов из-за этого увеличивается. Для решения этой проблемы для русского языка была предложена усложненная структура правил, а также предполагается в качестве контекста использовать все предложение [7]. С учетом этого замечания было разработано программное средство разрешения функциональной омонимии, которая для некоторых типов дает точность распознавания, равную 100% при тестировании не менее 100 примеров, в наихудших случаях – точность не менее 95% [7].

При исследовании омонимии в татарском языке в центре внимания были лексические омонимы. Тем не менее, есть несколько работ, посвященных и грамматической омонимии [9–11]. Но до настоящего времени специальные исследования и классификации грамматической омонимии практически не проводились [1].

В работе [12] приведены основные формально-грамматические модели словосочетаний в татарском языке (15 основных, 80 частных типов) с указанием главного и зависимого слов. Актуальной задачей является проверка возможности использования этих моделей в качестве основы для определения разрешающих контекстов. Определенная строгость агглютинативной синтаксической структуры позволяет рассчитывать на обнаружение четких контекстных ограничений [1].

Для разрешения морфологической многозначности на основе контекстных правил в татарском языке в НИИ «Прикладная семиотика» Академии наук Республики Татарстан создан программный инструментарий для разработки и тестирования контекстных правил. Первые результаты экспериментов по построению контекстных правил показали работоспособность метода, однако для окончательных выводов требуются дополнительные исследования [23].

Подход, основанный на правилах, является чрезвычайно трудоемким, требует проведения тщательной лингвистической экспертизы каждого типа омонимии. Полная классификация типов омоформ является прагматически неоправданной задачей, так как татарский язык относится к агглютинативным языкам, для которых количество присоединяемых к основе морфем теоретически не ограничено. Например, в указанном корпусе татарских текстов объемом более 21 млн. словоупотреблений число типов омоформ превышает 7000 [1]. Здесь под типом морфологической многозначности (т. н. типом омоформ) подразумевается комбинация возможных аффиксальных цепочек, соответствующих слову. Например, тип, состоящий из «N» (сущ.) и «V+Neg» (глагол в повелительном наклонении с отрицанием), приписан словам «*алма», «басма», «тартма*» и др.

Чрезмерная трудоемкость этого подхода требует поиска более оптимальных путей решения. Одним из направлений является попытка комбинирования данного подхода со статистико-вероятностными методами и методами машинного обучения.

СТАТИСТИКО-ВЕРОЯТНОСТНЫЕ МЕТОДЫ

В тех же 1950–1960-х годах вслед за контекстными методами для задач дизамбигуации стали использовать статистико-вероятностные методы. Отсутствие больших информационных ресурсов и языковых баз данных значительно осложняло эксперименты и их дальнейшее применение.

После появления репрезентативных электронных корпусов эксперименты с вероятностно-статистическими методами показали достаточно хорошие результаты. Например, для английского языка, как было отмечено, задача снятия морфологической омонимии сводится, как правило, к проблеме разрешения многозначности на уровне частей речи (так называемого POS-теггинга). При этом используются алгоритмы, основанные на статистических моделях, таких, как скрытая марковская модель HMM [13] и марковская модель максимальной энтропии MEMM [14], учитывающие вероятность появления тега той или иной части речи в данном контексте. Для английского языка эти алгоритмы дают приемлемый результат с точностью не менее 96% [15].

Среди известных методов, применяемых при снятии морфологической многозначности в текстах английского языка, следует также отметить метод опорных векторов (Support Vector Machines, SVM) и деревья решений (Decision Trees). Например, точность SVM составила 97.2% при тестировании на текстах новостных статей из корпуса The Wall Street Journal, что является достаточно хорошим результатом [16].

Статистические методы для разрешения морфологической омонимии применительно к русскому языку стали использоваться сравнительно недавно. Зеленков и др. [17] предложили алгоритм, предназначенный для разрешения морфологической омонимии слов, которые совпадают лишь в нескольких грамматических формах. Метод основан на использовании автоматически полученного словаря контекстов, выведенного из уже размеченных текстов [16].

При адаптации к русскому языку некоторых методов необходимо учесть некоторые особенности языка. Во-первых, морфологическая омонимия в русском языке в отличие от английского языка не сводится к частеречной омонимии, а охватывает большое количество различных грамматических признаков. Во-вторых, хорошая работа статистических моделей на материале английских текстов объясняется тем, что в английском языке существует фиксированный порядок слов. Это обстоятельство упрощает создание модели, так как позволяет, к примеру, опираться только на локальный контекст слова (соседние слова) без учета дальних зависимостей. Именно поэтому для морфологической дизамбигуации в

английском языке часто успешно используются алгоритмы, основанные на марковских моделях и учитывающие зависимость каждого набора тегов только от одного элемента контекста – непосредственно предшествующего ему набора тегов [15].

В русском языке, напротив, порядок слов свободный, так что предполагается, что количество возможных контекстов из-за этого увеличивается, и эффективность обучения простой модели, основанной на локальных зависимостях, снижается. Поэтому, наряду с марковскими моделями, для снятия морфологической омонимии в русском языке используются более сложные статистические модели или гибридные системы [15], в которых статистика дополняется набором правил (см., например, Transformation-Based Learning [18], а также [17]).

Алгоритм, основанный на использовании скрытой марковской модели (HMM), требует предварительного обучения системы на уже размеченной выборке текстов большого объема. Предварительные результаты экспериментов показали точность работы алгоритма для русского языка не менее 95% [15, 19].

В [15] отмечается, что при сравнительном анализе алгоритмов, основанных на скрытой марковской модели и марковской модели максимальной энтропии, оба алгоритма неплохо (точность не менее 95%) справляются с задачей частеречной дизамбигуации, но значительно хуже снимают омонимию по расширенному набору грамматических тегов. Как правило, алгоритмы ошибаются при разметке имен собственных, местоимений, римских цифр, инициалов и сокращений. Помимо этого, модели не работают со случаями субстантивации прилагательных и выбором некоторых падежных форм: в первую очередь, с разграничением между номинативом и аккузативом, что связано с особенностями порядка слов в русском языке. В заключении этой работы делается вывод, что алгоритм MEMM в целом работает лучше в применении к задаче POS-теггинга, чем HMM [15].

Для агглютинативных языков, таких, как венгерский [20] и финский [21], метод, основанный на HMM, также дает не менее 97% точности. В работе [24] утверждается о достижении 98% точности разрешения морфологической многозначности для турецкого языка при использовании HMM совместно с персептронным алгоритмом (англ, Perceptron Algorithm [25]).

В [16] проанализирована применимость метода опорных векторов для задач снятия многозначности. Основная идея SVM-метода заключается в поиске разделяющей гиперплоскости с максимальным зазором между векторами двух различных классов. Для нахождения разделяющей гиперплоскости потребуется уже размеченный набор текстов. Механизм метода опорных векторов довольно прост, и как показывает практика, эффективен. Гибкость алгоритма позволяет успешно сочетать его с уже существующими методами определения частей речи и снятия омонимии [16].

В работе [22] описан интересный подход с генерацией правил разрешения из размеченного корпуса со снятой многозначностью. Метод применялся для турецкого языка, эксперименты показали точность не менее 96%. Отличительной особенностью подхода является выявление контекстных ограничений не для всей аффиксальной цепочки в целом, а для каждой морфемы отдельно. Турецкому языку, как и татарскому, свойственна возможность теоретически неограниченно присоединять морфемы к основе, что приводит к многообразию форм слов, а это с свою очередь, – к разреженности данных при обучении. Данный подход в определенной мере способствует решению проблемы с разреженностью данных.

Проблема разреженности данных стоит и для татарского языка. На данный момент языковой корпус татарского языка находится на стадии разработки. Морфологическая разметка осуществляется автоматически адаптированным морфологическим анализатором на базе двухуровневой модели морфологии татарского языка [26]. Снятие морфологической многозначности выполняется экспертами вручную, поэтому объем корпуса со снятой многозначности незначителен. Поэтому экспериментальная проверка применимости всех описанных статистико-вероятностных методов для татарского языка в настоящее время не представляется возможной ввиду отсутствия размеченного корпуса. Тем не менее, типологическая и генетическая близость турецкого и татарского языка дает основание полагать, что статистические методы способны показать хорошие результаты для татарского языка.

Таким образом, текущими задачами являются подготовка татарского размеченного корпуса и применение описанных методов для решения морфологической многозначности. В первую очередь предполагается применять те методы,

которые показали хорошие результаты для близкородственных языков.

ЗАКЛЮЧЕНИЕ

В настоящей работе представлен аналитический обзор основных методов разрешения морфологической многозначности. Точность работы описанных методов составляет не ниже 95%. В основном методы являются языконезависимыми, но точность разрешения варьируется в зависимости от конкретного языка (см. табл. 1).

Таблица 1

Класс метода	**Методы**	**Язык**	**Точность**
Контекстные методы	-	английский	99,5% [19]
	-	русский	95% [7]
Статистико-вероятностные методы	HMM	английский	96% [15]
		русский	95% [15, 19]
		финский	97% [21]
		венгерский	97% [20]
		турецкий	98% [25]
	MEMM	английский	96% [14, 15]
		русский	95% [15]
	SVM	английский	97,2% [16]
		русский	95,7% [16]
	GPA	турецкий	96% [22]

Для английского языка, имеющего *бедную морфологию*, проблема разрешения морфологической многозначности, как правило, сводится к разрешению многозначности на уровне частей речи (POS-теггинг), что, в свою очередь, заметно

облегчает задачу. В агглютинативных языках, таких, как турецкий, венгерский и татарский, к основе слова присоединяются морфемы, которые, кроме семантики, определяют и синтаксические связи. Морфологическая многозначность в этих языках проявляется разнообразными формами. В некоторых случаях для разрешения морфологической многозначности могут потребоваться как синтаксический, так и семантический анализ.

С другой стороны, жесткий порядок слов в предложениях на английском языке позволяет использовать минимальный размер контекста, тогда как для русского языка иногда требуется в качестве контекста использовать все предложение [7], тем самым усложняя задачу поиска разрешающего контекста. Размер минимального контекста для татарского языка еще предстоит исследовать. Тем не менее, есть основания полагать, что определенная строгость синтаксическая структуры позволит рассчитывать на обнаружение четких контекстных ограничений в ближайшем контексте [1].

Применение статистических алгоритмов для снятия многозначности позволило сместить акценты разработки на подготовку размеченных корпусов для обучения статистико-вероятностных моделей.

Несмотря на указанные сложности, можно констатировать, что для английского, русского и турецкого языков проблема разрешения морфологической многозначности, в основном, решена. Используя различные надстройки над алгоритмами (либо увеличивая обучающую выборку для статистических методов), точность методов разрешения можно довести до уровня не ниже 97%.

Типологическая и генетическая близость турецкого и татарского языка дает основание полагать, что данные методы способны дать приемлемые результаты и для татарского языка.

СПИСОК ЛИТЕРАТУРЫ

1. *Хакимов Б.Э., Гильмуллин Р.А., Гатауллин Р.Р.* Разрешение грамматической многозначности в корпусе татарского языка // Учен. зап. Казан. ун-та. Сер. Гуманит. науки. 2014. Т. 156, кн. 5. С. 236–244.

2. *Турдаков Д.Ю.* Методы и программные средства разрешения лексической многозначности терминов на основе сетей документов: автореф. дис. … канд. тех. наук: 05.13.11. Москва, 2010. 20 с.

3. *Бобичев В.Л.* Автоматическое снятие морфологической многозначности при разметке корпуса // Тр. междунар. конф. «Корпусная лингвистика–2008». СПб.: СПбГУ, 2008. С. 45–49.

4. *Tufiş D., Popescu O.A.* Knowledge-based approach to morpho-lexical processing of natural language // in Proceedings of the International Conference for Young Computer Scientists, Beijing, 1991. P. 405–408.

5. *Harper K.E.* Contextual analysis // Mech. Translation. 1956. V. 4, No 3. P. 70–75.

6. *Caplan A.* An experimental study of ambiguity and context // Mech. Translation. 1955. V. 2, No 2. P. 39–46.

7. *Зинькина Ю.В., Пяткин Н.В., Невзорова О.А.* Разрешение функциональной омонимии в русском языке на основе контекстных правил // Труды межд. конф. Диалог'2005. М.: Наука, 2005. С. 198–202.

8. *Кобзарева Т.Ю., Афанасьев Р.Н.* Универсальный модуль предсинтаксического анализа омонимии частей речи в РЯ на основе словаря диагностических ситуаций // Труды междунар. конференции Диалог'2002. М.: 2002. С. 258–268.

9. *Курбатов Х.Р.* Грамматические омонимы в татарском языке // Татар теле һәм әдәбияты. Казан: Татар. кит. нәшр., 1959. Б. 307–311.

10. *Салахова Р.Р.* Омонимичные суффиксы татарского языка. Казань: Gumanitarya, 2007. 204 с.

11. *Салимгараева Б.С.* Омонимы в современном татарском языке. Автореф. канд. дис. Уфа, 1971. 82 с.

12. Татарская грамматика. Казань: Татар. книж. изд-во, 1993. Т. II. Морфология. 397 с.

13. *Weischedel Ralph M.* Coping with ambiguity and unknown words through probabilistic models // Computational Linguistics. Cambridge, MA, USA: MIT Press, 1993. V. 19, Issue 2. P. 361–382.

14. *Ratnaparkhi A.* Maximum entropy model for part-of-speech tagging // Proceedings of the Empirical Methods in Natural Language Processing. Philadelphia, PA, USA, 1996. P. 133–142.

15. *Лакомкин Е.Д., Пузыревский И.В., Рыжова Д.А.* Анализ статистических алгоритмов снятия морфологической омонимии в русском языке. URL: http://aist-conf.org/stuff/aist2013/submissions/aist2013_submission_33.pdf.

16. *Ткаченко М.В.* Модель и алгоритм улучшения распознавания частей речи в текстах, содержащих ошибки. СпбГУ, 2010. 20 с. URL: http://se.math.spbu.ru/SE/YearlyProjects/2010/list.

17. *Зеленков Ю.Г., Сегалович И.В., Титов В.А.* Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара Диалог'2005. М.: Наука, 2005. С. 616.

18. *Brill E.* A simple rule-based part of speech tagger // Proceedings of the third conference on Applied natural language processing (ANLC’92). Stroudsburg, PA, USA: Association for Computational Linguistics, 1992. P. 152–155.

19. *Сокирко А.В., Толдова С.Ю.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп). URL: http://www.aot.ru/docs/RusCorporaHMM.htm.

20. *Orosz G., Novak A.* PurePos 2.0: a hybrid tool for morphological disambiguation // In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria, 2013. INCOMA Ltd. Shoumen, Bulgaria. P. 539–545.

21. *Kristen Linden, Tommi Pirinen.* Weighted finite-state morphological analysis of finnish compounding with HFST-LEXC // In Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. Editors: Kristiina Jokinen and Eckhard Bick. NEALT Proceedings Series, 2009. V. 4. P. 89–95.

22. *Deniz Yuret, Ferhan Ture.* Learning morphological disambiguation rules for turkish // Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. New York, 2006. P. 328–334.

23. *Гатауллин Р. Р., Гильмуллин Р.А.* Контекстные правила для разрешения морфологической многозначности в корпусе татарского языка // Открытые семантические технологии проектирования интеллектуальных систем OSTIS-2016 (OpenSemantic Technologies for Intelligent Systems). Материалы V международной научно-технической конференции (Минск, 18–20 февраля 2016 года). Минск: БГУИР, 2016. С. 389–392.

24. *Hasim Sak, Tunga Gongur, Murat Saraclar*. Morphological disambiguation of turkish text with perceptron algorithm // Computational Linguistics and Intelligent Text Processing, 8th International Conference CICLing, Mexico City, Mexico, February 2007. P. 107–118.

25. *Collins M.* Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms // Proceedings of EMNLP, 2002. P. 1–8.

26. *Сулейманов Д.Ш., Гильмуллин Р.А.* Двухуровневое описание морфологии татарского языка // Тезисы докладов Международной научной конференции «Языковая семантика и образ мира». Казань: Изд-во Казан. гос. ун-та, 1997. Книга 2. С. 65–67.

REVIEW OF MORPHOLOGICAL DISAMBIGUATION METHODS

R.R. Gataullin

Institute of Computational Mathematics and Information Technologies.
Kazan Federal University
Institute of Applied Semiotics of the Tatarstan Academy of Sciences
ramil.gata@gmail.com

Abstract

This paper describes the morphological disambiguation methods and their application for the Tatar language. The state-of-the art technology is discussed. We analyze the contextual and statistical methods and their evaluations for different languages.

***Keywords**: morphological disambiguation, contextual method, statistical method, Tatar language*

REFERENCES

1. *Khakimov B.E., Gilmullin R.A., Gataullin R.R.* Razresheniye grammaticheskoy mnogoznachnosti v korpuse tatarskogo yazyka // Uchen. zap. Kazan. un-ta. Ser. Gumanit. nauki. 2014. T. 156. kn. 5. S. 236–244.

2. *Turdakov D.Yu.* Metody i programmnyye sredstva razresheniya leksicheskoy mnogoznachnosti terminov na osnove setey dokumentov: avtoref. dis. … kand. tekh. nauk. 05.13.11. Moskva, 2010. 20 s.

3. *Bobichev V.L.* Avtomaticheskoye snyatiye morfologicheskoy mnogoznachnosti pri razmetke korpusa // Tr. mezhdunar. konf. «Korpusnaya lingvistika–2008». SPb.: SPbGU, 2008. S. 45–49.

4. *Tufiş D., Popescu O.A.* Knowledge-based approach to morpho-lexical processing of natural language // In Proceedings of the International Conference for Young Computer Scientists, Beijing, 1991. P. 405–408.

5. *Harper K.E.* Contextual analysis // Mech. Translation. 1956. V. 4, No 3. P. 70–75.

6. *Caplan A.* An experimental study of ambiguity and context // Mech. Translation. 1955. V. 2, No 2. P. 39–46.

7. *Zinkina Yu.V., Pyatkin N.V., Nevzorova O.A.* Razresheniye funktsionalnoy omonimii v russkom yazyke na osnove kontekstnykh pravil // Trudy mezhd. konf. Dialog'2005. M.: Nauka. 2005. S. 198–202.

8. *Kobzareva T.Yu., Afanasyev R.N.* Universalnyy modul predsintaksicheskogo analiza omonimii chastey rechi v RYa na osnove slovarya diagnosticheskikh situatsiy // Trudy mezhdunar. konferentsii Dialog'2002. M., 2002. S. 258–268.

9. *Kurbatov Kh.R.* Grammaticheskiye omonimy v tatarskom yazyke // Tatar tele hem edebiyaty. Kazan: Tatar. kit. neshr. 1959. B. 307–311.

10. *Salakhova R.R.* Omonimichnyye suffiksy tatarskogo yazyka. Kazan: Gumanitarya, 2007. 204 s.

11. *Salimgarayeva B.S.* Omonimy v sovremennom tatarskom yazyke. Avtoref. kand. dis. Ufa, 1971. 82 s.

12. Tatarskaya grammatika. Kazan: Tatar. knizh. izd-vo, 1993. T. II. Morfologiya. 397 s.

13. *Weischedel Ralph M.* Coping with ambiguity and unknown words through probabilistic models // Computational Linguistics. Cambridge, MA, USA: MIT Press, 1993. V. 19, Issue 2. P. 361–382.

14. *Ratnaparkhi A.* Maximum entropy model for part-of-speech tagging // Proceedings of the Empirical Methods in Natural Language Processing. Philadelphia, PA, USA, 1996. P. 133–142.

15. *Lakomkin E.D., Puzyrevskiy I.V., Ryzhova D.A.* Analiz statisticheskikh algoritmov snyatiya morfologicheskoy omonimii v russkom yazyke. URL: http://aistconf.org/stuff/aist2013/submissions/aist2013_submission_33.pdf.

16. *Tkachenko M.V.* Model i algoritm uluchsheniya raspoznavaniya chastey rechi v tekstakh soderzhashchikh oshibki. SpbGU, 2010. 20 s. URL: http://se.math.spbu.ru/SE/YearlyProjects/2010/list.

17. *Zelenkov Yu.G., Segalovich I.V., Titov V.A.* Veroyatnostnaya model snyatiya morfologicheskoy omonimii na osnove normalizuyushchikh podstanovok i pozitsiy sosednikh slov // Kompyuternaya lingvistika i intellektualnyye tekhnologii. Trudy mezhdunarodnogo seminara Dialog'2005. Kazan, 2005. C. 616.

18. *Brill E.* A simple rule-based part of speech tagger // Proceedings of the third conference on Applied natural language processing (ANLC’92). Stroudsburg, PA, USA: Association for Computational Linguistics, 1992. P. 152–155.

19. *Sokirko A.V., Toldova S.Yu.* Sravneniye effektivnosti dvukh metodik snyatiya leksicheskoy i morfologicheskoy neodnoznachnosti dlya russkogo yazyka (skrytaya model Markova i sintaksicheskiy analizator imennykh grupp). URL: http://www.aot.ru/docs/RusCorporaHMM.htm.

20. *Orosz G., Novak A.* PurePos 2.0: a hybrid tool for morphological disambiguation // In Proceedings of the International Conference on Recent Advances in Natural

Language Processing (RANLP 2013), Hissar, Bulgaria, 2013. INCOMA Ltd. Shoumen, Bulgaria. P. 539–545.

21. *Kristen Linden, Tommi Pirinen.* Weighted finite-state morphological analysis of finnish compounding with HFST-LEXC // In Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. Editors: Kristiina Jokinen and Eckhard Bick. NEALT Proceedings Series, 2009. V. 4. P. 89–95.

22. *Deniz Yuret, Ferhan Ture*. Learning morphological disambiguation rules for turkish // Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. New York, 2006. P. 328–334.

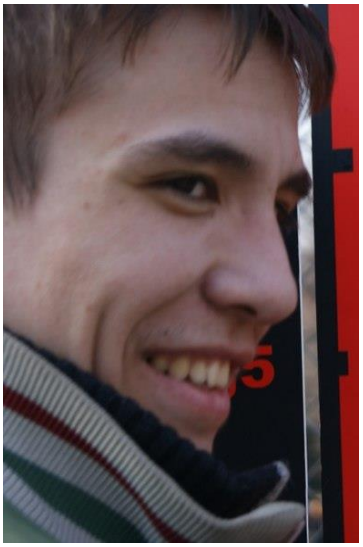
23. *Gataullin R.R., Gilmullin R.A.* Kontekstnyye pravila dlya razresheniya morfologicheskoy mnogoznachnosti v korpuse tatarskogo yazyka // Otkrytyye semanticheskiye tekhnologii proyektirovaniya intellektualnykh sistem OSTIS-2016 (OpenSemantic Technologies for Intelligent Systems). Materialy V mezhdunarodnoy nauchno-tekhnicheskoy konferentsii (Minsk, 18–20 fevralya 2016 goda). Minsk: BGUIR, 2016. S. 389–392.

24. *Hasim Sak, Tunga Gongur, Murat Saraclar*. Morphological disambiguation of turkish text with perceptron algorithm // Computational Linguistics and Intelligent Text Processing, 8th International Conference CICLing, Mexico City, Mexico, February 2007. P. 107–118.

25. *Collins M.* Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms // Proceedings of EMNLP, 2002. P. 1–8.

26. *Suleymanov D.Sh., Gilmullin R.A.* Dvukhurovnevoye opisaniye morfologii tatarskogo yazyka // Tezisy dokl. Mezhdunarodnoy nauchnoy konferentsii "Yazykovaya semantika i obraz mira". Kazan: Izd-vo Kazan. gos. un-ta, 1997. Kniga 2. S. 65–67.

СВЕДЕНИЯ ОБ АВТОРЕ



ГАТАУЛЛИН Рамиль Раисович – аспирант Института вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета.

Ramil Raisovich GATAULLIN, received MS degree in Mathematics from Kazan Federal University (2012). Currently is a graduate student at the Institute of Computational Mathematics and Information Technologies of Kazan Federal University. Current scientific interests: natural language processing, data mining, knowledge extraction technologies.

email: ramil.gata@gmail.com

Материал поступил в редакцию 18 марта 2016 года

УДК 004.5

ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА СОЗДАНИЯ ЭЛЕКТРОННЫХ ВЕРСИЙ ОБУЧАЮЩИХ МАТЕРИАЛОВ

А.Ф. Хусаинов[1], А.Х. Хусаинова[2], Р.А. Гильмуллин[3]

[1,2,3] *Казанский (Приволжский) федеральный университет*

[1,3] *Институт прикладной семиотики Академии наук Республики Татарстан*

[1] khusainov.aidar@gmail.com, [2] alfirahamzovna@gmail.com,

[3] rinatgilmullin@gmail.com@gmail.com

Аннотация

Описана технология, которая позволяет создавать электронные версии обучающих материалов. Данные материалы представляют собой часть общей образовательной среды, построенной на основе принципов Smart Education – современного метода обучения, базирующегося на облачных технологиях и обеспечивающего интерактивность учебного процесса. В электронных обучающих материалах полностью представлено содержимое печатного учебника, необходимых дополнительных интерактивных материалов; созданные с помощью набора алгоритмов электронные учебники могут быть интегрированы в учебный процесс как в виде интернет-ресурса, так и в виде мобильных приложений для наиболее популярных ОС.

Ключевые слова*: электронный учебник, кроссплатформенность, образовательная среда, Smart Education*

ВВЕДЕНИЕ

В рамках компетентностного подхода к образованию важно построить обучение студентов таким образом, чтобы выпускник стал обладателем профессиональных компетенций, адаптированных к быстро меняющейся информационной среде, носителем актуальных знаний и прикладных навыков.

Развитие электронного образования можно условно разделить на этапы

(рис. 1), сроки прохождения которых зависят от национальной системы образования. На смену прежним технологиям электронного обучения пришло «умное», smart-обучение [1].

	ИКТ в обучении	Электронное обучение	Повсеместное обучение	SMART-Обучение
Образовательные платформы	Компьютеризированного обучения	Система управления обучением (LMS), на основе веб	Мобильные LMS	Платформа для гибкого обучения
Технологии	CD, Дискеты	Интернет, локальная сеть, электронные учебные материалы	Электронные книги, мобильный контент, дополненная реальность	Открытые образовательные ресурсы, интеллектуальные системы
Оборудование	Настольные ПК	ПК, проводной Интернет	Ноутбуки, планшетные компьютеры, беспроводной Интернет	Смартфоны, планшеты, устройства для чтения, IP-телевидение
Период	С 1995	С 2000	С 2005	С 2010

Рис. 1. Этапы развития электронного образования

Smart Education («умное обучение») – это обучение в интерактивной образовательной среде с наличием свободного доступа к источникам информации, находящимся в свободном доступе; обучение, легко адаптируемое под потребности каждого студента. Преимущества использования smart-технологий заключаются в обеспечении доступности образования и максимальной индивидуальности траектории обучения для каждого обучаемого.

Основной задачей преподавателя при этом становится задача организации и управления учебным процессом. 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 smart-обучения необходимы быстрый доступ к интернету и устройство для просмотра информации (компьютер, ноутбук, планшет, смартфон и др.).

Все разнообразие доступных ресурсов может быть объединено на основе какой-либо платформы, выбор которой зависит от потребностей и предпочтений организатора учебного процесса. На выбранной платформе создается образовательная среда учебной дисциплины (рис. 2).

Рис. 2. Структура smart-образовательной среды учебной дисциплины

В выбранной архитектуре образовательной среды одной из основных составляющих является smart-учебник [2]. На него возлагается функциональность по обеспечению обучающихся необходимым образовательным контентом, а также интерактивными элементами, предоставляющими доступ к избыточному количеству источников информации, позволяя студенту выбрать наиболее доступный и привлекательный контент.

Помимо учебного контента важное место в smart-образовательной среде уделяется практическим 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 выбор более сложного уровня через балльную систему оценок, соревновательный момент или совместную деятельность вместе с преподавателем и другими студентами. Данные возможности могут быть реализованы как в виде отдельного модуля, так и встроены в существующие smart-учебники.

Контроль и самоконтроль результатов обучения легко организуются через набор тестовых заданий.

ОСНОВНЫЕ ХАРАКТЕРИСТИКИ ИНТЕРАКТИВНОГО ЭЛЕКТРОННОГО УЧЕБНИКА

Основываясь на smart-подходе к образовательной деятельности, сформулируем основные требования, предъявляемые к электронным формам учебных материалов. Для ускорения процесса создания электронных учебников разработан специализированный набор программных инструментов.

Определим набор формальных требований к электронной форме учебников:

- по содержанию, структуре и художественному оформлению должны соответствовать печатной форме;
- в полном объеме содержать иллюстрации, имеющиеся в печатной форме;
- содержать мультимедийные и интерактивные элементы;
- содержать средства контроля и самоконтроля;
- быть представлены в общедоступных форматах, не имеющих лицензионных ограничений для пользователя;
- иметь удобный и интуитивно понятный интерфейс;
- предоставлять возможность работы в офлайн-режиме;
- иметь номера страниц, соответствующие номерам страниц печатной версии учебника.

Разработанный электронный учебник доступен для использования на компьютерах со следующими операционными системами:

- планшетные компьютеры: Android версии 4.0 и выше; iOS версии 7 и выше; Windows Phone версии 8.1 и выше;
- стационарные и переносные компьютеры: Windows версии XP и выше.

ИНТЕРФЕЙС ЭЛЕКТРОННОГО УЧЕБНИКА

Интерфейс приложения показан на рис. 3, где:

1 – редактируемое поле с текущим номером страницы;

2/3/4 – кнопки перехода на предыдущую/следующие страницу;

5 – кнопка отображения меню, состоящего из эскизов страниц учебника;

6 – кнопка отображения интерактивного оглавления;

7 – кнопка отображения интерактивного меню;

8 – кнопка для поиска текста в учебнике;

9 – кнопка для перехода в полноэкранный режим;

10 – кнопка работы с закладками;

11 – кнопка работы с заметками;

12 – кнопка просмотра информации об используемых в электронном учебнике иконках.

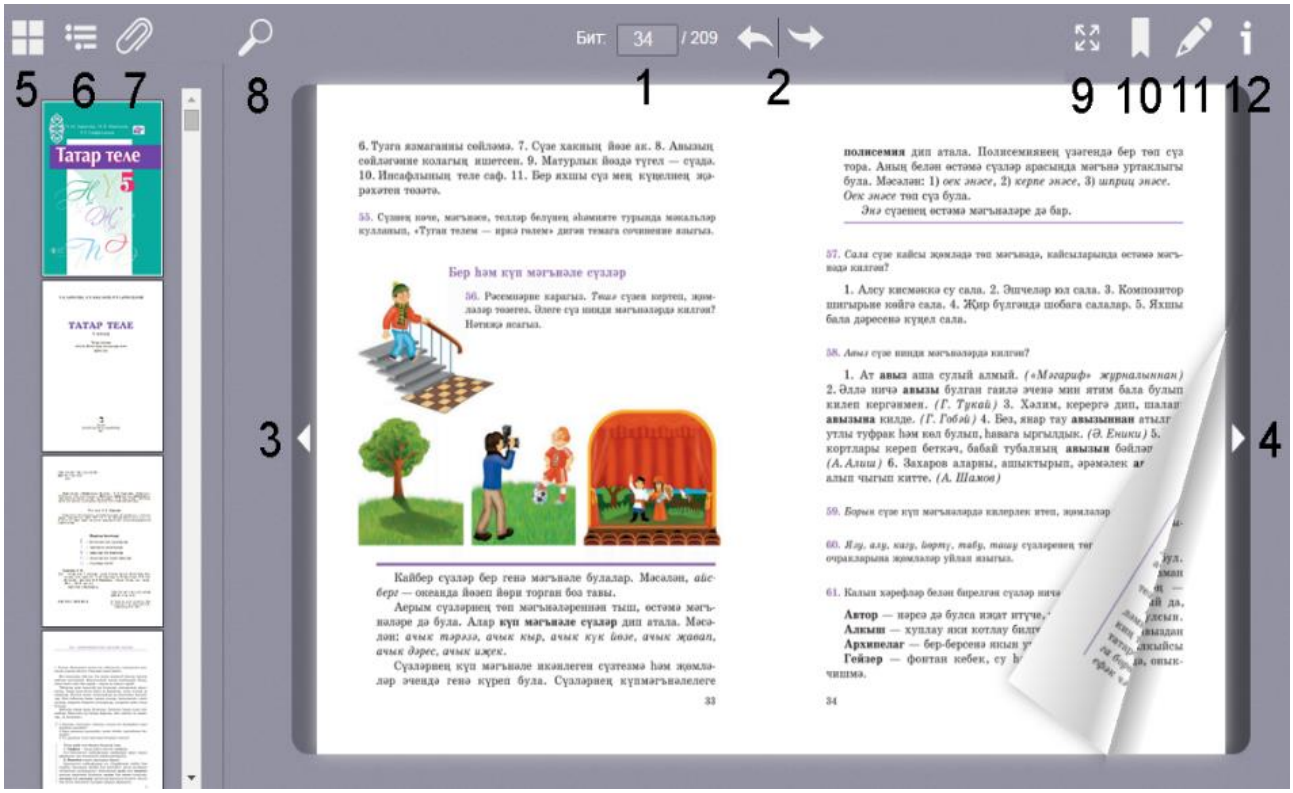


Рис. 3. Интерфейс приложения

Электронный учебник содержит следующие мультимедийные и интерактивные элементы:

Элемент	Иконка в электронном учебнике
аудиофрагменты	
изображения	
тест-тренажёр	
тест-контроль	
это надо знать	
это интересно	
задания	
закладка	
заметка	

ПРОЦЕСС СОЗДАНИЯ ИНТЕРАКТИВНОГО ЭЛЕКТРОННОГО УЧЕБНИКА

Процесс создания электронного учебника задействует 4 основных инструмента (JPDF2HTML5, Turn.js, PhoneGap, собственный инструмент), а также множество вспомогательных алгоритмов. Большинство необходимых процедур было автоматизировано, однако некоторые этапы требуют ручной работы редакторов (например, перенос информации о содержании учебника).

Исходными данными при создании электронных учебников являются следующие материалы:

- электронная версия учебника в pdf-формате;
- необходимые тестовые материалы в текстовом формате;

- набор вспомогательных элементов в текстовом формате (изображения, дополнительная информация, задания, заметки, закладки).

Общий алгоритм работы по созданию электронного учебника представлен на рис. 4. Возможность перелистывать страницы учебника, проводя мышкой (для стационарных устройств) или пальцем (для планшетных компьютеров) по краю или углам страниц, обеспечивается библиотекой Turn.js [3].

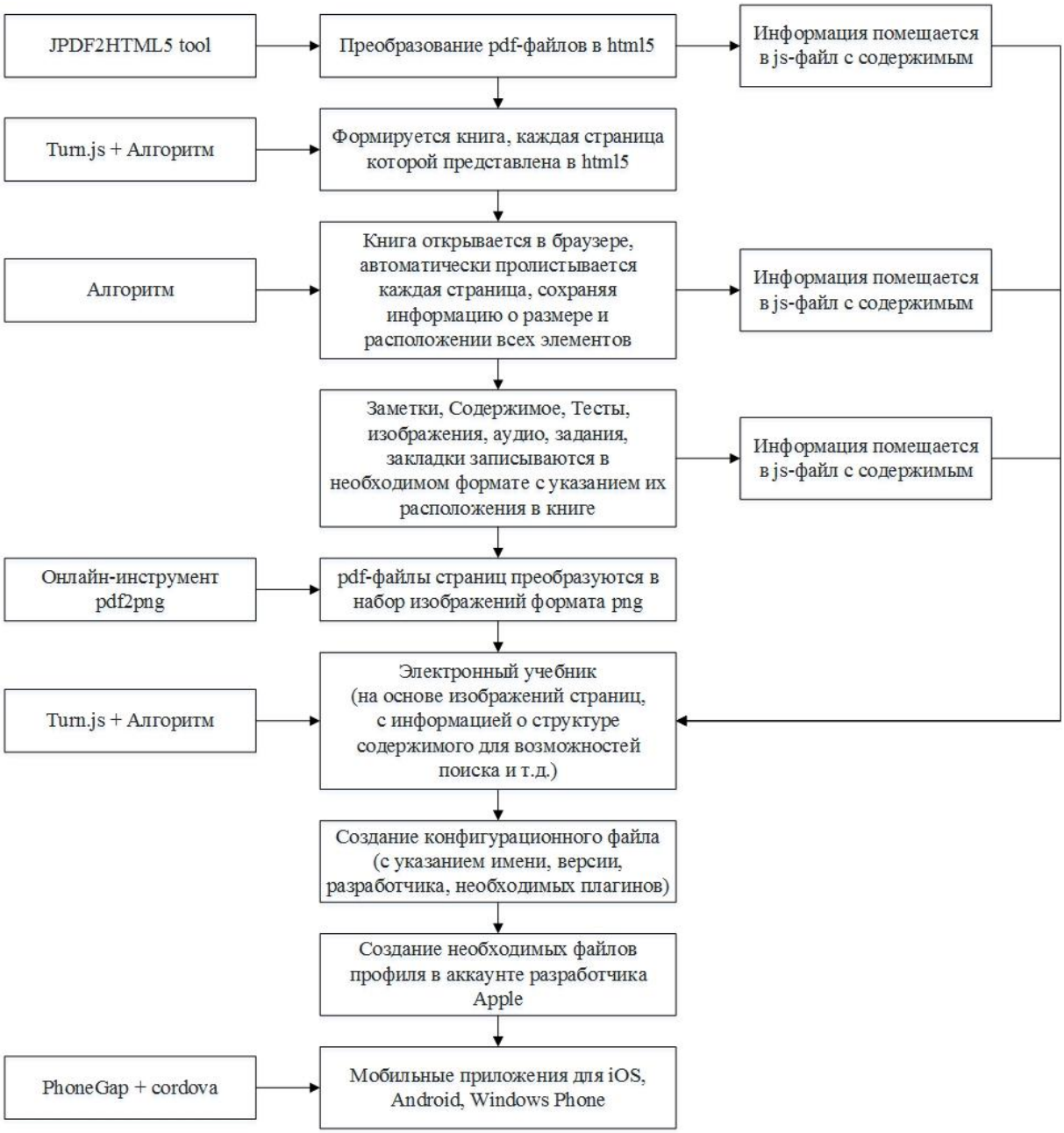


Рис. 4. Алгоритм создания электронного учебника

Изначально предполагалось использовать один из двух подходов: когда каждая страница представляет собой либо изображения, либо содержимое в формате html5. Преимущество первого подхода заключается в скорости загрузки, отображения и работы с электронным учебником, второй подход на распространённых устройствах показывал заметные зависания при подгрузке. Однако подход на основе изображений не обеспечивал возможность текстового поиска по учебнику, копирования или выделения отдельных фрагментов для создания заметок, подсветки важных фрагментов и т. д. Таким образом, было решено совместить два подхода: при отображении на экран используются изображения страниц, однако в отдельном файле хранится заранее созданный алгоритмом список с размерами и расположением всех элементов страницы. Для создания такого списка версия на основе html5-контента открывается в браузере. Алгоритм производит автоматическое пролистывание всех страниц и сохраняет позиции всех элементов, определённые браузером.

Важной характеристикой электронных учебников является возможность их использования на большинстве мобильных устройств в виде native-приложений, не требующих подключения к интернету или наличия вспомогательных программ для просмотра. Такая возможность достигается за счет использования технологий PhoneGap [4] и Cordova [5]. При разработке удалось адаптировать алгоритмы для корректной работы учебника, включая проигрывание мультимедиа-файлов, на всех мобильных устройствах. Дополнительно предоставляется возможность удобного масштабирования с помощью привычных жестов (double tap, pinch-to-zoom).

ВНЕДРЕНИЕ

Технология создания электронных учебников была использована ГУП «Татарское книжное издательство» при создании двух серий учебников по татарскому языку: для школ с русским и татарским языками обучения (учебники для 5–9 классов). Согласно требованиям, предъявляемым к формированию федерального перечня учебников, созданные электронные учебники прошли общественную и педагогическую экспертизы. Педагогическую экспертизу проводили Российская академия образования и Российская академия наук, а общественную экспертизу – ряд некоммерческих организаций, среди которых Российский книжный

союз, Русская школьная библиотечная ассоциация, НП «Лига образования».

ЗАКЛЮЧЕНИЕ

Разработанная технология позволяет в короткие сроки создавать электронные версии обучающих материалов. Данные материалы могут быть интегрированы в общую образовательную среду, построенную на основе принципов Smart-обучения.

Электронные учебники создаются на основе электронного представления содержимого учебника с введением дополнительных интерактивных элементов. Разработанные электронные учебники могут быть интегрированы в учебный процесс как в виде интернет-ресурса, так и в виде мобильных приложений для наиболее популярных ОС.

СПИСОК ЛИТЕРАТУРЫ

1. Smart учебное пособие по математике для высшей школы. URL: https://sites.google.com/site/ucebnyj123455/zadanie-3.

2. Smart-учебники в smart-образовании. Новая парадигма контента. URL: http://www.slideshare.net/pnevostruev/smart-congress.

3. Make a flipbook with HTML5. URL: http://www.turnjs.com/.

4. Adobe. Build amazing mobile apps powered by open web tech. URL: http://phonegap.com/.

5. Apache. Mobile apps with HTML, CSS & JS. URL: https://cordova.apache.org/.

TOOL FOR CREATING ELECTRONIC EDUCATIONAL MATERIALS

A.F. Khusainov[1], A.H. Khusainova[2], R.A. Gilmullin[3]

[1,2,3]*Kazan Federal University*

[2,3] *Institute of Applied Semiotics of the Tatarstan Academy of Sciences*

[1] khusainov.aidar@gmail.com, [2] alfirahamzovna@gmail.com, [3] rinat-gilmullin@gmail.com@gmail.com

Abstract

A technology for creation of electronic educational materials is described. It allows to use the result electronic textbook as a part of smart-education environment. Developed algorithms make it possible to work with all kind of devices from PC to tablets without any limitations in functionality.

Keywords*: electronic textbook, cross-platform, smart education environment*

REFERENCES

1. Smart uchebnoe posobie po matematike dlya vysshey shkoly. Smart-uchebniki v smart-obrazovanii. Novaya paradigm kontenta. URL: http://www.slideshare.net/pnevostruev/smart-congress.

2. Make a flipbook with HTML5. URL: http://www.turnjs.com/.

3. Adobe. Build amazing mobile apps powered by open web tech. URL: http://phonegap.com/.

4. Apache. Mobile apps with HTML, CSS & JS. URL: https://cordova.apache.org/.

СВЕДЕНИЯ ОБ АВТОРАХ

ХУСАИНОВ Айдар Фаилович – кандидат технических наук, старший научный сотрудник НИИ «Прикладная семиотика» Академии наук Республики Татарстан.

Aidar Failovich KHUSAINOV, Ph.D. in Technical Sciences. Senior researcher, Institute of Applied Semiotics of the Tatarstan Academy of Sciences. Current scientific interests: speech recognition, speech synthesis.

E-mail: khusainov.aidar@gmail.com

ГИЛЬМУЛЛИН Ринат Абрекович – кандидат физико-математических наук, зам. директора НИИ «Прикладная семиотика» Академии наук Республики Татарстан.

Rinat Abrekovich GILMULLIN – Ph.D. of Physics and Mathematics. Deputy Director of the Institute of Applied Semiotics of the Tatarstan Academy of Sciences. Current scientific interests: computer linguistics.

E-mail: rinatgilmullin@gmail.com

ХУСАИНОВА Альфира Хамзовна – старший преподаватель Института филологии и межкультурной коммуникации Казанского (Приволжского) федерального университета.

Alfira Hamzovna Khusainova – Senior lecturer at Kazan Federal University. Current scientific interests: e-learning.

E-mail: alfirahamzovna@gmail.com

Материал поступил в редакцию 20 марта 2016 года