V МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ ПО КОМПЬЮТЕРНОЙ ОБРАБОТКЕ ТЮРКСКИХ ЯЗЫКОВ «TURKLANG 2017»

Труды конференции

Том 1

КАЗАНЬ
2017

УДК 004.8+81’32
ББК 81.1

Организаторы:

Академия наук Республики Татарстан
Институт прикладной семиотики

Казанский (Приволжский) федеральный университет
Высшая школа информационных технологий
и информационных систем
Институт вычислительной математики
и информационных технологий

Евразийский национальный университет имени Л.Н. Гумилёва Министерства образования и науки Республики Казахстан
НИИ «Искусственный интеллект»

Международная Тюркская академия

Российская ассоциация искусственного интеллекта

Издание осуществлено при финансовой поддержке
Российского фонда фундаментальных исследований
(проект №17-47-161033)

Научные редакторы:
академик АН РТ, профессор, д.т.н. Д.Ш. Сулейманов,
к.т.н. А.Р. Гатиатуллин

Пятая Международная конференция по компьютерной обработке тюркских языков «TurkLang 2017». – Труды конференции. В 2-х томах. Т 1. – Казань: Издательство Академии наук Республики Татарстан, 2017. – 300 с.
ISBN 978-5-9690-0406-1

Сборник содержит материалы Пятой Международной конференции по компьютерной обработке тюркских языков «TurkLang-2017» (Казань, Татарстан, Россия, 18–21 октября 2017 г.)

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

УДК 004.8+81’32
ББК 81.1

ISBN 978-5-9690-0406-1

ПРЕДИСЛОВИЕ

В сборник материалов включены статьи участников Пятой Международной конференции по компьютерной обработке тюркских языков «TurkLang 2017» (Казань, Татарстан, Россия, 18–21 октября 2017 г.). Подготовка и издание сборника осуществлено при финансовой поддержке Российского фонда фундаментальных исследований, проект № 17-47-161033.

Участниками конференции, учеными и специалистами из России (Академия наук Республики Татарстан (Казань, Татарстан), Казанский федеральный университет, Московский государственный университета имени Ломоносова (Москва), Институт языкознания Российской академии наук (Москва), Высшая школа экономики (Москва), Институт проблем информатики Российской академии наук (Москва), Санкт-Петербургский государственный университет (Санкт-Петербур), Новосибирский государственный университет (Новосибирск), Институт истории, языка и литературы Уфимского научного центра Российской академии наук (Уфа, Башкортостан), Северо-Восточный федеральный университет имени М.К. Аммосова (Якутск, Саха), Чувашский государственный университет имени И.Н. Ульянова (Чебоксары, Чувашия), Тувинский государственный университет (Кызыл, Тува), Крымский федеральный университет имени В.И Вернадского (Симферополь, Крым) и др.), Синьцзянский университет (Урумчи, Китай), Университет Цинхуа (Бейджин, Китай), Бакинский Евразийский университет (Баку, Азербайджан), Оксбридж академия (Баку, Азербайджан), Евразийский национальный университет имени Л.Н. Гумилева (Астана, Казахстан), Назарбаев Университет (Астана, Казахстан), Казахский национальный университет имени аль-Фараби (Алматы, Казахстан), Кыргызский технический университет имени И. Раззакова (Бишкек, Кыргызстан), Институт теоретической и прикладной математики Национальной академии наук Кыргызстана (Бишкек, Кыргызстан), Кыргызский государственный университет строительства, транспорта и архитектуры имени Н.Исанова (Бишкек, Кыргызстан), Бишкекский государственный университет имени Карасаева (Бишкек, Кыргызстан), Ошский технологический университет (Ош, Кыргызстан), Университет Витаутаса Великого (Каунас, Литва), Стамбульский технический университет (Стамбул, Турция), Университет узбекского языка и литературы имени Алишера Навои (Ташкент, Узбекистан), Аризонский Университет (Тусон, Аризона, США), Суортмор колледж (Суортмор, Пенсильвания, США) были представлены доклады, посвященные актуальным проблемам компьютерной и когнитивной лингвистики для тюркских языков.

Активно и плодотворно обсуждались вопросы разработки формальных лингвистических моделей, электронных корпусов, систем машинного перевода, речевых технологий, а также проблемы, связанные с функциони-

рованием национальных языков в Интернет-технологиях. Участники отметили конструктивность обсуждения на секциях и круглых столах проблем разработки общей терминологии, общей системы обозначений лексико-грамматических категорий, использования для реализации своих национальных проектов аналогичных подходов, методов и технологий, особенно с учетом близости тюркских языков практически во всех компонентах, включая лексику, морфологию, синтаксис и семантику. Учитывая особую важность морфологической составляющей для тюркских языков и наибольшей теоретической и практической разработанности и представленности на конференции, соответствующие материалы, описывающие модели, методы и технологии обработки этой языковой компоненты собраны в отдельный том.

Тематика конференции находится в постоянном развитии. В список новых обсуждаемых тем включена проблема унификации систем грамматической аннотации в корпусах тюркских языков, которая подробно обсуждалась на семинаре Uniturk (проект "Унификация систем грамматической разметки в электронных корпусах тюркских языков"). В настоящее время не имеется единой унифицированной системы разметки для тюркских языков, включая стандартные теги для морфем и морфологических категорий. Вместе с тем, как показывает обсуждение имеющегося опыта и задач по этой проблематике, а также путей их решения, унификация систем аннотирования корпусов не является тривиальной практической задачей и требует теоретического пересмотра многих традиционных грамматических описаний. Для выполнения работ по унификации систем разметки в электронных корпусах тюркских языков в качестве программного инструментария уже в настоящее время можно эффективно использовать многофункциональный многоязычный Интернет-сервис на основе модели тюркской морфемы, разработанный в Институте прикладной семиотики АН РТ.

Организаторы и участники конференции будут и далее работать над превращением площадки конференции TurkLang в пространство согласованных лингвистических исследований, в пространство содействия разработке лингвистических ресурсов и эффективных систем и технологий обработки тюркских языков. Актуальной задачей является создание открытой платформы для размещения информационных ресурсов для тюркских языков (баз данных, терминологических и толковых словарей, тезаурусов), программных средств для обработки тюркских языков, прежде всего таких, как морфологические, синтаксические анализаторы и другие утилиты.

Организаторы конференции выражают благодарность директору Высшей школы Информационных технологий и информационных систем КФУ Хасьянову А.Ф., директору Института вычислительной математики и информационных технологий Казанского федерального университета (КФУ) Мосину С.Г., а также сотрудникам Научно-исследовательского института прикладной семиотики Академии наук Татарстана за их вклад в организацию и успешное проведение конференции «TurkLang 2017».

ПРОГРАММНЫЙ КОМИТЕТ

1. Сулейманов Джавдет Шевкетович (Казань, Татарстан, Россия) – председатель
2. Шарипбаев Алтынбек Амирович (Астана, Казахстан) – сопредседатель
3. Ешреф Адалы (Стамбул, Турция)
4. Дархан Кыдырали (Астана, Казахстан)
5. Алтынбек Гулила (Урумчи, Китай)
6. Дыбо Анна Владимировна (Москва, Россия)
7. Желтов Валериан Павлович (Чебоксары, Чувашия, Россия)
8. Замалетдинов Радиф Рифкатович (Казань, Татарстан, Россия)
9. Исраилова Нелла Амантаевна (Бишкек, Кыргызстан)
10. Кубединова Ленара Шакировна (Симферополь, Крым, Россия)
11. Мамедова Масума Гусейновна (Баку, Азербайджан)
12. Офлазер Кемаль (Доха, Катар)
13. Садыков Ташполот (Бишкек, Кыргызстан)
14. Салчак Аэлита Яковлевна (Кызыл, Тыва, Россия)
15. Сиразитдинов Зиннур Амирович (Уфа, Башкортостан, Россия)
16. Татевосов Сергей Георгиевич (Москва, Россия)
17. Торотоев Гаврил Григорьевич (Якутск, Саха, Россия)
18. Тукеев Уалишер Ануарбекович (Алматы, Казахстан)
19. Хасьянов Айрат Фаридович (Казань, Татарстан, Россия)

ОРГАНИЗАЦИОННЫЙ КОМИТЕТ

1. Гатиатуллин Айрат Рафизович
2. Невзорова Ольга Авенировна
3. Альменова Акмарал Байжановна
4. Аюпов Мадехур Масхутович
5. Галиева Альфия Макаримовна
6. Галимянов Анис Фуатович
7. Гатауллин Рамиль Раисович
8. Гильмуллин Ринат Абрекович
9. Курманбакиев Марат Ильдарович
10. Мухамедшин Дамир
11. Хакимов Булат Эрнстович
12. Хусаинов Айдар Фаилович
13. Якубова Диляра Джавдетовна

СЕКЦИЯ 1 | **СЕМАНТИЧЕСКИЕ ТЕХНОЛОГИИ**

УДК 81’33

BUILDING AN AUTOMATIC DICTIONARY IN THE EXPERT SYSTEM OF TEACHING SCIENTIFIC VOCABULARY

A. Alizade[1], Z. Guliyeva[2]
[1]EUROASIAN University, Baku AZ 1073, Azerbaijan
[2]Oxbridge Academy, Baku AZ 1009, Azerbaijan
guliyeva_z_y@hotmail.com

Artificial Intelligence Systems aimed at the achievement of the maximum simulation of human brain activities give opportunities for application of information computer technologies in foreign languages learning as well as the creating of automatic training courses and self-training apparatus. Paper illustrates classification of computer tutoring systems (CTS), their structure and also functions of such constituents as interface, databases, knowledge bases, diagnostic and control units. On the basis of conceptual approach authors offers the technique of automatic dictionary compiling within the computer tutoring system intended for educating of English lexis required to the academic or scientific staff. The principles of automatic dictionary building as the database of CTS are based on the results of modern ESL methods analysis together with the comparative analysis of the linguistic features essential for formalization of appropriate linguistic information compiled within automatic dictionary (AD).

Keywords: applied linguistics; expert tutoring system; automatic dictionary; English lexis, artificial intelligence.

ПОСТРОЕНИЕ АВТОМАТИЧЕСКОГО СЛОВАРЯ В ЭКСПЕРТНОЙ СИСТЕМЕ ОБУЧЕНИЯ НАУЧНОЙ ЛЕКСИКЕ

А. Ализаде[1], З. Кулиева[2]
[1] Бакинский Евроазийский Университет, Баку AZ 1073, Азербайджан
[2] Школа Оксбридж Академи, Баку AZ 1009, Азербайджан
guliyeva_z_y@hotmail.com

Системы Искусственного Интеллекта, нацеленные на достижение максимальной имитации деятельности человеческого мозга, дают возможность

применения информационных компьютерных технологий в обучении иностранным языкам, создания автоматических учебных курсов и тренажеров для самообучения. В статье предложена классификация компьютерных обучающих систем, в соответствии с их назначением и функциями. На основе концептуального подхода предлагается методика построения автоматического словаря в компьютерной обучающей системе, предназначенной для обучения лексике английского языка научному персоналу. Принципы построения автоматического словаря в качестве базы данных КОС, изложенные в статье, основаны на результатах анализа современных методов обучения иностранному языку, а также сравнительного анализа лингвистических признаков, необходимых для формализации и компьютеризации дидактических процессов того или иного языка. Структура контекстологического словаря, методика подачи и обучения материала, формализация данных в словарной статье описываемой системы базируются на подходе, основанном на коллокациях (collocation-based approach).

Ключевые слова: прикладная лингвистика; экспертная обучающая система; автоматический словарь; лексика английского языка; искусственный интеллект.

1. Введение

Высокие темпы развития информационных технологий создают предпосылки для изменений в требованиях к образованию, которое направлено на удовлетворение потребностей современного рынка труда. Динамично развивающаяся экономика, научный и технический прогресс обуславливают необходимость в квалифицированных кадрах, обучение которых возможно только в условиях постоянного обновления знаний посредством непрерывного образования. Компьютерные обучающие системы превратились в основное звено непрерывного образования, предоставляющие равные возможности повышения уровня знаний, умений и навыков самостоятельно или в массовом порядке. Процесс обучения представляет собой двусторонние отношения между обучающим и обучаемым, действия каждого их которых, обусловлены ответным действием второй стороны. Инициатором в процессе обучения является обучающий, который анализируя действия обучаемого, осуществляет постановку новых целей, заданий и условий, которые должны быть выполнены обучаемым. Компьютеризация процесса обучения на базе современных компьютерных технологий создает благоприятные условия для разработок компью-

терных обучающих систем (КОС) различного назначения, в особенности, систем, как для самостоятельного, так и для массового изучения иностранных языков.

Для изучения иностранного языка вне языковой среды особенно важна возможность создания и использования компьютерных программ в рамках изучаемого языка, что позволяет максимально использовать изучаемый язык, и является одной из главных междисциплинарных проблем как компьютерной лингвистики, так и современной лингводидактики. В зависимости от целей обучения, деятельность педагога или компьютерной обучающей системы базируется на содержании учебного материала, методах обучения, формах и средствах обучения.

Структура КОС определяется, в первую очередь, такими критериями как, возрастной фактор пользователей, цель использования, предметная область, доступность и простота использования системы, индивидуализация и интерактивность. Процесс обучения иностранному языку можно представить в качестве совокупности следующих этапов:

– Первичная диагностика знаний и установление уровня знаний обучаемого;

– Предоставление материала в соответствии с установленным уровнем знаний;

– Определение степени освоенности пройденного материала обучаемым;

– Контрольное тестирование для выявления слабых звеньев освоенного материала;

– Закрепление пройденного материала;

– Заключительное тестирование;

– Переход к следующему этапу обучения.

Учитывая вышесказанное, создание КОС сводится к построению компьютеризированной версии этапов процесса обучения, с учетом требований, предъявляемых для получения тех или иных результатов обучения.

2. Классификация КСО

Обучающая языковая система может быть создана как на основе специально разработанных программных продуктов, включающих в свою структуру программы разных типов, так и на основе

тщательно подобранного комплекса обучающих, прикладных и инструментальных программ.

Основной задачей разработчиков является создание наиболее продуктивной структуры КСО, которая будет обладать всеми возможностями для реализации вышеуказанных этапов процесса обучения в автоматизированной среде. В результате исследований в области существующих разработок компьютерных обучающих систем была разработана классификация КСО, ранжирующаяся по различным критериям и функциональным признакам:

1) В соответствии с *функциями обучения*, выполняемыми системой, их можно классифицировать на следующие виды:

– *Тренировочные КСО – системы,* в которых учащимся не предлагается теоретический материал, а предлагаются только вопросы и задачи в случайной последовательности;

– *Контролирующие КСО,* предназначенные для проверки умений и навыков обучаемого до начала обучения или в процессе обучения;

– *Информативные КСО*, где задачи и вопросы, в которых служат для организации человеко-машинного диалога и для управления ходом обучения, когда учащимся предлагается только теоретический материал для изучения;

– *Имитационные КСО* – представляют собой симуляционные модели обучения, имитирующие реальные ситуации, в которых обучаемые применяют те или иные знания и навыки;

– *Игровые КСО*, в которых учащиеся обеспечиваются использованием предоставляемых программой средств для реализации возможностей, связанных с изучением мира игры и деятельности в этом мире, что приводит к развитию определенного навыка обучаемого;

– *Диагностические КСО* применяются для определения уровня знаний обучаемого в соответствующей предметной области, где периодически производится повторный контроль знаний для выявления слабо усвоенного материала; индивидуализируется контроль знаний по определенному предмету для каждого пользователя.

2) По *структурным признакам взаимодействия обучающей системы с пользователем* КСО подразделяются на два базовых класса:

– прямые системы (без обратной связи) в которых, учащийся не взаимодействует, а лишь получает информацию и выполняет задания.

– циркулярные системы (с обратной связью), направленные на установление уровня знаний учащихся в определенный период учебного процесса, с возможностью развития определенных навыков и повторной проверки уровня знаний.

3) В соответствии с *обучаемыми языковыми навыками, методическим назначением определенных лингвистических аспектов и видами учебной деятельности* КСО можно поделить на следующие виды:

– Для развития артикуляционных навыков и приобретения фонетической базы

– Для изучения лексики (развития словарной базы)

– Для изучения грамматики

– Для развития навыков восприятия речи

– Для развития академического письма

– Для развития навыков разговорной речи

– Для развития техники чтения и орфографического письма

4) В контексте *алгоритмического построения* КСО можно поделить на следующие типы:

- КСО, построенные на линейном алгоритме Скиннера
- КСО, построенные на разветвленном (не линейно) алгоритме Кроудера
- КСО, построенные на адаптивном (интеллектуальном) алгоритме Паскома
- Комбинированные КСО, построенные на применении каждого из вышеуказанных алгоритмов на определенном отрезке обучения.

5) По способу *применения принципов гипертекстовых и гипермедийных технологий* при создании интеллектуальной обучающей среды, КОС можно разделить на следующие виды:

- Мультимедийные ОС
- Экспертные ОС
- Дистанционные ОС

Мультимедийные КОС являются наиболее популярным видом обучающих технологий, что обусловлено представлением информации не только в текстовом и графическом виде, но и посредством гипертекста, анимации, видео- и аудио- сопро-

вождения и интерактивных графических тестов. Применение аудио-визуальных форм информации активизирует одновременно несколько каналов восприятия, что облегчает понимание излагаемого материала, делает его интересным, наглядным и запоминающимся.

Экспертные КОС представляют собой вид интеллектуальных систем, осуществляющих сбор, хранение и обработку формализованной информации, основанной на знаниях экспертов – высоко квалифицированных методистов, лингвистов и педагогов. Структура экспертных систем позволяют использовать их как для профессионального обучения, так и для диагностики, контроля и проверки знаний обучаемых.

Дистанционные МОС – это системы, реализующие процесс обучения на расстоянии, который позволяет эффективно внедрять непрерывное образование, обучение вне зависимости от географии обучающегося, а также получение знаний в интерактивном режиме. Дистанционные системы позволяют расширить диапазон преподаваемых курсов без снижения их качества, сократить время, затрачиваемое на получение образования, обеспечивая обучение на рабочем месте, по месту жительства.

3. Предназначение ЭОС

Цели и задачи, поставленные перед ЭОС, предопределяют ее структуру и последовательность функционирования ее составных частей. Некоторые ЭОС предназначены для работы с отдельными элементами обучения, которые способствуют усвоению отдельных тем, текстов; другие представляют собой автоматизированные учебные курсы, и в зависимости от поставленных целей КОС может:

- контролировать знания учащихся
- содержать в себе элементы «учебного тренажа»
- содействовать овладению новым учебным материалом
- стимулировать интерес учащихся к изучаемому предмету.

Обучение иностранному языку предполагает формирование у учащихся иноязычной коммуникативной компетентности *(Щеглова Н.В.,2011)*. Учитывая коммуникативный подход, характерный для современных методов обучения английскому языку как иностранному, в центре внимания многих лингвистов находится понятие коммуникативной компетентности.

В последние десятилетия формирование коммуникативной компетентности рассматривается в отечественной и зарубежной методике преподавания иностранных языков в качестве цели и результата коммуникативного обучения *(Матвиенко В.Э., 2014).* Во многих дидактических словарях коммуникативная компетентность определяется как знание психологических, страноведческих, социальных факторов, которые определяют использование речи в соответствии с социальными нормами поведения. Однако навыки коммуникативной компетентности не могут внедряться отдельно от письма, чтения или же грамматики, с учетом современных требований лингводидактики, которые предписывают процессу обучения взаимосвязь всех указанных аспектов. Иными словами, такое многоуровневое понятие как коммуникативная компетентность базируется как на грамматической точности, беглости речи, так и на восприятии речи и корректного академического письма.

Целью курса разрабатываемой ЭОС было обучение лексике иностранного языка, и соответственно развития коммуникативной компетенции, в силу своей актуальности. Ссылаясь на необходимость изучения иностранных языков научными работниками, вне зависимости от раздела науки, в которой они специализируются, разработчики АС в ЭСО ориентировались на аудиторию, включающую в себя научные кадры, научно-техническую интеллигенцию и работников образовательных учреждений. В настоящее время практическое овладение одним или несколькими иностранными языками является не только требованием времени, а превратилось в прерогативу современности. Развитие и интеграция ИКТ во все сферы деятельности человека обусловили необходимость изучения иностранных языков не только в частных и совместных предприятиях, а также в научной и образовательной сфере. Интенсификация международного обмена информацией о полученных результатах, приобретенном опыте и проделанных инновационных работах вынуждает представителей всех функционирующих сфер на данном этапе развития страны изучать иностранные языки, в частности, английский язык.

4. Методы построения Контекстологического словаря как базы данных ЭСО

При создании автоматического словаря в компьютерной системе обучения иностранному языку, в первую очередь, необхо-

димо следовать следующим принципам, оптимизирующим процесс организации словаря, структурирования словарных статей и отбора соответствующей информации для определенных частей речи. Методика построения автоматического словаря основана на следующих этапах:

1. Определение целей и задач курса обучения, включающих такие особенности как обучаемая аудитория, предметная область и обучаемые лингвистические знания и навыки.

2. Отбор методов и приемов преподавания иностранного языка, оптимальных при автоматизации процесса обучения, направленного на выработку определенных знаний и навыков

3. Создание сценария КСО и соответствующей ему текстовой базы данных, основанной на содержании поурочного материала

4. Отбор лемм и построение словника автоматического словаря в КОС

5. Формализация данных, отобранных в словарную статью автоматического словаря для различных частей речи

6. Представление формализованных данных АС в различных функциональных блоках КОС.

Определение сценария работы системы обучения ИЯ напрямую связано с теми знаниями, которыми обучаемый будет обладать по прохождении данного курса. Выработка коммуникативной компетенции базируется, в первую очередь, на изучении, закреплении и правильном применении лексики изучаемого языка, а следовательно, на знании слов, сочетаний и выражений данного языка.

5. Особенности текстовой базы и словника ЭСО

В языковых обучающих системах в большинстве случаев функционирует строго-фиксированное количество словоформ, закрепленное за каждым отдельным обучающим модулем, включая тексты, упражнения, аудио и видео материалы, фото и иллюстрации, и возможность выхода за пределы предлагаемого контекста зависит от вида КОС, ее структуры и назначения. Важный обуславливающий фактор сосредоточен в способности процесса обучения любой предметной области изменяться от простого к сложному, так в процессе обучения иностранному языку при подаче материала переводное соответствие трансформируется по

иерархической линии от словоформы к словосочетанию, затем к предложению и, наконец, к гипертексту (Рис.1). Стратегия обучения иностранного языка, методология, используемая при подаче материала, закрепления и контроля полученных обучаемым знаний, наряду со структурой самой системы, а также со структурой обучаемых модулей влияют на информацию, хранимую с автоматическом словаре КОС. Решение таких трудно формализуемых проблем, с которыми разработчики сталкиваются при создании автоматических систем обработки текста, как омонимия и многозначность, не всегда возможны на уровне словоформ и словосочетаний.

В этой связи, обучающие модули более высоких языковых уровней предоставляют различные возможности преодоления трудностей, связанных с омонимией и многозначностью на уровне контекста, путем включения в словарную статью одного или нескольких предложений для создания более точного восприятия контекстной ситуации. Корпус словаря включает **слова**, **словосочетания**, **словоизменительные и словообразовательные элементы**. Невозможно составлять корпус словаря не принимая во внимание лингводидактические требования, предъявляемый материал, методы подачи и закрепления материала.

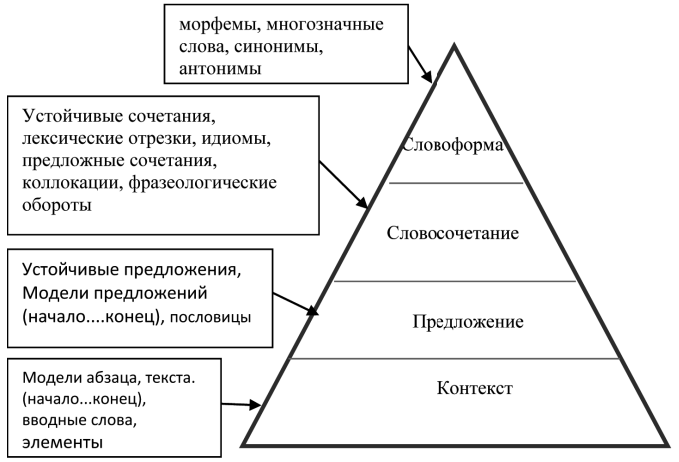


Рис. 1. Трансформация слов в рамках контекста

В соответствии с курсом обучения лексике языка, словарь, представляет собой неотъемлемый ингредиент коммуникативных способностей и самого процесса общения. Изучение слов – это сложный процесс, в котором приобретенный словарный запас должен постепенно заучиваться, закрепляться и повторно перерабатываться, что в данном случае обозначает, что слово (а иногда некоторые группы слов до 10 раз) должно преподноситься несколько раз в новых контекстах.

Фразеологические обороты, идиомы и устойчивые словосочетания специфического назначения входят в состав словосочетаний, которые вносятся в более расширенную базу словарного корпуса. Более нейтральные и упрощенные словосочетания, применяются чаще, чем сложные сочетания, у которых усложненная идиоматичность. Для научной лексики характерны фразовые глаголы, соединительные союзы и элементы, а также коллокации, своевременность и правильность употребления которых имеет первостепенное значение для коммуникативной компетентности как в устной, так и в письменной речи. Фразовые глаголы, терминологические и лексикализованные словосочетания, составные наречия, предлоги, идиомы и фразеологизмы – все являются разновидностями словосочетаний.

Для более точного представления о лексическом составе второго уровня словника АС приведем классификацию типов словосочетаний, входящих в состав контекстологической базы КСО:

1. Полислова (polywords) – əlbətdə ki/ of course
2. Chunk*s* – *у*стойчивые выражения – əlavə olaraq/in addition
3. Коллокации – təhsil almaq/receive education
4. Фразеологические обороты – mahiyyəti təffərüatdadır/the devil is in the details
5. Идиомы – başlanğıcını/sasını qoymaq/break new ground
6. Фразовые глаголы – nəzəriyyə irəli sürmək/put forward
7. Терминологические выражения – tələb və təklif/ supply and demand

Предлагаемые принципы построения АС могут применяться при разработке интеллектуальной системы обучения научной лексике любого языка. Структура контекстологического словаря, методика подачи и обучения материала, формализация данных в словарной статье данной системы базируются на подходе, основанном на коллокациях (collocation-based approach). Совершенно

новая концепция подхода на основе коллокаций дает возможность более точного и облегченного деления контекста, предъявления и формального представления данных текстовой базы и АС, при обучении устной речи и грамотному письму.

ЛИТЕРАТУРА

1. Алисейчик С.Г., Вашик К., Кнап Ж., Кудрявцев В.Б., Строгалов А.С., Шеховцов П.А. Компьютерные обучающие системы. www.intsys.msu.ru/magazine/archive/v8(1-4)/strogalov-005-044.pdf

2. Бурдаев В.П. ПИОС − почти интеллектуальная обучающая система // Искусственный интеллект, 2009, №4.www.nbuv.gov.ua/portal/natural/ii/2009_4/7%5C00_Burdaev.pdf

3. Варинская Виктория Михайловна. Контекстологический словарь как элемент обучающих систем : диссертация 10.02.21. – Москва, 2005. – 182 с. : ил. + Прил. (130с.). РГБ ОД.

4. Деркач А.А., Щербак С.Ф. Педагогическая эвристика: Искусство овладения иностранным языком. М.: Педагогика, 1991, – 219 с.

5. Домрачев В.Г., Ретинская И.В. О классификации образовательных информационных технологий // Информационные технологии, 1996, № 2. – с. 10–13.

6. Зубов. А.Н. Information Technology in Linguistic. http:www.amozon.com/information

7. Крюкова. О.П. Самостоятельное изучение иностранного языка в компьютерной среде (на примере английского языка). М.: Издательская корпорация «Логос», 1998.

8. Кулиева. З. Ю. Модели Автоматических обучающих систем, их структура и классификация. İnformasiya texnologiyaları problemləri, №1(5), 2012, 89-9690 www.jpit.az

9. Лобанова.М.А, Проектирование АОС. Материалы первой студенческой научно-практической конференции. 2008

10. Марчук.Ю.Н. Основы Компьютерной лингвистике. М. 1999.

11. Матвеенко В.Э., Лингводидактическая система обучения иностранных студентов-филологов национально-окрашенной лексике с использованием аудио-видеосредств. Диссертация. Москва – 2014

12. Нелюбин.Л.Л. Контекстологический словарь как элемент обучающей системы. http:www.planetadisser.com/see/dis_8307html

13. Павлова.И.П. Контекстологический словарь как элемент обучающей системы. http:www.planetadisser.com/see/dis_8307html

14. Петрушин В.А. Экспертно-обучающие системы АН УССР. Ин-т кибернетики. – Киев.

15. Пиотровская. К.Р. Современная компьютерная лингводидактика. – ВИНИТИНТИ. серия 2 № 4, 1991.

16. Потапова Р.К. 2002.2003. Computer Assisted Language Learning-Wikipedia, the free encyclopedia. http:www. Enwikipedia.org/…/Computer-assisted_language.com

17. Смирнов Ю.М., Андреев А.М., Березкин Д.В., Друшляков Г.И. Компьютерные системы в обучении русскому языку как иностранному. www.inteltec.ru/publish/articles/others/at_rus2.shtml

18. Талызина Н.Ф. Теоретические проблемы программированного обучения. – М.: Изд-во МГУ, 1969. – 133 с.

19. Халеева.И.И Контекстологический словарь как элемент обучающей системы. http:www.planetadisser.com/see/dis_8307html

20. Щеглова Н. В. Формирование коммуникативной компетенции в процессе обучения иностранным языкам./Историческая и социально-образовательная мысль.2011 https://cyberleninka.ru/article/n/

21. Intelligent Tutoring System For Marathi. Dissertation.Karnataka State Open University December, 2005.

22. Kudryavcev V.B. and others. Modeling educational process using expert system. html//intsys.msu.ru/en/staff

23. Levy M. (1997) CALL: context and conceptualization, Oxford: Oxford University Press.

24. Marty, F (1981). “Reflections on the use of computers in second language acquisition”. System 9 (2): 85–98.

25. Skinner B.F. The science of learning and art of teaching. // Harward Education Review, Spring, 24, 1954. – p. 86–97.

УДК 004.048:519.765

APPLICATION OF THE RHETORICAL STRUCTURE THEORY IN AUTOMATIC TEXT PROCESSING SYSTEMS

A. Bakiyeva[1], T. Batura[2]
[1]Novosibirsk State University,
Novosibirsk, 630090, Russia
m_aigerim0707@mail.ru

[2]Novosibirsk State University,
A.P. Ershov Institute of Informatics Systems SB RAS,
Novosibirsk, 630090, Russia
tatiana.v.batura@gmail.com

In this paper, we investigated the possibility of applying the theory of rhetorical structures to the analysis of texts with scientific and technical topics in Russian and Kazakh languages, and described some of the formal features of rhetorical relations.

Automatic detection of rhetorical relations in texts allows you to locate the nuclear and the satellite. Since the nuclear contains the most important part of the statement, it is necessary to compile a short text annotation in order to convert the source text. The system will perform different actions depending on different markers and discursive relations. For a formal description of actions, we use the first and second order predicate logic formulas.

Based on the proposed methods, we developed the system to determine rhetorical relations. With the help of it, an experiment was conducted to find out the relations characterizing the texts of scientific and technical topics. We collected and analyzed 168 articles in Russian and 207 articles in Kazakh, with an average length of 7-12 pages. In the experiment, we considered 11 relations and about 40 markers in each language. The precision of the results is quite high.

Keywords: automatic text processing; rhetorical structure theory; discourse marker; text analysis; rhetorical relations; semantics.

ПРИМЕНЕНИЕ ТЕОРИИ РИТОРИЧЕСКИХ СТРУКТУР В СИСТЕМАХ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

А.М. Бакиева[1], Т.В. Батура[2]
[1]Новосибирский государственный университет,
г. Новосибирск, 630090, Россия
m_aigerim0707@mail.ru

[2]Новосибирский государственный университет,
Институт систем информатики им. А.П. Ершова СО РАН,
г. Новосибирск, 630090, Россия
tatiana.v.batura@gmail.com

Исследована возможность применения теории риторических структур для анализа текстов научно-технической тематики на русском и казахском языках. Описаны некоторые формальные признаки риторических отношений.

Автоматическое определение риторических отношений в тексте позволяет установить местоположение ядра и сателлита. Для составления краткой аннотации необходимо преобразовать первоначальный текст, исходя из предположения, что ядро содержит наиболее важную часть высказывания. В зависимости от разных маркеров и дискурсивных отношений будут выполняться разные действия. Для формального описания преобразования текста используются формулы логики предикатов первого и второго порядка. Предложенный в статье метод может применяться в системах автоматического реферирования и извлечения информации.

Ключевые слова: автоматическая обработка текста; теория риторических структур; дискурсивный маркер; анализ текста; риторические отношения; семантика.

Введение

Ввиду стремительного увеличения объемов текстовой информации в Интернете исследования в области компьютерной лингвистики сохраняют свою актуальность. Разработка алгоритмов и создание систем автоматического реферирования, поиска и извлечения информации, классификации и кластеризации текстовых документов по-прежнему являются сложными задачами.

В последнее время все чаще встречается мнение, что языковые явления не могут быть адекватно поняты и описаны вне их употребления, без учета их дискурсивных аспектов [1]. Дискурс

часто отождествляется с текстом, состоящим из предложений (коммуникативных единиц языка) и их объединений в более крупные единства, находящиеся в непрерывной смысловой связи. Другими словами, дискурс – это не только связная последовательность предложений, противопоставляемая изолированному предложению, но и определенное семантическое единство, обладающее семантической связностью [2], и как следствие, содержащее знания о мире, о ситуации, социальные и другие виды знаний.

В работе [3] описан опыт создания корпуса на русском языке, содержащего дискурсивную разметку. В корпус включены тексты различных жанров (научные, научно-популярные, новостные), он является общедоступным. Прежде, чем применять теорию риторических структур, ее следует адаптировать для конкретного языка. Это связано с грамматическими особенностями. Авторы привели в своей статье иерархию риторических отношений, которые, как показало их исследование, удобнее и правильнее учитывать при разметке русских текстов.

Существуют попытки применения дискурсивного анализа для решения различных задач компьютерной лингвистики. Подробный обзор литературы, представленный в работе [4], показывает, что в большинстве случаев дискурсивный анализ способен улучшить качество автоматических систем на 4–44 % в зависимости от конкретной задачи. В то время как для английского языка разработки в данной области выходят на достаточно высокий уровень, для русского языка подобных исследований мало [4-6]. Для казахского языка таких исследований ранее не проводилось.

1. Семантический анализ и формальные признаки риторических отношений

Теория риторических структур (Rhetorical Structure Theory) – одна из наиболее известных теорий организации текстов [7]. Согласно ей, сначала текст делится на непересекающиеся фрагменты, называемые элементарными дискурсивными единицами (ЭДЕ).

Далее последовательные ЭДЕ соединяются между собой риторическими отношениями. Эти части являются элементами, из которых строятся более крупные фрагменты текстов и целые тек-

сты. Каждый фрагмент по отношению к другим фрагментам выполняет определенную роль. Текстовая связность формируется посредством тех отношений, которые моделируются между фрагментами внутри текста.

В теории риторических структур определено два типа ЭДЕ. Ядро рассматривается в качестве наиболее важной части высказывания, тогда как сателлиты поясняют ядра и являются вторичными. Ядро содержит основную информацию, а сателлит содержит дополнительную информацию о ядре. Сателлит часто бывает непонятным без ядра. В то время как выражения, где сателлиты были удалены могут быть поняты в определенной степени.

Ниже приведены маркеры, соответствующие им риторические отношения и примеры предложений с ними.

Пример 1.

Маркер: оған қоса (кроме того)

Название отношения: Elaboration

Текст на казахском: Үй жаман емес көрінді. ***Оған қоса***, бағасы да тиімді болды.

Текст на русском: Дом выглядел неплохо. ***Кроме того***, цена была подходящая

Для удобства дальнейшего изложения введем следующие обозначения.

Пусть

x – ядро;

y – маркер;

z – саттелит;

S (x) – ЭДЕ, являющееся ядром;

S' (x) – ЭДЕ с заглавной буквы, являющееся ядром;

S (z) – ЭДЕ, являющееся сателлитом;

S' (z) – ЭДЕ с заглавной буквы, являющееся сателлитом ;

y' – маркер с заглавной буквы;

p () – знак пунктуации, аргументом может быть «.», «,», «:», «;».

Теперь рассмотренный пример 1 может быть представлен в виде формулы логики исчисления предикатов: $S'(x) \wedge p(.) \wedge y' \wedge S(z) \wedge p(.)$.

Пример 2.

Маркер: себебі (потому что)

Название отношения: Cause-Effect

Текст на казахском: Ол өте білімді, ***себебі*** бар ынтамен оқып, барлық тапсырмаға ұғынып жатты.

Текст на русском: Она очень умная, ***потому что*** училась усердно, вникала во все задания.

Формульное представление: $S'(x) \wedge p(,) \wedge y \wedge S(z) \wedge p(.)$.

Пример 3.

Маркер: оған қарамастан (несмотря на)

Название отношения: Concession

Текст на казахском: Дәрігерлердің қарсылықтарына ***қарамастан***, ол ана болуды шешті

Текст на русском: ***Несмотря на*** все противопоказания врачей, она решила стать матерью.

Формульное представление: $S'(z) \wedge y \wedge p(,) \wedge S(x) \wedge p(.)$.

Примр 4.

Маркер: дегенмен (хотя)

Название отношения: Concession

Текст на казахском: Анизотропты және изотроптық жағдайларға арналған контурлық кернеулерді бөлу сапалық жағынан ұқсас, ***дегенмен*** кейбір сандық айырмашылықтар бар.

Текст на русском: Распределение контурных напряжений для анизотропного и изотропного случаев качественно подобно, ***хотя*** и имеется некоторое количественное различие.

Формульное представление: $S' (x) \wedge p(,) \wedge y \wedge S(z) \wedge p(.)$.

Пример 5.

Маркер: мысалы (например)

Название отношения: Example

Текст на казахском: Шинаны бөлшектеу алдында оны металл корды мен синтетикалық кордтан босатып алу керек. ***Мысалы***, резина ұнтағында металл бөлшектері массаның 0,01–0,03 % аспауы шарт.

Текст на русском: Перед демонтажом шины ее необходимо снять с металлического шнура и синтетического шнура. ***Например***, резиновый порошок не должен превышать 0,01–0,03% частиц металла.

Формульное представление: $S'(x) \wedge p(.) \wedge y' \wedge p(,) \wedge S(z) \wedge p(.)$.

Пример 6.

Маркер: сондықтан (поэтому)

Название отношения: Evidence

Текст на казахском: Жоғарыда келтірілген барлық параметрлердің әсерін қабылдауға және суффозия мен кольматацияны болжау мүмкін емес. ***Сондықтан***, формулалар қатаң анықталған жағдайлар үшін пайдаланылады және құм көші-қон процессі тек қана ұңғыма айналасындағы аймақта сипатталады.

Текст на русском: Невозможно предсказать влияние всех вышеперечисленных параметров и предсказать кальцификацию и успокоение. ***Поэтому*** формулы используются для строго определенных ситуаций, и процесс миграции песка характеризуется только в области вокруг скважины.

Формульное представление: $S'(z)\wedge p(.)\wedge y'\wedge S(x)\wedge p(.)$.

2. Формальное описание преобразования текста

Чтобы автоматически получить краткий реферат, необходимо сначала обнаружить ядерные ЭДЕ в тексте. Затем преобразовать высказывания, содержащие эти ядерные ЭДЕ так, чтобы текст формируемого реферата получился связным. В зависимости от разных маркеров и дискурсивных отношений, эти преобразования будут разные. Ниже описаны некоторые из рассмотренных нами. Для формального описания действий, выполняемых системой, было решено использовать логику предикатов первого и второго порядка.

2.1. Предикаты первого порядка

Согласно обозначениям, введенным в предыдущем разделе, можно описать действия, выполняемые системой, следующим образом.

В примере 1 с маркером y = “Оған қоса” необходимо удалить сателлит вместе с маркером и оставить предыдущее предложение, являющееся ядерным ЭДЕ, то есть

$$S'(x)\wedge p(.)\wedge y'\wedge S(z)\wedge p(.)\rightarrow S'(x)\wedge p(.)\wedge\neg(y'\wedge S(z)\wedge p(.)). \qquad (1)$$

В примере 2 с маркером y = “себебі” надо удалить маркер с сателлитом и оставить первую часть предложения до маркера, то есть

$$S'(x)\wedge p(,)\wedge y\wedge S(z)\wedge p(.)\rightarrow S'(x)\wedge p(.)\wedge\neg(y\wedge S(z)\wedge p(.)). \qquad (2)$$

В примере 3 с маркером y = “оған қарамастан” надо первую часть предложения удалить и оставить вторую часть предложения, то есть

$$S'(z)\wedge y\wedge p(,)\wedge S(x)\wedge p(.)\rightarrow\neg(S'(z)\wedge y\wedge p(,))\wedge S'(x)\wedge p(.). \quad (3)$$

В примере 4 с маркером y = “дегенмен” используется действие, аналогичное примеру 2:

$$S'(x)\wedge p(,)\wedge y\wedge S(z)\wedge p(.)\rightarrow S'(x)\wedge p(.)\wedge\neg(y'\wedge S(z)\wedge p(.)). \quad (4)$$

В примере 5 с маркером y = «Например» необходимо оставить ядро, т.е. то предложение, которое предшествует маркеру. При этом сам маркер и сателлит надо удалить, то есть

$$S'(x)\wedge p(.)\wedge y'\wedge p(,)\wedge S(z)\wedge p(.)\rightarrow S'(x)\wedge p(.)\wedge\neg(y'\wedge p(,)\wedge S(z)\wedge p(.)). \quad (5)$$

В примере 6 с маркером y = «Поэтому» необходимо удалить сателлит и сам маркер, и оставить ядерное ЭДЕ с заглавной буквы, то есть

$$S'(z)\wedge p(.)\wedge y'\wedge S(x)\wedge p(.)\rightarrow\neg(S'(z)\wedge p(.)\wedge y)\wedge S'(x)\wedge p(.). \quad (6)$$

2.2. Предикаты второго порядка

Случаи вложенных ЭДЕ, когда ЭДЕ более низкого уровня вкладываются в ЭДЕ более высокого уровня, удобнее описать при помощи предикатов второго порядка. Причем для каждого маркера вводится отдельный предикат. Краткий список подобных предикатов приведен ниже в таблице 1.

Таблица 1. **Предикаты для маркеров**

Название отношения	**Предикат**	**Маркер (на казахском)**	**Маркер (на русском)**
Elaboration (Детализация)	*El1* *El2*	оның салдарынан Сонымен қатар	Вследствие (того, чего,) Кроме того
Contrast (Контраст)	*Cont1* *Cont2* *Cont3*	Алайда қарамастан Ескерту	Однако Несмотря на то, что Обратите внимание
Evidence (Обоснование)	*Ev1* *Ev2* *Ev3*	Әлбетте, бұл Оған дәлел болып Осылайша	Очевидно, что Доказательством тому Таким образом

Cause-Effect (Причина)	*CEf1* *CEf2* *CEf3* *CEf4*	Неліктен кесірінен Өйткені Сондықтан	Почему Из-за Так как Поэтому

Чтобы пояснить, как происходит преобразование текста в случаях вложенных ЭДЕ, рассмотрим два примера.

Пример 7.

Текст на казахском: ***Сонымен қатар***, мұздатқыш камерасына енген әуе кіріс ауаның жылу жүктемесі шамамен 50% құрайды тоңазытқыш жүйесі док станциясын, пайдаланып °C 1,5 суыған. ***Осылайша***, таза әсері 1-0,7 * 0.5 = 0.65 ретінде шамамен 65% -ға қондыру жеткізу инфильтрация мұздатқыш жүктеме қысқарту суытылады. Таза табыс мұздатқыштың инфильтрация жүктемені азайту және тоңазытқыш жүктеме тасымалдау док арасындағы айырма болып табылады. ***Ескерту*** бұл док тоңазытқыштар әлдеқайда жоғары температура (°C 1.5 °C орнына -23) жұмыс, және салқындату сол сомаға әлдеқайда аз қуатты тұтынады.

Текст на русском: ***Кроме того***, воздух, который поступает в морозильную камеру уже охлаждают до 1.5°C с помощью холодильной установки док-станции, которая составляет около 50% тепловой нагрузки поступающего воздуха. ***Таким образом***, чистый эффект охлаждаемой док доставки является уменьшение инфильтрации нагрузки морозилку примерно на 65% поскольку 1-0,7*0,5=0,65. Чистая прибыль равна разнице между уменьшением инфильтрации нагрузки морозильной камеры и холодильной нагрузки доке судоходства. ***Обратите внимание***, что док холодильники работают при значительно более высоких температурах (1.5°C вместо -23°C), и, потребляют значительно меньше энергии на ту же сумму охлаждения.

Формульное представление примера на казахском языке:

$$S'(z)\wedge p(.)\wedge S'(x)\wedge p(.)\wedge S'(x)\wedge p(.)\wedge S(z)\rightarrow\neg S(El2\wedge S(x)\wedge p(.))\wedge$$
$$\wedge S'(\neg(Ev3\wedge p(,))\wedge S'(x)\wedge p(.))\wedge S'(x)\wedge\neg S(Cont3\wedge S(z)). \qquad (7)$$

Пример 8.

Текст на казахском: ***Алайда***, тіпті қалыпты заң үшін арифметикалық орта орташа мәнді бағалау болып табылмайды, ал медиана шығарындылардың болуында эмпирикалық орташа мәнді

бағалауға мүмкіндік береді. ***Сондықтан***, медианаңы пайдалану регрессиялық тәуелділіктің параметрлерін баяулатуға мүмкіндік бергеніне ***қарамастан***, орташа мәнді эмпирикалық бағалаулар медианы пайдалану регрессиялық тәуелділіктің параметрлерін баптау процедурасын жасайды.

Текст на русском: ***Однако*** даже для нормального закона среднее арифметическое не является робастной оценкой среднего значения, в то время как медиана позволяет оценивать эмпирическое среднее при наличии выбросов. ***Поэтому*** для построения параметрических регрессионных зависимостей также используются эмпирические оценки среднего при помощи медианы, ***несмотря на то, что*** использование медианы делает процедуру настройки параметров регрессионной зависимости более медленной.

Формульное представление примера на казахском языке:

$$S'(z)\wedge p(.)\wedge S(x)\wedge p(.)\rightarrow\neg(S(Cont1\wedge p(,)\wedge S(z))\wedge p(.))\wedge \\ \wedge\, S(\neg(CEf4\wedge S'(z)\wedge p(,)\wedge Cont2)\wedge S'(x))\wedge p(.)). \qquad (8)$$

3. Описание экспериментов и полученные результаты

Автоматическое определение риторических отношений в тексте позволяет установить местоположение ядра и сателлита. Для формирования реферата необходимо выполнять определенные действия с текстом в зависимости от разных маркеров и дискурсивных отношений. Поскольку ядро содержит наиболее важную часть высказывания, то предложенный метод можно использовать в системах автореферирования и извлечения информации из текстов.

На основе предложенных методов был создан инструмент для определения риторических отношений. С помощью него было решено провести эксперимент по обнаружению отношений, характеризующих тексты научно-технической тематики. В ходе эксперимента было проанализировано 168 статей на русском языке, средняя длина которых 7–12 страниц. На казахском языке была собрана коллекция из 207 статей. В эксперименте рассматривалось в общей сложности 11 отношений и около 40 маркеров на каждом языке. Распределение рассмотренных риторических отношений в текстах показано в таблице 2.

Таблица 2. **Рассмотренные риторические отношения в текстах научно-технической тематики**

	Название отношения	**Кол-во примеров на казахском**	**Кол-во примеров на русском**
1.	Условие, Condition	1854	3458
2.	Причина, Cause-Effect	759	2123
3.	Пример, Example	1492	1329
4.	Переформулировка, Restatement	761	1274
5.	Контраст, Contrast	291	1112
6.	Уступка, Consession	223	802
7.	Цель, Purpose	3584	671
8.	Детализация, Elaboration	631	629
9.	Обоснование, Evidence	47	259
10.	Сравнение, Comparison	92	77
11.	Источник информации, Attribution	15	32

В результате можно сделать вывод, что для обоих языков научно-технические тексты в большей мере характеризуются следующими отношениями: *Условие, Причина, Пример, Переформулировка*. Кроме того, в статьях на казахском языке в большом количестве представлено отношение *Цель*, а в статьях на русском – *Контраст*. Значит, в дальнейших исследованиях научно-технических текстов имеет смысл детальнее изучить

именно их. Отношения *Источник информации* и *Сравнение* практически не представлены в собранных коллекциях.

Для оценки точности предложенного метода использовалась экспертная оценка. Исходя из полученных данных, было решено рассмотреть наиболее часто встречающиеся отношения. Точность была оценена для каждой коллекции по формуле:

$$Precision = \frac{TP}{TP + FP},$$

где TP – истинно положительное решение;

FP – ложно положительное решение.

В таблице 3 приведена оценка точности определения наиболее часто встречающихся риторических отношений.

Предположительно, возникшие ошибки связаны с особенностями словоупотребления, а также возможным наличием опечаток и ошибок в исходных текстах. Для более детального анализа требуется проведение дополнительных более масштабных экспериментов.

Таблица 3. **Оценка точности определения риторических отношений**

	Название отношения	**Точность (коллекция на казахском)**	**Точность (коллекция на русском)**
1.	Условие, Condition	0,92	0,896
2.	Причина, Cause-Effect	1,0	0,987
3.	Пример, Example	0,98	0,981
4.	Переформулировка, Restatement	0,963	0,968
5.	Контраст, Contrast	0,973	1,0
6.	Цель, Purpose	1,0	0,989

Заключение

В данной работе исследована возможность применения теории риторических структур для анализа текстов научно-технической тематики. Предпринята попытка формально описать признаки

некоторых отношений, на основании которых удавалось бы однозначно установить соответствие. Для формального описания действий, осуществляемых с текстом, используется язык логики предикатов первого и второго порядка.

Автоматическое определение риторических отношений в тексте позволяет установить местоположение ядра и сателлита. Поскольку ядро содержит наиболее важную часть высказывания, то предложенный метод может применяться в системах извлечения информации и автоматического реферирования.

ЛИТЕРАТУРА

1. Прокошенкова Л.П., Гецкина И.Б. (2006). Дискурсивный анализ и его роль в современной лингвистике. №4. URL: http://cyberleninka.ru/article/n/diskursivnyy-analiz-i-ego-rol-v-sovremennoy-lingvistike (дата обращения: 20.01.2017).

2. Темнова Е.В. (2004). Современные подходы к изучению дискурса. *Язык, сознание, коммуникация: сб. статей*. М.: МАКС Пресс. Вып. 26. С. 24–32.

3. Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A. (2017). Towards building a discourse-annotated corpus of Russian. *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference "Dialogue 2017"*. Iss. 16 (23). V. 1. pp. 194–204.

4. Ананьева М.И., Кобозева М.В. (2016). Разработка корпуса текстов на русском языке с разметкой на основе теории риторических структур. *Труды Международной конференции «Диалог». Студенческая сессия*. URL: www.dialog-21.ru/media/3460/ananyeva.pdf (дата обращения: 18.01.2017).

5. Batura T.V., Bakiyeva A.M., Yerimbetova A.S. Mit'kovskaya M.V. Semenova N.A. (2016). Methods of constructing natural language analyzers based on Link Grammar and rhetorical structure theory. *Bulletin of the Novosibirsk Computing Center. Series: Computer Science*. Novosibirsk: NCC Publisher, Is. 40. pp. 37–51.

6. Тревгода С.А. (2009). Методы и алгоритмы автоматического реферирования текста на основе анализа функциональных отношений. *Автореферат диссертации на соискание ученой степени кандидата технических наук*. Санкт-Петербург. С. 15.

7. Mann W., Thompson C. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*. V. 8. N 3. P. 243–281.

УДК 81`33

MORPHOLOGICAL DISAMBIGUATION IN TATAR LANGUAGE USING PUREPOS

R. Gilmullin, R. Gataullin
Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic Kazan, Russia
rinatgilmullin@gmail.com

This article describes the results of relevant studies devoted to the morphological disambiguation of texts in the Tatar language. As an instrument for the disambiguation, the open software product PurePos 2.0 based on hidden Markov networks was selected. As training data for the experiments, the annotated and disambiguated subcorpus of the Tatar language with a volume of 2,5 million lexical units was used.

Keywords: PurePos, lemmatization; POS-tagging; morphological disambiguation; Tatar language.

РАЗРЕШЕНИЕ МОРФОЛОГИЧЕСКОЙ МНОГОЗНАЧНОСТИ ТЕКСТОВ НА ТАТАРСКОМ ЯЗЫКЕ НА ОСНОВЕ ИНСТРУМЕНТАРИЯ PUREPOS

Р.А. Гильмуллин, Р.Р. Гатауллин
Институт прикладной семиотики Академии наук Республики Татарстан, Казань, Россия
rinatgilmullin@gmail.com

В данной статье описываются результаты актуальных исследований, посвященных разрешению морфологической многозначности текстов на татарском языке. В качестве инструментария разрешения многозначности выбран открытый программный продукт PurePos 2.0, основанный на скрытых Марковских сетях. Для проведения экспериментов в качестве обучаемых данных использовался размеченный подкорпус татарского языка объемом 2,5 миллионов лексических единиц со снятой многозначностью.

Ключевые слова: PurePos, лемматизация, морфологическая многозначность, татарский язык.

1. Введение

Центральной проблемой интеллектуального анализа текстов является проблема разрешения многозначности. К настоящему времени сформирована основная парадигма методов снятия многозначности, которая включает методы, основанные на правилах; методы машинного обучения, использующие вероятностные модели; гибридные методы. Создание электронного корпуса татарского языка «Туган тел» (http://tugantel.tatar/) и репрезентативного подкорпуса со снятой вручную морфологической многозначностью данных дали возможность исследования данной задачи с применением статистико-вероятностных методов.

Анализ открытых программных продуктов, разработанных для этой задачи, показал, что наиболее перспективным программным продуктом является инструментарий PurePos (https://github.com/ppke-nlpg/purepos), реализующая гибридную модель на основе скрытых Марковских моделей (HMM - Hidden Markov model). Скрытая Марковская модель − модель процесса, в которой процесс считается Марковским, причем неизвестно, в каком состоянии находится система (состояния скрыты), но каждое состояние может с некоторой вероятностью произвести событие, которое можно наблюдать. Другими словами, изучается Марковский процесс с неизвестными параметрами, и задачей является распознавание неизвестных параметров на основе наблюдаемых.

Инструмент PurePos разрабатывался для языков со сложной морфологией (в том числе для агглютинативных языков) и с малой ресурсной базой (low resource languages). Основные эксперименты (*Orosz, G., Novák, 2012*; *Orosz, G. and Novák, A., 2013*) проводились для венгерского языка. Результаты по распознаванию POS-тегов слов имеют оценки точности 97%. Код инструмента открыт и распространяется под лицензией LGPL. Авторами выдвинуто предположение, что модель должна работать и для других языков (в том числе для языков тюркской группы).

Нами проведены эксперименты по применению инструмента PurePos для распознавания морфологической многозначности для татарских текстов. Для проведения экспериментов в качестве обучемых данных использовался размеченный подкорпус татарского языка объемом 2,5 миллиона лексических единиц со снятой многозначностью.

2. Описание экспериментов

2.1 Разрешение морфологической многозначности

При использовании PurePos нас интересовало решение двух задач:

Точность разрешения морологической многозначности, а также распознавание разметки NR (лемма, Part-of-Speech(POS) и вся морфологическая цепочка), где NR (Not Recognized) – это обозначение тега для нераспознанной морфологическим модулем словоформы.

Начальная выборка размеченных данных была разделена на выборку для обучения (90%) и тестовую выборку (10%). Модель обучалась только на обучающей выборке, тестовая выборка использовалась только для тестирования. Позже из текстов корпуса, не вошедших ни в одну из предыдущих выборок, была сформирована выборка для валидации.

Для оценки качества работы модели был использован текст как в исходной форме без разметок, например, предложение (“Без барабыз”), так и с предварительной морфологической разметкой, например, ([(“Без”, [(“без”, “[PN]”, 0.5), (“без”, “[N][Sg][Nom]”, 0.5)]), (“барабыз”, [(“бар”, “[V][PRES(Й)][1PL(бЫз)]”)]), “.”]).

Результаты экспериментов, представленные в таблицах 1 и 2 свидетельствуют о том, что при использовании морфологической разметки, показатель точности разметки и разрешения повышается.

Таблица 1. **Результаты работы лемматизатора инструмента PurePos 2.0**

	Без морфоанализатора	**С морфоанализатором**	**С увеличением данных**
Точность лемматизации	8767/9251 = 94.77%	9289/9510 = 97.68%	12822/13004 = 98.60%
POS-теггинг	*Нет данных*	9369/9510 = 98.52%)	12814/13004 = 98.54%
Точность определения морф.цепочки	8618/9251 = 93.16%	9141/9510 = 96.12%)	12622/13004 = 97.06%

Разрешение многозначности (только лемма)	1818/2112 = 86.08%	1947/2168 = 89.81%	2956/3138 = 94.20%
Разрешение многозначности (только POS)	*Нет данных*	2027/2168 = 93.50%	2948/3138 = 93.95%
Разрешение многознач-ности (только морфологическая цеп.)	1682/2112 = (79.64%)	1799/2168 = (82.98%)	2756/3138 = (87.83%)

Таблица 2. **Примеры разрешения многозначности инструмента PurePos 2.0**

Контекст	**Разрешаемая многозначная словоформа**		**Разрешено как**	**Коммен-тарий**
ибраһимовның чынбарлыкка романтик мөнәсәбәте ‘ «йөз ел элек» хикәяләрендә дә чагылды.	йөз	йөз+N+Sg+Nom; йөз+Num; йөз+V+IMP_SG();	йөз+Num	Разрешено правильно
аш-су осталары һәр көн саен төр сыйлар әзерләп, ханның табынына куйган.	куйган	куй+V+PCP_PS(ГАн); куй+V++PST_INDF(ГАн);	куй+V++PCP_PS(ГАн)	Разрешено неправильно, в данном контексте должно быть «куй+V++PST_INDF(ГАн); «
– юк, матурым, өмет итү, көтү, хыяллану бик ләззәтле мизгелләр.	итү	и+V+CAUS(т)++VN_1(у/ү/в)+Nom; ит+V+VN_1(у/ү/в)++Nom;	ит+V+VN_1(у/ү/в)++Nom	Разрешено правильно

мәхәббәттән әйләнгәнгәме, бу уй моңа күңеленә карамады.	уй	уй+N+Sg+Nom; уй+V+IMP_SG();	уй+N+Sg+ +Nom	Разрешено правильно
уйларым, читлеккә эләккән кош, бәргәләнергә тотынды: нишләргә, нишләргә?	эләккән	эләк+V+PCP_PS(ГАн); эләк+V+ +PST_INDF(ГАн);	эләк+V+ +PCP_PS(ГАн)	Разрешено правильно
син ренегат, безгә елан: олы юлбашчыбызның шалашында яттың, өмет итеп буарга җай булмасмылыкны син, ренегат, күптән кобра: тәүбә итеп ничә йөз тапкыр цика тирәсендә шуыштың син һәм ахырда кировны чактың (7, 152) өзек хх гасырның 20–30 нчы югарыдагы сәяси, сталинга троцкий, зиновьевларның бастыра.	елан	ела+V+REFL(Ын)+ +IMP_SG(); елан+N+Sg+Nom;	елан+N+ +Sg+Nom	Разрешено правильно
нигә күзләреңне ачы белән чылатасың?	ачы	ач+Adj+Sg+ +POSS_3(СЫ)+Nom; ачы+Adj;	ач+Adj+Sg+ +POSS_3(СЫ)+ +Nom	Разрешено неправильно, в данном контексте должно быть «ачы+Adj;»

ибраһимовның кызу темплар белән үсә һәм байый барган шикелле үк, аның әдәби-эстетик да урында катып калмадылар, интенсив төстә үстеләр һәм тирәнәйделәр.	кызу	кыз+V+VN_1(у/ү/в)+ +Nom; кызу+Adj; кызу+Adv;	кызу+Adj	Разрешено правильно
кунакларның атлары өчен лапас башына күп итеп печән китереп өйгән-нәр икән, без шул кибеп тә өлгермәгән хуш исле печәнгә палас җәеп, ике мендәр ташлап яттык.	исле	ис+N+ +ATTR_MUN(лЫ); исле+Adj;	ис+N+ +ATTR_ MUN(лЫ)	Разрешено правильно
ә сезнең соравыгызга, карыйм.	ка-рыйм	кара+V+ +HOR_SG(Йм); кара+V+PRES(Й)+ +1SG(м);	кара+V+ +PRES(Й)+ +1SG(м)	Разрешено правильно
хатыны чи-бәр, шуның матурдан-матур кәнизәкләре ханның назлы карашын көтеп, аның тирәсендә бөтерелгән.	шу-ның	шу+PN+GEN(нЫң); шу+PN+ +POSS_2SG(Ың)+ +Nom;	шу+PN+ +GEN(нЫң)	Разрешено правильно
– әй, карт, – дигәннәр аңа, – син бәхетлеме?	карт	карт+Adj; карт+N+Sg+Nom;	карт+Adj	Разрешено неправиль-но, в данном контексте должно быть «карт+N+ +Sg+Nom;»

2.2 Распознавание нераспознанных словоформ

Кроме разрешения морфологической многозначности, инструмент PurePos может быть использован для распознавания нераспознанных морфоанализатором словоформ (с тегом NR). Так, предварительные эксперименты показали, что точность при полной разметке NR (правильная разметка леммы и морфологической цепочки) составила 45% и 65% при частичной разметке, а также 79% при разметке только леммы. Примеры разметки нераспознанных словоформ приведены в таблице 3.

Таблица 3. **Примеры разметки нераспознанных словоформ**

Словоформа	**Предсказанный тег**	**Предсказанная лемма**	**Комментарий**
җылый	[V][PRES(Й)]	җылы	Словоформа относится к диалекту языка, литературной формой будет словоформа «елый». Тем не менее, тег указан верно, но лемма должна быть «җыла».
советында	[N][Sg][POSS_3(СЫ)][LOC(ДА)]	совет	Тег и лемма указаны верно
һәркайсына	[N][Sg][POSS_3(СЫ)][DIR(ГА)]	һәркайсы	Тег частично установлен верно, ошибка в части речи. В данном случае слово является местоимением(PN), а не существительным (N). Лемма указана верно.
шаһитлар	[N][PL(ЛАр)][Nom]	шаһит	Тег и лемма указаны верно
гобәйдуллин	[PROP][Sg][Nom]	гобәйдуллин	Тег и лемма указаны верно
уңаеннан	[N][Sg][ABL(ДАн)]	уңаен	Ошибочно указаны и лемма, и тег. В данном случае словоформу следует рассматривать как единое целое («уңаеннан») и размечать как POST. Также теоретически возможен вариант

			со следующим разбором: уңай+N+Sg+ +POSS_3(СЫ)+ABL(ДАн);
фәхрет-динов	[PROP][Sg][Nom]	фәхрет-динов	Тег и лемма указаны верно
саймә	[N][Sg][Nom]	саймә	Лемма указана верно, часть речи должен быть установлен как имя собственное (PROP), а не как имя существительное (N).
юхиди	[Adj]	юхиди	Лемма указана верно, часть речи должен быть установлен как имя собственное (PROP), а не как прилагательное (Adj).
коми-тетына	[N][Sg][POSS_3(СЫ)] [DIR(ГА)]	комитет	Тег и лемма указаны верно

3. Заключение

В данной работе представлены результаты работ по разрешению морфологической многозначности татарского языка с использованием инструментария PurePos 2.0. Результаты экспериментов показали достаточно высокие показатели точности для разрешения морфологической многозначности по лемме, части речи и морфологической цепочки до 96,1%, 89,8% и 93,5% соответственно. При этом необходимо отметить, что использование морфологического анализа повышает точность разрешения в среднем более, чем на 3%. Кроме того, важным условием повышения точности разрешения является качество и объем обучаемый данных, на основе которых, исследование будет продолжено.

ЛИТЕРАТУРА

1. Orosz, G. and Novák, A. 2013. PurePos 2.0: a hybrid tool for morphological disambiguation. Proceedings of Recent Advances in Natural Language Processing, pages 539–545, Hissar, Bulgaria, 7–13 September 2013. Online version: http://aclweb.org/anthology//R/R13/R13-1071.pdf

2. Orosz, G., Novák, A.: PurePos – an open source morphological disambiguator. In Sharp, B., Zock, M. (eds.) Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science, page 53–63, Wroclaw, 2012.

УДК 811.161+811.512.111]:130.3

DECISION-MAKING AS INTELLECTUAL ACTIVITY IN THE SITUATION OF CAUSE AND INVESTIGATE DEPENDENCIES (ON THE MATERIAL OF RUSSIAN AND CHUVASH LANGUAGE)

A. Gubanov, G. Gubanova
Chuvashia State University I.N. Ulyanova
alexgubm@gmail.com

The article is devoted to the correlation of the conceptual world of decision making on the basis of causal (causal) relations with the language picture of the world, reflecting the interaction of the human factor in the language. As a result of the analysis, the article states that the decision-making process is connected with the intensional state of the subject (cognitive state) of the setting-utterance with the perceptual (sensory) predicate and setting the utterance with the predicates of the thought relation; emotional state (set-statements with an axiological predicate), etc. Based on the factual material of the Russian and Chuvash languages, the conceptual space of the concept "decision-making", verbal and non-verbal composition

Keywords: language picture of the world, concept, decision-making, intellectual activity, situation, cause-effect dependence, causality, feature space of the concept "decision-making", subject's intensional state, propositional attitudes, verbal and non-verbal components, strategic levels, Russian and Chuvash languages.

ПРИНЯТИЕ РЕШЕНИЙ КАК ИНТЕЛЛЕКТУАЛЬНАЯ ДЕЯТЕЛЬНОСТЬ В СИТУАЦИИ ПРИЧИННО-СЛЕДСТВЕННЫХ ЗАВИСИМОСТЕЙ (НА МАТЕРИАЛЕ РУССКОГО И ЧУВАШСКОГО ЯЗЫКОВ)

А.Р. Губанов, Г.Ф. Губанова
Чувашский государственный университет имени И.Н. Ульянова
alexgubm@gmail.com

Статья посвящена корреляции концептуального мира принятия решения на основе причинно-следственных (каузальных) отношений с языковой картиной мира, отражающей взаимодействие человеческого фактора в языке. В результате анализа в статье констатируется, что процесс принятия решения связан с интенсиональным состоянием субъекта (когнитивное состояние) установки-высказывания с перцептивным (сенсорным) предикатом и установки высказывания с предикатами мыслительного отношения; эмоциональное состояние (установки-высказывания с аксиологическим

предикатом) и др. На основе фактографического материала русского и чувашского языков рассмотрены: признаковое пространство концепта «принятие решения», вербальные и невербальные компоненты в реализации иллокутивной цели акта принятия решения и стратегические уровни принятия соответствующего иллокутивного акта.

Ключевые слова: языковая картина мира, концепт, принятие решений, интеллектуальная деятельность, ситуация, причинно-следственная зависимость, каузальность, признаковое пространство концепта «принятие решения», интенсиональное состояние субъекта, пропозициональные установки, вербальные и невербальные компоненты, стратегические уровни, русский и чувашский языки.

Фрагменты языковой картины мира формируются на основе взаимодействия индивидуума с предметным миром, проявляются как итог осмысления феноменологического знания и маркируются однопорядковыми языковыми единицами, что подтверждается данными национальных корпусов тех или иных языков [4,6,13,14]. Корреляция концептуального мира принятия решения на основе причинно-следственных (каузальных) отношений с языковой картиной мира отражает взаимодействие человеческого фактора в языке и языкового фактора в человеке [11].Любую человеческую деятельность можно представить как цепочку принятия решений [2, 4, 5].

Принятие решений – мыслительный процесс, предполагающий предварительное осознание цели и способа действий, проработку различных вариантов. Важнейшая особенность этого процесса – его волевой характер. Процесс принятия решения во многом зависит от установки интенсионального состояния субъекта. Я. Хинтикка в одной из своих работ, посвященной пропозициональным установкам, лишь перечисляет типы установок, но не разграничивает их [9]. Мы, в свою очередь, используем таксономию установок, предложенную Н.Д. Артюновой, полагающей, что установки делятся на перцептивные (Я вижу/слышу что…), аксиологические (эмотивные) – Я люблю, боюсь…), эпистимические (Я знаю, думаю…), иллокутивные (Я скажу, говорю, что…), волитивные (Я хочу, желаю, чтобы…) [3]. Ситуации-установки отражают отношение говорящего к объективной действительности. В их семантике представлены такие этапы мыслительного акта: а) первый этап – Я слышу, вижу что…; второй этап: Я знаю, думаю, что…; третий этап: Я рад, сожалею, боюсь, что …; четвертый этап: Я говорю, скажу, что…; пятый этап: Я хочу, желаю, чтобы…

Каждый этап является предпосылкой следующего этапа, “приобретенное” при этом в первой ситуации лежит в основе второй и т.д. Правда, возможны при этом и различные варианты. Чтобы точнее определить инвариантное значение ситуаций-установок, рассмотрим их поочередно. Прежде всего отметим, что значение предикатов пропозициональных установок относится к одной референтной области – психической деятельности субъекта.

Психическая деятельность представляет собой единство трех процессов (психологической триады) – познание, чувства, воля. Эти три аспекта психической деятельности с разной степенью выраженности присутствуют в каждом акте деятельности, хотя традиционно рассматриваются отдельно: когнитивные аспекты интеллектуальных процессов отделяются от эмоциональных и волевых компонентов, собственно эмоциональные компоненты чувств абстрагируются от их когнитивных аспектов, а регуляционные функции волевых процессов отделяются от тех познавательных и эмоциональных структур, которые эту регуляцию осуществляют.

Чувственное восприятие является первой ступенью диалектического процесса познания человеком объективного мира. Под ситуацией с перцептивной установкой мы понимаем построение в сознании субъекта следующей модели реальной ситуации: Я вижу/ слышу, что… В качестве значимых элементов соответствующих установок выступают воспринимающий, воспринимаемое и отношения между ними. Субъект – воспринимающий, как и во всех других ситуациях-установках, обозначается именами лиц. «Если предложение сообщает о восприятии, существует субъект этого восприятия, и язык находит различные способы указания на него» [7; 218]. Лексико-синтаксическим ядром предиката ситуаций с перцептивной установкой являются сенсорные глаголы *слышать, видеть, смотреть, слушать*. Перцептивные предикаты по своему значению неоднородны, указывая на различные характеристики ситуаций: одни указывают на восприятие вообще, другие – на восприятие с помощью определенных органов чувств. В число обязательных аргументов данных ситуаций входит, как было выше сказано, позиция объекта: «Глаголы зрительного восприятия амбивалентны: они могут соединяться как с предметным, так и с пропозитивным объектом» [6. С. 127].

Рассмотрим установки, выражающие состояние мнения (эпистимическое состояние). Эпистимическое ментальное состояние

это такое состояние, когда выражается отношение мыслящего субъекта к содержанию суждения. Эпистимические предикаты относятся к экзистенционально-результативным предикатам. По ее мнению, этим предикатам свойственны такие признаки, как невозможность употребления в конкретно-процессном значении, отсутствие конкретной временной локализованности, описание неизменяющегося «положения вещей», не ограниченного определенным промежутком времени, наличие нахождения в пространстве. Различия между основными эпистимическими предикатами не носят денотативного характера, они не в том, что эти глаголы обозначают разные мысленные состояния… при описании одного и того же «мысленного состояния» субъекта возможно употребление разных эпистимических предикатов. Различия заключаются в коммуникативной перспективе соответствующих установок в их иллокутивной парадигматике. Эпистимические пропозициональные ситуации-установки непременно предполагают наличие субъекта. Семантическая функция этого субъекта – партиципанта не является агентивной: она является носителем определенной формы мысли. Для реализации семантики состояния мнения очень важен признак инактивности субъекта, поскольку интеллектуальное состояние не может существовать независимо от своего источника. Например, в высказываниях с глаголами знания субъект в силу значения самих глаголов представлен как пассивный обладатель информации. Иными словами, необходимы дополнительные условия контекста для нейтрализации активности субъекта.

Семантическая структура эмотивных установок складывается из следующих элементов: субъекта-носителя определенного эмоционального состояния, предиката эмоционального отношения и объекта состояния. К числу эмотивных предикатов относятся глаголы, пропозициональная семантика которых указывает на определенное психическое состояние. Таковы, в частности, глаголы, обозначающие положительное эмоциональное состояние (радоваться) и глаголы, обозначающие отрицательное эмоциональное состояние (огорчаться), также при предикатах эмоционального состояния объект понимается как источник состояния или каузатор, т.е. лицо, явление или событие, прямо или косвенно вызывающее изменение в состоянии другого лица. Одним из главных дифференциональных признаков предикатов эмотивных установок является наличие в них эмоционального оценочного значения, т.е. оценки «хорошо/ плохо».

Непосредственно связаны с эпистимическими установками иллокутивные установки, описывающие ситуации передачи и получения информации. Можно говорить о следующих семантических ролях участников иллокутивных ситуаций: субъект – говорящий (агентив), собеседник (адресат речи), объект речи (содержания высказывания). Субъект-говорящий в речевом акте является не производителем действия, а источником информации. Если же речь принадлежит первому лицу, то субъект предиката речи и говорящий совпадают (Я говорю, что…).

В тематическом отношении группа предикатов речи весьма разнообразна. Исходя из принципа коммуникативной обусловленности, предикаты речи делят на следующие группы: а) предикаты, обозначающие участие одного из говорящих лиц (говорить, рассказать, спрашивать); б) предикаты, обозначающие участие двух или нескольких лиц (беседовать, спорить, дискутировать).

Считается, что специфика коммуникативных предикатов состоит в том, что они обозначают целесообразное контролируемое действие – сознательную каузицию тех или иных ПУ у адресата коммуникации. В общем случае коммуникативный предикат описывает сложную ситуацию, в которой «Отправитель является одной или нескольких ПУ – коммуникативных намерений (= «интенций» = «коммуникативных установок») и предпринимает усилия, направленные на достижение определенного результата – перлокутивного эффекта, состоящего в том, что Адресат стал, в свою очередь, носителем тех же самых или иных ПУ» [7. С. 72].

К волитативным установкам относятся установки с типовым значением «субъект считает, что пропозиция желательно/нежелательно». Внутри волитативности различаются собственно волитативность (Я хочу) и фактитивность /Я делаю так, чтобы он работал)». Собственно волитативным установкам присущ признак «интенсиональности», указывающий на процесс формирования у субъекта некоторой интенции или намерения совершить определенное действие [8]. Для установок, указывающих на интенсиональное состояние субъекта, характерен признак «сознательная целеустановка на совершения действия».

Таким образом, следует различать установки: 1) выражающие когнитивное состояние (установки-высказывания с перцептивным сенсорным предикатом и установки высказывания с предикатами мыслительного отношения; 2) выражающие эмоциональное со-

стояние (установки-высказывания с аксиологическим предикатом); 3) выражающие интенсиональное состояние.

Следует обратить внимание на то, что некоторые установки (перцептивные и иллокутивные) называют психические акты, тогда как другие установки не называют соответствующих актов, а лишь указывают на аспект психической деятельности. Это свидетельствует, что статические типы предикатов перцептивных и иллокутивных установок занимают «промежуточное положение» между «каноническими» пропозициональными установками.

Таким образом, для вышерассмотренных установок для принятия решения оказывается существенным фактор ментальности под ментальными (семиотическими) факторами нами понимается психическое состояние субъекта, обуславливающее реализацию или нереализацию потенционального действия; во-вторых, кроме ментального фактора в установках присутствуют эндогенные факторы, т.е. любое состояние субъектом непосредственно переживается, ощущается, испытывается, и данный субъект всегда страдательный; в-третьих, важным признаком ситуаций-установок является наличие в них прессупозитивного значения.

На основе анализа фактографического материала выявлены типы пресуппозитивных фреймов, способствующих или препятствующих акту принятия решения: а) физическое или эмоциональное состояние человека: Стыд за свое неопределенное положение (она все не верила, что Григорий ушел навсегда, и, прощая, ждала его) толкнул ее на следующий поступок: решила послать тайком от домашних в Ягодное к Григорию, чтобы узнать, совсем ли ушел он и не одумался ли (Шолохов); б) мораль: Если б Григорий ходил к жалмерке Аксинье, делая вид, что скрывается от людей, если б жалмерка Аксинья жила с Григорием, блюдя это в относительной тайне, и в то же время не чуралась бы других, то в этом не было бы ничего необычного, хлещущего по глазам (Шолохов); в) чувства: С величайшей тревогой Ильинична ждала возвращения Натальи. Старику решила не говорить, боясь попреков и нареканий (Шолохов); Он было схватил с Бурдовским другой экипаж, тут же случившийся, и бросился было в погоню, но раздумал, уже дорогой, что “во всяком случае поздно! Силой не воротишь”. – Да и князь не захочет! – решил потрясенный Бурдовский (Достоевский); Он хотел решиться, но с ужасом чувствовал, что не было у него в этом случае той решимости, кото-

рую он знал в себе и которая действительно была в нем. Пьер принадлежал к числу тех людей, которые сильны только тогда, когда они чувствуют себя вполне чистыми. А с того дня, как им овладело то чувство желания, которое он испытал над табакеркой у Анны Павловны, несознанное чувство виноватости этого стремления парализировало его решимость (Толстой); г) безысходность: Егорушка стоял на коленях или, вернее, сидел на сапогах. Когда дождь застучал по рогоже, он подался туловищем вперед, чтобы заслонить собою колени, которые вдруг стали мокры; колени удалось прикрыть, но зато меньше чем через минуту резкая, неприятная сырость почувствовалась сзади, ниже спины и на икрах. Он принял прежнюю позу, выставил колени под дождь и стал думать, что делать, как поправить в потемках невидимую рогожу. Но руки его были уже мокры, в рукава и за воротник текла вода, лопатки зябли. И он решил ничего не делать, а сидеть неподвижно и ждать, когда всё кончится (Чехов); д) вдохновение:Но Пацюк взглянул и снова начал хлебать галушки. Ободренный кузнец решился продолжать: – К тебе пришел, Пацюк, дай боже тебе всего, добра всякого в довольствии, хлеба в пропорции! (Гоголь); е) крайняя необходимость: Эта находка так его обрадовала, что он позабыл все и, стряхнувши с себя снег, вошел в сени, нимало не беспокоясь об оставшемся на улице куме. Чубу показалось между тем, что он нашел дорогу; остановившись, принялся он кричать во все горло, но, видя, что кум не является, решился идти сам (Гоголь); ж) неизбежность: Генерал, принимая приглашение полковника на турнир храбрости, выпрямив грудь и нахмурившись, поехал с ним вместе по направлению к цепи, как будто все их разногласие должно было решиться там, в цепи, под пулями. Они приехали в цепь, несколько пуль пролетело над ними, и они молча остановились (Толстой); з) отсутствие выбора: - Вы не так поняли, – сказала она, – я с вами не пришла… ссориться, хотя я вас не люблю. Я… я пришла к вам… с человеческою речью. Призывая вас, я уже решила, о чем буду вам говорить, и от решения не отступлюсь, хотя бы вы и совсем меня не поняли. Тем для вас будет хуже, а не для меня (Достоевский).

Рассмотрим признаковое пространство концепта «принятие решения». Одним из главных компонентов ситуации «принятия решения» является субъект акта решения, ибо в соответствующем акте субъект может решать что-либо или принимать решение

по поводу чего-либо «за себя» и «за другого»: Я почувствовал себя утомленным, и прилег, не раздеваясь, на кровать, Я думал, что мне вовсе не удастся заснуть в эту ночь и что я до утра буду в бессильной тоске ворочаться с боку на бок, поэтому я решил лучше не снимать платья, чтобы потом хоть немного утомить себя однообразной ходьбой, по комнате (Куприн). Речевой акт решения обычно актуализируется полифонически (совместное, коллективное или индивидуальное решение). В контекстах рассматриваемых ситуаций встречаются следующие типы субъектоа решений: а) определенный субъект: За минувший день, обдумав все, твердо решил он всячески противодействовать дальнейшему продвижению сотни на Петроград; лежа, размышлял, каким образом склонить назад к своему решению, как на них подействовать (Шолохов); б) неопределенный субъект: Мы поцаловались горячо, искренно – и таким образом все было между нами решено (Пушкин); Обстоятельства, понудившие Добровольческую армию уйти из Ростова, вам известны. Вчера у нас был совет. Принято решение идти на Кубань, имея направление на Екатеринодар (Шолохов); в) метафорический субъект: Марья Ивановна предчувствовала решение нашей судьбы; сердце ее сильно билось и замирало (Пушкин).

Рассмотрим вербальные и невербальные компоненты в реализации иллокутивной цели акта принятия решения.

Ситуация принятия решения выступает как многокритериальня оценка воспринимаемых явлений объективной действительности, как оценка результатов действий субъекта или оценка альтернатив. Встраивание процедуры оценки регулирующего воздействия в процесс принятия решения достигается базовым предикатом «решить» (шутлама) – сделать выбор: В первый же день, как только болезнь свалила девочку с ног, Аксинье вспомнилась горькая Натальина фраза: «Отольются тебе мои слезы…» и она решила, что это ее бог наказывает за то, что тогда глумилась над Натальей (Шолохов); Предикаты типа решить в своей семантической структуре имеют рациональный оценочный компонент. Выражение непроизвольности установки, столь характерное в целом для семантики предикатов, выражающих эмоциональную оценку, оказывается застертым для данных предикатов. Ср.: Услышав о вашем прибытии и узнав, что вы изволили отправиться на берега нашего пруда, решил, если вам не будет

противно, предложить вам свои услуги (Тургенев); Надо было видеть, как одни, презирая опасность, подлезали под нее, другие перелезали через, а некоторые, особенно те, которые были с тяжестями, останавливались, искали обхода, или ворочались назад, или по хворостинке добирались по моей руке и, кажется, намеревались забраться под рукав моей курточки (Толстой). Эквивалентами предикатов рациональной оценки в чувашском языке выступают аналогичные по семантике русским предикатам глаголы: *шухăшла, шутла, тăрăш, хăтлан, хатĕрлен, пуçтарăн, пăх, пар,* но первая группа эквивалентов сочетается с формой инфинитива на *–ма/-ме*, а вторая – с деепричастием на *-са/-се*. Например: 1) Вначале Пантелей Прокофьевич думал даже повозку везти на санях (Шолохов). – Пантелей Прокофьевич малтанах çуна çине карман лартса кайма та шутланăччĕ; Леонтий решил оторваться от земли (Садай). – Леонтий çĕртен хăпма шутларĕ; 2) Пробовал даже свернуть из лоскутков куклу, но тут у него ничего не вышло (Шолохов). – Çитсă татăкĕнчен пукане тума та хăтланса пăхнăччĕ, анчах кунта нимĕн те тухмарĕ; А потом, на уроке вышивания попробуйте вышить на холсте (Артемьев). – Кайран вара, тĕрлемелли урокра тĕрлесе пăхăр; Вот, попробуйте-ка! (Садай). – Кăларса пăхар-ха! На вашей свадьбе, если она даже завтра, все-равно сплясать попробую (Садай). – Сирĕн туйăрта, вăл час пулас пулсан та, ташласа та пăхăп-ха. В случаях выражения в русском языке оттенков рационально-оценочного значения при помощи предиката намереваться в чувашском языке употребляется трехкомпонентный эквивалент, в состав которого входят: будущее причастие на -ас(-ес), деепричастие на -са(-се) (образованный от эквивалента предикату состояния хотеть – те и глагол шутла (решить): Зная упрямство дядьки моего, я решил убедить его лаской и искренностью (Пушкин). – Савельич хăй тăвас тенине çав тери тума юратнине, тӳрккессине пĕлсе тăнипе, чунтан каласа ӳкĕте кĕртес тесе шутларăм; В случаях, когда в русском языке предикат выражается глаголом решился: чувашский эквивалент выражается следующей конструкцией: инфинитивом на *-ма/-ме,* существительное *хал* и фазовый глагол çитер: Я решился последовать совету Зурина, оправить Марью Ивановну в деревню и остаться в его отряде (Пушкин). – Зурин каланă пек тума шутларăм эпĕ, Марья Ивановна яла ярас та хам ун отрядĕнче юлас терĕм; Как известно, между семантическими компонентами ситуации и компонентами

поверхностной семантической структуры складываются отношения соответствия. Однако в ряде случаев типичная денотативная ситуация может подвергаться другой форме актуализации, которая обусловлена номинализацией вторичной денотативной ситуации, вследствие чего номинализованным трансформом глагольного предиката выступает сочетание, в состав которого входят девербативный (пропозитивное имя) и вербальный (или нулевой) компоненты. Ср.: решил поехать – принял решение все-таки поехать. В результате такого изменения в семантической структуре рассматриваемых предложений складываются следующие отношения: отношения экзистенции, рефлексивные отношения и отношения идентификации. В предложениях, где коммуникативный акцент делается на показателе бытия, характерно употребление показателя экзистенциональности, а также пропозитивного имени. Они характеризуются скользящим коммуникативным фокусом, который может падать на экзистенциональный показатель или на имя бытующей «пропозиции»: Как показывает языковой материал, пропозитивное имя может выражаться: а) лексемой «решение» (йышӑну), «решимость» как субстанция в результирующей стадии: Окружной мировой судья 7-го участка, по указу Его Императорского Величества, приказал: всем местам и лицам, до коих сие может относиться, исполнить в точности настоящее решение, а властям местным, полицейским и военным оказывать исполняющему решение приставу надлежащее по закону содействие вез малейшего отлагательства (Шолохов); Ещё в поезде она приняла твёрдое решение никогда больше сюда не возвращаться; варианты характеристик решения: Деланное оживление Натальи потухло искрой на ветру. Бабы перекинулись в разговоре на последние сплетни, на пересуды. Наталья вязала молча. С трудом высидев до конца, она ушла, унося в душе неоформленное решение (Шолохов); б) лексемой «решительный» (хӑюлли): Натыкаясь на решительный отказ, он иногда обижал ее резким словом: – Ты не имеешь права так издеваться надо мной! (Шолохов); в) краткими формами причастий типа «решено», прилагательных типа «готов»: Решено было в неделю проводить беседы (Шолохов). Предикатное выражение данных конструкций в чувашском языке передается эквивалентом, в состав которого входят: а) инфинитив на -ма (-ме), деепричастие на -са(-се) от глагола шутла, глагол хур (положить): Решено было держаться и впредь в университете,

несмотря на то, что в этом случае я в первый раз расходился во мнениях со своим другом (Толстой). – Ку тĕлĕшпе мана пĕрремĕш хут хамăн тусăм пек мар шухăшлама тиврĕ пулин те, малашне эпĕ университетра çавнашкал пулма шутласа хутăм; б) инфинитив на ма(-ме): вспомогательный глагол пул (быть): Ими решено приехать сюда на подмогу (Ажаев). – Вĕсем кунта пулăшма килме пулчĕç; (особенность данных конструкций заключается в том, что они выражают такое волевое состояние субъекта, когда данный субъект знает предшествующее «положение дел» и не сомневается в претворении в жизнь потенциального действия; интенциональные конструкции с предикатами, выраженные краткими прилагательными типа готов в чувашском языке, как и в русском, обозначают такую статичную ситуацию, которая выступает как прямой результат сознательного действия или спонтанного субъекта (эквивалентом названных конструкций в чувашском языке выступает конструкция с *хатĕр* в сочетании с инфинитивом на *-ма(-ме):* Однажды, когда нам удалось как-то рассеять и прогнать довольно густую толпу, наехал на казака, отставшего от своих товарищей, и я готов был ударить его своею турецкою саблею, как вдруг он снял шапку и закричал… (Пушкин). – Пĕрре тепле майпа эпĕ юлташĕсенчен тăрса юлнă пĕр казак çине çитсе тăрăнтăм, эпĕ ăна алăри турка хĕççипе пĕррех туртса касма хатĕрччĕ. Как видно из примеров, в чувашском языке в данном случае эквивалентной формой выступает, как и в русском языке, пропозитивные имена.

А теперь рассмотрим невербальные компоненты в реализации перлокутивного эффекта в акте принятия решения. Компонентами дополнительного модуля рассматриваемого дискурса выступают такие родовые понятия, как: а) вид, движение (действие), поза: Вдруг Мерцалов быстро поднялся с сундука, на котором он до сих пор сидел, и решительным движением надвинул глубже на лоб свою истрепанную шляпу (А. Куприн); б) голосовые акустические характеристики, губы: Он был меньше среднего роста, сухой, жилистый, очень сильный. Лицо его, с покатым назад лбом, тонким горбатым носом и решительными, крепкими губами (А. Куприн); в) лицо, глаза, взгляд: Полковой командир, покраснев, подбежал к лошади, дрожащими руками взялся за стремя, перекинул тело, оправился, вынул шпагу и с счастливым, решительным лицом, набок раскрыв рот, приготовился крикнуть. Полк встрепенулся, как оправляющаяся птица, и замер (Толстой).

Рассмотрим стратегические уровни принятия оптимального решения. Анализ текстовых фрагментов «принятия решений» позволяет вычленить различные модели-стратегии тактических действий: 1) стратегии иллокутивного типа: а) «совет»: Вот что, Пантелей Прокофьевич, – начал хозяин, переглянувшись с женой, – посоветуйте вы, и мы посоветуем промеж себя, семейно. А потом уж и порешим дело, будем мы сватами аль не будем (Шолохов); б) «внушение»: В любовных делах, а особенно в женитьбе, внушение играет большую роль. Все – и товарищи и дамы – стали уверять Беликова, что он должен жениться, что ему ничего не остается в жизни, как жениться; все мы поздравляли его, говорили с важными лицами разные пошлости, вроде того де, что брак есть шаг серьезный; к тому же Варенька была не дурна собой, интересна, она была дочь статского советника и имела хутор, а главное, это была первая женщина, которая отнеслась к нему ласково, сердечно, – голова у него закружилась, и он решил, что ему в самом деле нужно жениться (Чехов); в) «предупреждения»: Усевшись рядом с нею, я почувствовал чрезвычайную неловкость и решительно не знал, о чем с ней говорить. Когда молчание мое сделалось слишком продолжительно, я стал бояться, что бы она не приняла меня за дурака, и решился во что бы то ни стало вывести из такого заблуждения на мой счет (Толстой); е) «убеждения»: Это еще и хорошо, что я знаю, что не скрыто от меня это; а то бы я умерла от подозрений. Да, Ваня! Я уж решилась: если я не буду при нем всегда, постоянно, каждое мгновение, он разлюбит меня, забудет и бросит. Уж он такой; его всякая другая за собой увлечь может (Достоевский); 2) стратегии – прагматические коннотации: а) «выяснения»: Сергей Платонович спал ночью плохо, ворочался, одолеваемый бестолковыми мыслями и неосознанными желаниями; уснул за полночь, а утром, прослышав, что в Ягодное приехал с фронта к отцу Евгений Листницкий, решил съездить туда, чтобы поговорить, выяснить подлинное положение и снять с души горькую накипь тревожных предчувствий. Емельян, посасывая трубку, запряг в городские сани маштака, повез хозяина в Ягодное (Шолохов); б) «расчет»: Как ни велико было желание поехать домой, он решил остаться в Новочеркасске, чтобы отдохнуть, не теряя времени на переезды (Шолохов); Мать поняла, что надо бежать, и свистнула. Но он еще не понимал или был слаб. Она попробовала подтолкнуть его в спину губами. Он покачнул-

ся. Она решила обмануть собаку, чтобы та за ней погналась, а теленка уложить и спрятать в траве. Так он и замер в траве, весь осыпанный и солнечными и своими «зайчиками». Мать отбежала в сторону, встала на камень, увидела Тайгу. Чтобы обратить на себя внимание, она громко свистнула, топнула ногой и бросилась бежать (Пришвин). в) «поиск альтернативы»: Большая белая собака, смоченная дождем, с клочьями шерсти на морде, похожими на папильотки, вошла в хлев и с любопытством уставилась на Егорушку. Она, по-видимому, думала: залаять или нет? Решив, что лаять не нужно, она осторожно подошла к Егорушке, съела замазку и вышла (Чехов); г) «риск»: Он тотчас стал одеваться, сердито сопя. Ленька извинился и вышел. Он уже успел разглядеть в дверную щелку лицо Рыбникова, и хотя у него оставались кое-какие сомнения, но он был хорошим патриотом, отличался наглостью и не был лишен воображения. Он решил действовать на свой риск (Куприн); д) «преодоления препятствий»:Дома Егор ходил из угла в угол, что-то обдумывая. Курил. Время от времени принимался вдруг напевать: «Зачем вы, девушки, красивых любите?» Бросал петь, останавливался, некоторое время смотрел в окно или в стенку… И снова ходил. Им опять овладело какое-то нетерпение. Как будто он на что-то такое решался и никак не мог решиться. И опять решался. И опять не мог… Он нервничал (Шукшин); е) «игры»: Я не оттого плачу, что я низок и подл, но оттого, что через меня Наташа будет несчастна. Ведь я оставляю ее на несчастье… Ваня, друг мой, скажи мне, реши за меня, кого я больше люблю из них: Катю или Наташу? (Достоевский); «уступки»: Аксинья знала, о чем он думает. Помочь ему она ничем не могла. Она сама страдала, видя, как ему тяжело, и догадываясь о том, что надежды ее на совместную жизнь снова становятся несбыточными. Она ни о чем его не спрашивала. Пусть он решает все сам (Шолохов); к) «упрек»: – А что ж! – подхватил он вдруг, как будто раздраженный нашим молчанием, – чем скорей, тем лучше. Подлецом меня не сделают, хоть и решат, что я должен заплатить. Со мной моя совесть, и пусть решают. По крайней мере дело кончено; развяжут, разорят… Брошу все и уеду в Сибирь (Достоевский); л) «утешения»:Она как будто совсем не хотела говорить со мной, точно я перед ней в чем-нибудь провинился. Мне это было очень горько. Я даже сам нахмурился и однажды целый день не заговаривал с нею, но на другой день мне стало

стыдно. Часто она плакала, и я решительно не знал, чем ее утешить. Впрочем, она однажды прервала со мной свое молчание (Достоевский); м) «самоутешения»: Голова решился молчать, рассуждая: если он закричит, чтобы его выпустили и развязали мешок, – глупые дивчата разбегутся, подумают, что в мешке сидит дьявол, и он останется на улице, может быть, до завтра (Гоголь); н) «самооправдания»: Защити меня, спаси; передай им все причины, все как сам понял. Знаешь ли, Ваня, что я бы, может быть, и не решилась на это, если б тебя не случилось сегодня со мною! Ты спасение мое: я тотчас же на тебя понадеялась, что ты сумеешь им так передать, что по крайней мере этот первый-то ужас смягчишь для них (Достоевский); о) «самопожертвования»: Начал я с того, что решил пожертвовать в пользу голодающих пять тысяч рублей серебром. И это не уменьшило, а только усилило мое беспокойство. Когда я стоял у окна или ходил по комнатам, меня мучил вопрос, которого раньше не было: как распорядиться этими деньгами? (Чехов); – Меня чрезвычайно заботит теперь одно обстоятельство, Иван Петрович, – начал он, – о котором я хочу прежде всего переговорить с вами и попросить у вас совета: я уж давно решил отказаться от выигранного мною процесса и уступить спорные десять тысяч Ихменеву. Как поступить? (Достоевский); п) «стратегии вежливости»: Наташа однажды заболела, даже чуть было не пошла к ней сама. Но это был крайний случай. Сначала она даже и при мне не решалась выражать желание увидеться с дочерью и почти всегда после наших разговоров, когда, бывало, уже все у меня выспросит, считала необходимостью непременно подтвердить, что хоть она и интересуется судьбою дочери, но все-таки Наташа такая преступница, которую и простить нельзя (Достоевский).

Таким образом, для принятие решения оказывается существенным фактор ментальности (психическое состояние субъекта, обуславливающие реализацию или нереализацию потенционального действия -наличие в них прессупозитивного значения. Одним из главных компонентов ситуации «принятия решения» является субъект акта решения, способный решать что-либо или принимать решение в той или иной ситуации: В реализации иллокутивной цели акта принятия решения присутствуют вербальные и невербальные компоненты, которые соотносимы с теми или иными стратегическими уровнями принятия оптимального решения.

ЛИТЕРАТУРА

1. Амиров А. Обстоятельства причины и цели в современном узбекском литературном языке: автореф. дис. … канд. филол. наук. – Самарканд, 1967. – 20 с.

2. Арутюнова Н.Д. Предложение и его смысл: Логико-семантические проблемы. – М.: Наука, 1976. – 285 с.

3. Арутюнова Н.Д. Типы языковых значений: оценка, событие, факт. – М., 1988. – 280 с.

4. Губанов А.Р.Морфологический стандарт для систем автоматической обработки текстов на чувашском языке и архитектура грамматического словаря / Актуальные вопросы истории и культуры чувашского народа. Чебоксары, 2015. С. 146–161.

5. Губанов А.Р., Губанова Г.Ф., Свеклова О.В.Тезаурус чувашского языка (чăваш пĕлĕвĕн мулĕ) как языковая система знаний // Вестник Чувашского университета. № 2. 2017.

6. Губанов А.Р., Каюмова Д.Ф.Ключевые концепты культурно-национального мировидения через единицы фразеологического уровня английского и тюркских языков// Проблемы и перспективы развития многоуровневой языковой подготовки в условиях поликультурного общества материалы II межрегиональной заочной научно-практической конференции. Казанский государственный университет культуры и искусств. Казань 2015. С. 160–165.

7. Диев В.С. Управление. Философия. Общество // Вопросы философии.2010. № 8).

8. Золотова Г.А. Очерк функционального синтаксиса русского языка. – М.: Наука, 1973. – 351 с.

9. Левонтина И.Б. Целевые слова и наивная теология. Автореф. дис. … канд. филол. наук. – М., 1995. – 22 с.

10. Смурова Л.И. Вариантность финальных (целевых) синтаксем (на материале английского языка). Автореф. дис. канд. филол. наук. – Санкт-Петербург, 1995. – 21 с.

11. Рец Н.И., Губанов А.Р.Категория причинности: каузальная и каузативная связи//Вестник Чувашского университета. 2012. № 1. С. 244–249.

12. Хинтикка Я. Логико-эпистемологические исследования. – М.: Прогресс, 1980. – 448 с.

13. Zheltov P.V., Zheltov V.P., Gubanov A.R.Automation of lexical search in national corpora of chuvash language: methods of exploring space of literary texts/Russian linguistic Bulletin. 2016. № 3 (7). С. 58–60. 14. Zheltov P.V., Zheltov V.P., Gubanov A.R.Text analysis subsystem in a search engine for the national corpora of the chuvash language/Russian linguistic Bulletin. 2016. № 3 (7). С. 61–63.

УДК 811.512.145

SOME WORDS ABOUT MORPHOLOGICAL CLASSIFICATION OF WORDFORMS WITH AH INDEX -GAN IN THE TEXTS IN THE TATAR LANGUAGE

M. Dubrovina
Sankt-Petersburg State University, Sankt-Petersburg, Russia,
maggydu@rambler.ru

From a formal point of view, -gan is a morphological index, which comes after the root morpheme of the verb in the linear range of the formants within word forms. However, even a superficial analysis of texts in Tatar language indicates that this formant is used very often. There is a need for adequate morphological characterization of the index for efficient processing of text material. The author, doing the description and analysis of the semantics of practical applications, proposes a classification of -gan – like morpheme of the different homonymous forms, which are included in the different inflectional categories of the morphological subsystem of the Tatar language.

Keywords: form -gan participle with -gan, the morphology of the Tatar, the Tatar language.

О МОРФОЛОГИЧЕСКОЙ КВАЛИФИКАЦИИ СЛОВОФОРМ С ПОКАЗАТЕЛЕМ -ГАН В ТЕКСТАХ НА ТАТАРСКОМ ЯЗЫКЕ

М. Э. Дубровина
Санкт-Петербургский Государственный Университет,
Санкт-Петербург, 199034, Россия
maggydu@rambler.ru

С формальной точки зрения -ган это морфологический показатель, который в линейном ряду формантов словоформы следует за корневой морфемой смыслового глагола. Однако даже поверхностный анализ текстов на татарском языке свидетельствует о весьма частотном употреблении этого форманта. Для эффективной обработки текстового материала необходима его морфологическая квалификация. Автор через описание и анализ семантики речевых/текстовых употреблений предлагает классификацию -ган, как морфемы разных омонимичных форм, входящих в различные словоизменительные категории морфологической подсистемы татарского языка.

Ключевые слова: форма -ган, причастие на -ган, татарская морфология, татарский язык.

1. Введение

Любое научное лингвистическое исследование, как правило, опирается на определенную теоретическую базу, позволяющую рассматривать языковые факты под определенным теоретическим углом. На наш взгляд, перспективным направлением в языкознании может быть признан функционально-семантический подход к анализу языковых явлений. Перспективным этот метод является по причине того, что благодаря входящим в него установкам становится возможным объяснить функциональную специфику обнаруживаемых в языке форм, четко обрисовать их языковую значимость и отличительные особенности, вследствие чего предложить последовательную и стремящуюся к объективности классификацию.

2. Функционально-семантический подход

Функционально-семантический подход предполагает в качестве основы функциональную направленность грамматики и принцип, согласно которому объективное познание языка и его составляющих осуществимо не путем формального описания, а посредством изучения языка как функциональной системы. При таком подходе на первый план выходит изучение функций единиц языка и закономерности их функционирования. Поскольку функция представляет собой внешнее проявление свойств объекта в конкретной системе отношений, то совершенно естественно, что именно понятие **«функция»** является отправной точкой исследований в области функциональной грамматики. (Бондарко, 1999) В самом общем виде под **функцией** языковых средств подразумевается свойственная им в языковой системе способность к выполнению определенного назначения и к реализации этого назначения в речи. (Ахманова, 1999)

Путем выявления функции каждой конкретной единицы языка становится возможным определить место этой единицы, ее «относительную значимость» внутри всей языковой системы. (Матезиус, 1967) При описании языкового материала в рамках функциональной грамматики были выработаны две взаимодополняющие методологические установки: «от семантики к ее форме» («от функций к средствам») и «от формы к семантике» («от средств к функциям»). (Бондарко, 1988) Подобный двусторонний

подход к описанию языковых средств обусловлен, прежде всего, тем, что определенная функция может быть реализована разными языковыми средствами и, с другой стороны, одно и то же средство может обладать и обычно обладает разными функциями. На наш взгляд, именно это обстоятельство имеет место когда речь идет о татарской форме с показателем -ган.

3.Функции формы с показателем -ган

Рассматриваемая форма в речи имеет следующие алломорфы: -ган /-гән/-кан/-кән. Обращаясь к национальному корпусу татарского языка, исследователь обнаруживает функциональную нерасчленённость рассматриваемой формы. Однако, фактический материал текстов на татарском языке демонстрирует, что словоформы с показателем -ган имеют различную семантику и различную дистрибуцию.

3.1. *Форма -ган в текстах способна выступать в качестве определения*

1. *Өч улының кайсысы булса да кайчан да булса бер **килгән** поезддан төшеп авыл юлына чыган бит (ӘЕ, Х, 6). «Ведь хоть кто-нибудь из его трех сыновей хоть когда-нибудь сойдет с прибывшего поезда и направится к деревне».*

2. *...рәхимсез нужада торып **укыган** ярлы шәкерт гадәттә бик сәләтле булып чыга. (ӘЕ, М, 1) «Бедный ученик, **который учился**, находясь в жестокий нужде, обычно становится очень талантливым».*

3.2. О термине «причастие»

Традиционно в татарском языкознании эту форму в такой дистрибуции принято называть причастием. Однако существует мнение, что эта форма не только в татарском, но и в других тюркских языках, ведет себя не как чистое причастие. (Дубровина, 2016). Автор данной статьи вслед за многими другими тюркологами принимает следующую позицию: форма -ган не представляет собой причастия. Она, обладая специфическими чертами, не свойственными европейским причастиям, должна получить самостоятельный статус и самостоятельное название. Во-первых,

как писал С.Н. Иванов о турецких формах, что соответствует и татарской форме -ган, определяемое при форме -ган может обозначать не только субъект или объект действия, (как это имеет место у причастий русского языка), но также место и время действия. (Иванов, 1977).

3. *Ә монда, колхозның атаклы кешеләре* ***җыелган*** *бу йортта зур җәмәгать эшләре хәл ителә. (АШ, СӘ, 6). «А здесь, в этом доме,* ***где собирались*** *видные деятели колхоза, решаются важные общественные проблемы».*

Кроме того важной особенностью это формы является то, что она индифферентна к «агентивному» значению, т.е. отношения между действием и определяемым предметом могут быть различными, а зачастую они становятся известными только на основании всего контекста:

4. *һәм аның тавышында* ***мин көткән*** *уңайсызлану яки оялу ник кенә бер ишетелсә иде! (http://kitap.net.ru/eniki/6.php). «Ах, если бы в ее голосе было слышно то, что* ***я ожидал услышать****: чувство неловкости и стыда».*

Текстовой материал и теоретические изыскания подводят нас к выводу о том, что форма -ган и аналогичные ей формы в других тюркских языках представляют собой не аналоги европейских причастий, а особые **двуфункциональные** морфологические средства, в задачу которых входит выступать не только в роли глагольных атрибутивов, но и в роли глагольных субстантивов.

5. *Җир шарында бер хатын да Гөлчирә* ***күргәнне*** *күрмәсе! «Пусть не увидит ни одна женщина* ***того, что увидела*** *Гульчира».* (https://cyberleninka.ru/article/n/pridatochnye-predlozheniya-podchinennye-glavnomu-pri-pomoschi-padezhnyh-affiksov-na-materiale-tatarskogo-yazyka).

В представленном примере словоформа **күргәнне** включает в себя аффикс -гән. По традиции подобные случае в татарском языкознании рассматриваются как формы придаточных предложений, осложненных падежными окончаниями. Однако в теоретических исследованиях по тюркской грамматике уже были озвучены идеи, согласно которым подобные формы не следует воспринимать как финитные формы. Задачей таких средств является передать не предикат высказывания, а глагольное дополнение, т.е. в сущности в данном случае форма -ган выражает не предикативные отношения, а подчинительные (отношения дополнения). Таким образом,

в случае если после аффикса -ган следует аффикс одного из падежей, необходимо воспринимать эту форму как инфинитную, выступающую в качестве глагольного субстантива.

Две выше перечисленные функции: атрибутивная и субстантивная относятся к одной и той же форме, которую едва ли оправданно именовать причастием. Следовательно, предлагается ввести в научный оборот новый термин **«субстантивно-адъективная форма»**, сокращением которой является аббревиатура САФ, который активно используется учеными петербургской научной школы. (Гузев, 1976; Дениз-Йылмаз, 2006; Дубровина, 2016) В результате чего в качестве добавленного глосса для обозначения данных форм был введен знак SAF (substantive-adjective form/ субстантивно-адъективная форма).

3.3. Функция формы -ган в качестве «субстантивно-адъективной формы»

В высказываниях форма -ган в подобной дистрибуции выступает либо в роли подлежащего, либо в качестве второстепенных членов предложения: прямого, косвенного дополнений.

6. *Кайдадыр балаларның шатлыклы шаулашып **уйганнары** ишетелә. (Афзал Шамов. Сайланма Әсәрләр. Казан, 1954. С. 188). «Слышно, **как** где-то радостно крича, **играют** дети».*

7. *Нилəр **уйлаганым**, башыма нинди "этлекләр" **килгәнен** һич кенә дә искә төшерә алмыйм. Ә инде мич алдыннан шырпы алып кесәгә **яшергәнемне** яхшы хәтерлим. (http://kitap.net.ru/minnullin2.php). «Я никак не могу вспомнить, **о чем я думал**, какие «проделки» пришли мне в голову. Зато я прекрасно помню, **как я взял** из-за печи спички и спрятал их в карман».*

3.4. Функция формы -ган в качестве временной формы

В текстах словоформы с аффиксом -ган представляют собой форму, входящую в категорию времени, прежде всего, в том случае, если эти словоформы стоят в конце высказывания. Основаниями для такого вывода, необходимо считать заключения, основанные на сравнительно-историческом анализе фактического материала других тюркских языков, в которых аналогичная форма также отпочковалась от глагольно-именных форм и стала самостоятельной формой времени.

8. *Кызым барысын да сөйләде: ул кибеттән чыгып килгәндә* ***очрашкансыз.*** *(Информант) «Дочка мне все рассказала,* ***вы встретились****, когда она выходила из магазина».*

Таким образом, с синхронии современного татарского временная форма -ган, будучи омонимичной формой глагольно-именной форме, САФу -ган, имеет самостоятельное морфологическое окружение и свое значение.

3.5. Дистрибуция временной формы -ган:

1. Конечная позиция в высказывании
2. Употребление после аффикса -ган аффиксов сказуемости:

Ед. число	Мн. число
1 л. -мын/-мен	-быз/-без
2 л. -сың/ -сең	-сыз/-сез
3 л. -	-нар/ -нәр ң

Все формы времени составляют подкатегорию времени, которая в свою очередь, входит в категорию сказуемости. Таким образом, формы времени представляют собой финитные формы. Согласно последним разработкам в области теоретического тюркского языкознания, финитная форма это специализированное, морфологическое средство выражение мысли двучленной субъектно-предикатной структуры, т.е. суждения. С помощью татарской формы времени -ган коммуникант передает информацию одновременно и о субъекте высказывания, который на морфологическом уровне выражается личным показателем сказуемости, и о предикате – выраженном посредством самого глагола и грамматического маркера -ган.

3.6. Значение временной формы -ган

Значение этой формы складывается из двух составляющих: в значение присутствует 1) указание на то, что действие предшествует настоящему периоду (т.е. воспринимается как прошедшее), 2) модальная сема опосредованности, сигнализирующая о том, что информация получена из каких-то источников (пересказывательность, заочность, умозаключение).

В конкретных речевых высказываниях словоформы с временным показателем -ган могут иметь многочисленные конкретные смыслы, самыми частыми из которых являются следующие:

а). Говорящий не был очевидцем действия и либо пересказывает его, либо судит о нем со слов других людей:

9. *Кызым барысын да сөйләде: ул кибеттән чыгып килгәндә* ***очрашкансыз.*** *(Информант) «Дочка мне все рассказала,* ***вы встретились****, когда она выходила из магазина».*

б). Результат действия, совершенного в прошлом самим говорящим, но не осознанного им в момент его совершения, обнаруживаемого лишь после, по его результату:

10. *Ялгыш тукталышта* ***чыкканмын****, шуңа күрә адашып соңга калдым. (Информант) «Я* ***сошел*** *не на той остановке, поэтому заблудился и опоздал».*

3.7. Функция формы -ган в качестве плюсквамперфекта

Следующей формой, которая образуется на базе морфемы -ган, является форма плюсквамперфекта (преждепрошедшего времени). В этом случае обязательным грамматическим маркером выступает аналитический показатель ***иде.*** Таким образом, морфема плюсквамперфекта (преждепрошедшего времени) представляет собой сложное, но *единое, неразделяемое образование*: **-ган иде**. Аналитический показатель присоединяет к себе личные аффиксы. В речи словоформы с показателем *-ган иде* способны употребляться для выражения большого количества конкретных смыслов, одним из которых является следующий:

Форма передает действие, завершившееся раньше другого действия, которое совершилось в прошлом:

11. *Миң бу яңалыкны радиодан* ***ишеткән идем*** *инде. (Информант) «На тот момент я* ***уже слышал*** *эту новость по радио».*

3.8. Функция формы -ган в сложной морфеме 0с обстоятельственным значением

Формант -ган входит в морфемы самостоятельных форм с обстоятельственным значением. Полагаясь на татарские тексты и речь информантов, а также на сравнителные исследования по тюркским языкам, можно с уверенностью утверждать, что такие показатели как **-ганда, -ганчы** уже отпочковались от морфемы -ган и превратились в маркеры самостоятельных морфологических форм. Опираясь на результаты теоретических исследований, становится понятным, что данные формы представляют со-

бой деепричастия и составляют одну из категорий инфинитных глагольно-именных форм.

3.8.1. Форма -ганда/-гәндә/-канда/-кәндә

12. *Кызым барысын да сөйләде: ул кибеттән чыгып* ***килгәндә*** *очрашкансыз. (Информант). «Дочка мне все рассказала, вы встретились,* ***когда*** *она* ***выходила*** *из магазина».*

13. *– Мин* ***яшь булганда*** *бик көчле идем, – ди бер карт. Мин* ***барганда****, кешеләр «Ай-һай», дип әйтәләр иде. '****Когда*** *я* ***был молодым****, я был очень сильным,- говорит один старик. «Когда я шел, люди говорили «ах-ах!'.*

Исследуемая форма имеет обстоятельственное значение времени, соотносимое со значением русских союзов «когда», «в то время, когда».

3.8.2. Форма -ганчы/-гәнче/-канчы/-кәнче

14. *Анда* ***барганчы****, мин концертка барам. (Информант) «****До того, как пойти*** *туда, (вместо того, чтобы пойти туда), я пойду на концерт».*

Данная форма представляет собой деепричастие с временным значением, близким по семантике русским подчинительным союзам «до того, как», «вместо того, чтобы», «в соответствии с тем, что».

С позиций функциональной грамматики, показатели -ганда и -ганчы представляют собой неразложимые в синхронии языка образования, так как являются маркерами самостоятельных по семантике и функционированию морфологических форм. Эти формы входят в категорию деепричастий, несут обстоятельственные значения и должны идентифицироваться отдельно. В соответствии с принятым глоссированием при разметке данных форм можно применить знак AP (ADV.PTCP – adverbial participle/ деепричастие) или знак CNV (converb).

Вывод

Опираясь на результаты, полученные в ходе различных практических и теоретических исследований в области тюркского языкознания, допустимо утверждать, что в татарском языке обнаруживается несколько форм с омонимичным аффиксом -ган, которые входят в разные грамматические категории. Это обстоятельство необходимо учитывать при составление корпуса татар-

ского языка, путем помещения имеющихся примеров употребления данной формы в разные главы с различной морфологической разметкой.

ЛИТЕРАТУРА

1. Ахманова О.С. Словарь лингвистических терминов. М., Издательство «Советская энциклопедия» 1999.

2. Бондарко А.В. (1988) Направления функционально-грамматического описания “от формы” и “от семантики”//Функциональный анализ грамматических форм и конструкций. Л., ЛГПИ им. А.И. Герцена.

3. Бондарко А.В. (1999) Основы функциональной грамматики. Языковая интерпретация идеи времени. СПб., Изд-во СПбГУ.

4. Гузев В.Г. (1976) Система именных форм тюркского глагола как морфологическая категория (на материале староанатолийского и турецкого языков) //Turcologica. К семидесятилетию академика А.Н. Кононова. Л.: Издательство Наука.

5. Дениз-Йылмаз О. (2006) Категория номинализации действия в турецком языке. СПб.: Издательство С.-Петербургского ун.-та.

6. Дубровина М.Э. А. (2016). Об отличительных особенностях глагольно-именной формы с показателем -gan в узбекском языке// Вестник Санкт-Петербургского университета. Серия 13. Востоковедение, африканистика. Вып.3. Сентябрь. 2016. СПб.: Издательство СПбГУ.

7. Дубровина М.Э. Б. (2016). О термине «субстантивно-адъективная форма» (САФ) применительно к некоторым глагольно-именным формам тюркских языков //Актуальные вопросы тюркологических исследований / под ред. Н.Н. Телицина, Й.Н. Шена. СПб.: СПбГУ.

8. Иванов С.Н. (1977). Курс турецкой грамматики. Ч. 2: Грамматические категории глагола. JL: Издательство ЛГУ.

9. Матезиус В. (1967). О системном грамматическом анализе // Пражский лингвистический кружок. Сборник статей М., Издательство «Прогресс».

УДК 81’32

IDENTIFYING THE TONALITY OF TEXTS IN THE KAZAKH LANGUAGE ON THE BASIS OF THE DICTIONARY OF EMOTIONAL LEXIS

B. Ergesh
L.N. Gumilyov Eurasian National University
saturn_banu@mail.ru

The paper describes the sentiment analysis of texts in the Kazakh language based on linguistic methods. For this purpose, the dictionary of sentiment words of the Kazakh language has been created and the rules for determining tonalities using morphological categories have been formalized.

Keywords: Kazakh language; sentiment analysis; emotional vocabulary; the dictionary of sentiment words.

ОПРЕДЕЛЕНИЕ ТОНАЛЬНОСТИ ТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ НА ОСНОВЕ СЛОВАРЯ ЭМОЦИОНАЛЬНОЙ ЛЕКСИКИ

Б.Ж. Ергеш
Евразийский национальный университет имени Л.Н. Гумилева, Астана, 010000, Казахстан
saturn_banu@mail.ru

В статье описывается анализ тональности текстов на казахском языке на основе лингвистических методов. Для этого создан словарь лексических тональностей казахского языка и формализованы правила определения тональностей с помощью морфологических категорий.

Ключевые слова: казахский язык; сентимент анализ; словарь эмоциональной лексики.

1. Введение

Определение тональности текста или сентимент анализ это одно из интереснейших и сложных задач обработки естественного интеллекта. Сегодня исследованиями в этой сфере занимаются многие исследователи по всему миру. Сентимент анализ это процесс извлечения эмоции, мнений, настроений или отношения людей к продуктам, сервису, событиям и т.д. через анализ текстов(Bing Liu, 2015).

Решения задач сентимент анализа применяются во многих сферах жизни (Giannakopoulos, Theodoros et al., 2015; A. Makazhanov , D. Rafiei, 2013; M. Choy, M. L. F. Cheong, et al, 2011., www.smartinsights.com), например:

- Коммерческая сфера (оценка продуктов, сервисов, брендов);

Общественные события (прогноз результатов выборов, оценка обсуждаемых тем)];

- Развлечения (рейтинги фильмов, книг, звезд);
- Здравоохранение (рейтинги поликлиник, врачей);
- образование (рейтинги университетов, школ, детских садов);
- туризм (оценка гостиниц, туристических направлений, агентств);
- спорт (рейтинги спортсменов, тренеров, команд) и т.д..

В мире существуют множество решений, приложений для сентимент анализа текстов, но многие решения в основном работают с текстами на английском, итальянском, немецком языках. Например, Google cloud; Microsoft Azure Text Analytics API; Microsoft Azure Emotion API; Social Mention; Sentiment140; «SentiStrength»; Semantria; SentiFinder и другие.

Для задач сентимент анализа используются общеизвестные методы, такие как методы машинного обучения и лингвистические методы. Для определения тональности текстов на казахском языке был выбран лингвистическией метод основанный на словарях и правилах.

Для сентимент анализа текстов на казахском языке были решены следующие задачи:

- Построение словаря лексической тональности казахского языка;
- Построение и формализация правил определения сентимента в предложении;

• Разработки и реализация алгоритма сентимент анализа текстов на казахском языке на основе словарей и правил.

2. Словарь тональных слов казахского языка

Лексические русурсы, такие как, словари, тезаурусы имеют большую ценность для решения разных задач. Существуют разные лексические ресурсы для определения тональностей текстов, в основном для английского, итальянского, русского. Например, есть такие открытые ресурсы, как WordNet-Affect, SentiWordNet, SenticNet, MPQA Opinion Corpus для английского языка, для русского языка РуСентиЛекс объемом 12000 слов и словосочетании.

Для казахского языка отсутствуют открытые лексические ресурсы, в том числе и тональные словари. Для решения задачи автоматического определения тональности текста был создан словарь эмоциональной лексики казахского языка. Словарь был вручную создан и размечен по тональности по 5-бальной шкале (от -2 до 2).

word	meaning	POS	Semantic orientation
Абай болды.	а) Сақ болды. ә) Ес болды, бас-көз болды.		1
АБАЙЛАҒЫШ	Байқағыш, сезгіш.	сын.	1
АБАТ	Бай, берекелі.	ир., сын.	1
АБДЫРАҒЫШ	Асып-сасқыш, абыржығыш, қысылғыш.	сын.	-1
АБЗАЛ	1. Қадірлі, құрметті, ардакты. 2. Жон, дұрыс, орынды, қолайлы.	сын.	1
АБИЫРЛЫ	Абыройлы, беделді.	сын.	2
АБИЫРСЫЗ	1. Абыройсыз, беделсіз. 2. Ұятсыз, коргенсіз.	сын.	-2
АБҮЙЫРЛЫ	Абиырлы.	сын., сойл.	2
АБҮЙЫРСЫЗ	Абиырсыз.	сын., сойл.	-2
АБЫРОЙЛЫ	1. Әбүйірлі, жақсы атакты. 2. Ойдан шыккандай, көңілдегідей. 3. Сыпайы, әдепті.	сын.	2
АБЫРОЙСЫЗ	Қадірі жок, беделсіз.	сын.	-2
АҒАТ	Қате, теріс, жаңылыс.	сын.	-1
Ағат кетті	Қателесті, жаңылысты.	ет.	-1
АҒАТСЫЗ	Қатесіз, мұлтіксіз.	сын.	1
АҒАТТАУ	Әбестеу, терістеу, орынсыздау.	сын.	-1
АҒЫНДЫ	Қаркынды, екпінді.	сын.	1
АДА	Жұрдай, тұк жок.	ар., сын.	-1
Ада болды.	Таусылды, түгесілді, бітті.	ет.	-1
Ада етті, ада қылды	Тауысты, бітірді.	ет.	-1
АДАЛ	1. Ақ ниетті, таза ойлы, әділ. 2. Алдаусыз, арамдыксыз. 3. Нағыз жанашыр, таза сезімді. 4. Тура, ашық, шын. 5. Күнәсіз, жазығы жок.	ар., сын.	2
Адал жол	Дұрыс бағыт, орынды іс-әрекет.		2
Адалынан жолықты	Шын көңілдегідей болды, ойдағыдай болып шықты, таза, ақ ниетті болды.	ет.	2

Рис. 1. Фрагмент словаря эмоциональной лексики казахского языка

При создании словаря для казахского языка были выделены следующие классы:

- Прилагательные слова и словосочетания (оценки: [-2; 2]);
- Существительные (оценки: [-1;1]);
- Глаголы слова и словосочетания (оценки: [-1;1]);
- Наречия (усилительные).

Словарь содержит около 11000 тональных слов и словосочетании из разных частей речи (см. Рис.1).

3. Определение сентимента текста

Текст состоит из последовательностей лексических единиц, которыми являются слово, устойчивое словосочетание или другая единица языка, способная обозначать предметы, явления, их признаки и т.п.

Тональность текста – это эмоциональная оценка автора по отношению к какому- нибудь событию или объекту, представленном в тексте комментария, которая определяется тональностями его лексических единиц.

Тональность лексических единиц измеряется следующими величинами:

-2 – крайне негативный;
-1 – негативный;
0 – нейтральный;
1 – позитивный;
2 – крайне позитивный.

Тональность всего текста определяется как средне-арифметическая величин измерения тональностей лексических единиц (предложений) и правил их сочетания.

$$sent(L) = \frac{\sum_{i=1}^{n} sent_i(\omega)}{n}$$

В казахском языке тональность тексту придают такие части речи, как существительные (жауыздық, соғыс), глагол (тұтқындау, қуанды, ашуланды), прилагательное (әдемі/ көріксіз, жақсы/жаман), наречие (нағыз, ең, өте). Как показывает исследования, существительное является аспектом (объектом) обсуждения, а прилагательные в основном определяют семантическую ориентацию (полярность) текста. Тональность оценочного слова может

зависеть от контекста и предметной области. Также тональность может изменятся или усиливаться в зависимости от наречия, глагола и союзов. Фрагменты правил были описаны в предыдущих работах [Б.Ж. Ергеш, 2016; Yergesh B., Sharipbay A.(2016)]. На основе созданных словаря и правил был разработан и реализован алгоритм автоматического определения сентимента текстов на казахском языке. Алгоритм определения сентимента показан на рисунке 2.

Рис. 2. Алгоритм определения сентимента текстов

4. Заключение

В статье описан созданный словарь эмоциональной лексики казахского языка, размеченный по пятибальной шкале. Словарь содержит около 11000 тональных слов и словосочетании На основе этого словаря и правил определения сентимента построен и реализован алгоритм автоматического определения сентимента.

ЛИТЕРАТУРА

1. Bing Liu. (2015) Sentiment Analysis: mining opinions, sentiments, and emotions. Cambridge University Press.

2. Giannakopoulos, Theodoros et al. (2015) Visual sentiment analysis for brand monitoring enhancement. 9th International Symposium on Image and Signal Processing and Analysis (ISPA): 1-6.

3. A. Makazhanov and D. Rafiei (2013) Predicting political preference of Twitter users, 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), Niagara Falls, ON, , pp. 298–305.

4. M. Choy, M. L. F. Cheong, M. N. Laik, and K. P. Shung (2011) A sentiment analysis of Singapore presidential election 2011 using twitter data with census correction. CoRR, vol. abs/1108.5520.

5. WordNet-Affect http://wndomains.fbk.eu/wnaffect.html

6. SentiWordNet http://sentiwordnet.isti.cnr.it

7. SenticNet http://sentic.net/

8. MPQA Opinion Corpus http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/

9. РуСентиЛекс http://www.labinform.ru/pub/rusentilex/index.htm

10. Б.Ж. Ергеш, А.А. Шарипбай, Г.Т. Бекманова (2016) Роль имен прилагательных в определении тональности текста. Труды международной конференции по компьютерной и когнитивной лингвистике TEL-2016. – Казань: Изд-во Казан. ун-та, 2016. – стр 85–89.

11. Yergesh B., Sharipbay A., Bekmanova G., Lipnitskii S (2016) Sentiment analysis of Kazakh phrases based on morphological rules. Journal Of Kyrgyz State Technical University named after I.Razzakov, Theoretical And Applied Scientific Technical Journal, Bishkek, Kyrgyzstan, № 2 (38), pp. 39–42, ISSN 1694-5557.

УДК 004.42:811

INDEPENDENT COMPUTER PRESENTATION OF SPATIAL NOTIONS IN TURKIC LANGUAGES

S. Karabaeva[1], P. Pankov[2]
[1]Kyrgyz State University of Construction, Transportation and Architecture, Maldybayeva Str. 34, Bishkek 720020, Kyrgyzstan
[2]Institute of Mathematics, Chui prospect 265a, Bishkek 720071, Kyrgyzstan
pps50@rambler.ru

Supra, we proposed a definition of independent computer presentation of an object (the user can master the object without reference to similar objects). Particularly, it means that the user is to be able to master foundations of the language by using corresponding software (with interactive actions with feedback) without any other language as a media. Hence, the only software is needed for presentation of any natural language regardless of native languages of users. To implement such definition, definitions of "almost-closed or affectable objects" (including both humans and computers), of "commands" (low-energetic outer influences on affectable objects causing sufficiently various high-energetic reactions and consequences), of mathematical and computer models of notions of natural languages were introduced. A mathematical model of a notion consists of preliminaries (if the notion is not primary); description of media (objects presented as sets on display); permitted and prohibited relations between objects (overlapping, intersection, inclusion) and the user's actions (moving and transforming objects); command involving the notion with random generation of auxiliary words; temporal sequence of relations between objects to be done by the user in order to guess and fulfill the command.

We distinguished that spatial notions in Turkic languages are not "relations between objects" (as in other languages) but are "domains of space" (upper-space; interior; exterior; left-space; right-space; near-space etc.) defined by position of affectable object (watcher with sensors), directions of motions and gravitation. Experiments with native speakers specified these domains of space. Respondents' marking of points of "near-space" substituted L. Zadeh's conclusions on fuzzy sets but their marking of points of other domains discovered the phenomenon of discretization of opinions of different people for non-parallelepiped-like objects.

We propose the following definition: If some native speakers being said (separately) a *pseudo-word* (combination of phonemes that does not exist in the language but sounds like a word of any part of speech in this language) and corresponding grammatical question give similar responds (transformations of

pseudo-word) then it means existence of the algorithm of word-formation for this grammatical question; if such algorithms of word-formation for different grammatical questions have an essential common part then it is said to be the unified algorithm of word-formation in this language. It is well-known that a person proficient in one of Turkic languages not only can guess meaning of some word stems in another Turkic language but can also guess a grammatical sense of a simple phrase. This proves existence of a unified algorithm of word-formation in Turkic languages.

Basing on these experiments we propose mathematical and computer models of spatial notions in Turkic languages. Existence of unified algorithms of word-formation in Turkic languages facilitates computer implementation of these models as a constituent of their general independent interactive computer presentation and complex examination.

Keywords: independent computer presentation; interactive presentation; spatial notion; Turkic languages.

НЕЗАВИСИМОЕ КОМПЬЮТЕРНОЕ ПРЕДСТАВЛЕНИЕ ПРОСТРАНСТВЕННЫХ ПОНЯТИЙ В ТЮРКСКИХ ЯЗЫКАХ

С.Ж. Карабаева[1], П.С. Панков[2]

[1]Кыргызский государственный университет строительства, транспорта и архитектуры

[2]Институт математики Национальной академии наук КР

pps50@rambler.ru

Ранее, мы предложили определение независимого компьютерного представления объекта (пользователь может осваивать объект без ссылки на подобные объекты). В частности, это означает, что пользователь должен иметь возможность овладеть основами языка, используя соответствующее программное обеспечение (с интерактивными действиями с обратной связью) без какого-либо другого языка в качестве посредника. Следовательно, только одно программное обеспечение будет необходимо для представления любого естественного языка, независимо от родных языков пользователей. Чтобы реализовать такое определение «почти замкнутых или воздействуемых объектов» (включая людей и компьютеры), введено определение «команды» (малые по затратам энергии воздействия на объект вызывают существенно различные изменения во внутреннем состоянии объекта и большие по затратам энергии действия объекта). Были введены математические и компьютерные модели понятий естественных языков.

Ключевые слова: независимая компьютерная презентация; интерактивная презентация; пространственное понятие; Тюркские языки.

1. Introduction

One of main tasks of the up-to-date informatics is developing of interactive computer presentations of all familiar real and virtual objects to offer the user the opportunity to master them safely and effectively before real treating. If such computer presentation does not depend on the user's knowledge and skills on similar objects then we call it independent. By our opinion, independent computer presentations are more effective because the user becomes proficient in the object, immediately, without references to other objects in mind. (Definitions of terms are in Section 2).

Earlier, investigating and learning a living language were implemented with the assistance of persons who had a complete command of it. Invention of recording sounds gave possibility to fix examples of an oral language objectively. Invention of talking pictures fixed examples of phrases with connection to situations and actions. Computer games gave the user the opportunity to choose actions with corresponding phrases. Existing well-known software to learn languages base on languages native to the user, nevertheless some notions are presented independently.

We put the problem of completely independent presentations of natural languages and proposed to involve interactive actions with feedback by the user for this purpose).

For this purpose we introduced energetic definitions of "almost-closed or affectable objects" (including both humans and computers), of "commands" (Панков, Баячорова, Жураев, 2010).

Combining ideas of (Asher, 1965), (Winograd, 1972) and (Zadeh, 1975) we gave suggestions and have developed elements of such presentations (Pankov, Aidaraliyeva, Lopatkin, 1996), (Pankov, Alimbay, 2005), have implemented a primary version of an algorithmic language for such purposes (Pankov, Dolmatova, 2007), proposed presentation of a language as whole (Bayachorova, Pankov, 2009), (Pankov, Bayachorova, Juraev, 2012) including verbs, nouns, adjectives, prepositions and some abstract notions (Bayachorova, Pankov, 2013). Also, expanding computer simulators, we proposed the notion of multimedia complex computer examination and developed such examination on a natural language (Pankov, Dolmatova, 2009).

We distinguished that spatial notions in Turkic languages are not "relations" (as in other languages) but are "domains of space" defined

by position of the affectable object (watcher with sensors), directions of probable or actual motion and gravitation (Karabaeva, Dolmatova, 2014), (Karabaeva, 2016), (Карабаева, 2016).

The purpose of this paper is construction of mathematical and computer models of spatial notions in Turkic languages. Existence of unified algorithms of word-formation in Turkic languages (Панков, Исмаилов, Асилбеков, Карасев, Парахин, 1991), (Панков, 1992) facilitates computer implementation of these models as a constituent of their general independent interactive computer presentation and complex computer examinations.

2. Definitions

Definiton 1. If a computer presentation of an object does not depend on the user's

knowledge and skills on similar objects then we call it independent.

In our opinion, such presentations are more effective because the user can learn inductively – without referencing other objects in mind. In regards with learning a language, the user begins to thinking in it, without translation in mind.

The following notions are used by us in computer testing.

Definiton 2. Generativity means that a complete task must not exist before the testing and must be generated (randomly).

Generativity is used also for presentation of notions with random auxiliary objects.

Definiton 3. Uniqueness means that all examinees must obtain different versions of tasks (of the same level of difficulty).

Definiton 4. Concreteness means that the user's respond may be: number (exact or approximate), word, short phrase, action ("Drag-and-Drop" by a computer mouse).

Definition 5. If a computer examination meets Definitions 2, 3, 4, tests both knowledge and skills on a subject, involves multimedia and contains various tasks including ones with sound and graphical conditions then it is said to be complex.

Particularly, complex a computer examination on a language is to contain tasks on short dictation, text conclusions from short sound commands, graphical images and short video clips.

The following definitions describe a language (both of human and computer) from our standpoint.

Definition 6. If low energetic outer influences can cause sufficiently various reactions and changing of the inner state of the object (by means of inner energy of the object or of outer energy entering into object besides of commands) at any time then such (permanently unstable) object is an almost-closed or affectable object, or a subject, and such outer influences are commands.

Definition 7. A system of commands such that any subject can achieve desired efficiently various consequences from other one is a language.

Hypothesis 1. A human's genuine understanding of a text in a natural language can be clarified by means of observing the human's actions in real life situations corresponding to the text.

Definition 8. Let any notion (word of a language) be given. If an algorithm acting at a computer: generates (randomly) a sufficiently large amount of instances covering all essential aspects of the notion to the user, gives a command involving this notion in each situation, perceives the user's actions and performs their results clearly on a display, detects whether a result fits the command, then such algorithm is said to be a computer interactive presentation of the notion.

Certainly, commands are to contain other words too. But these words must not give any definitions or explanations of the notion.

Definition 9. If all words being used in Definition 8 are unknown to the user nevertheless s/he is be able to fulfil the meant action (because it is the only natural one in this situation) then the notion (word of a language) is said to be primary. If the user has to know supplementary words to complete the action then the notion is said to be secondary. Thus, there is a natural hierarchy of notions.

Using this method we can present not only real notions (objects and actions) but also notions which have imaginary concepts. Also, some computer games can be adapted to develop an "adequate" representations of fictional and abstract concepts or objects.

Hypothesis 2. A person learning a natural language without references to any other ones, hearing a notion in various situations begins to form a kind of mathematical_model in mind corresponding to this notion by means of trial and error method and attempts to fulfil operations similar to mathematical ones: closing and compactification. After successful completing such operations, the human feels "mastering" this notion.

Hypothesis 3. Any notion has a minimalistic model (involving minimal number of entities in Occam's sense).

3. On unified algorithms of word-formation in Turkic languages

Definition 10. Choose any "*pseudo-word*" (combination of phonemes) that does not exist in the language but sounds like a word of any part of speech in this language. If some native speakers being said (separately) this "*pseudo-word*" and corresponding grammatical question give similar responds (transformations of "*pseudo-word*") then it means existence of the algorithm of word-formation for this grammatical question.

If such algorithms of word-formation for different grammatical questions have an essential common part then it is said the unified algorithm of word-formation in this language.

For instance, it is well-known that such unified algorithm does not exist in Russian language. By this reason, many foreign words are not declined in Russian writing. In oral speech, Russian native speakers sometimes adapt such words to Russian usage and decline such "adapted word". For example, widespread female name *Çolpon* (*Şolpan*, *Çulpan*, *Sulpan*) is often said as *Çolpona* (nominative case) and is declined by rules for feminine nouns in Russian.

We implemented essential parts of unified algorithms of word-formation in Kyrgyz and Kazakh languages as follows: given a word stem and a primary form of affix. The algorithm elaborates (sometimes transformed) stem together with transformed affix.

A typical example: Kyrgyz native speakers were said: *Emne? (What-singular?) – Lömük* [such word does not exist]; *Emneler? (What-plural?).* All they responded: *Lömüktör.* It coincided with the result given by the computer program: *Lömük+lar = Lömüktör.*

It is well-known that a person proficient in one of Turkic languages not only can guess meaning of some word stems in another Turkic language but can also guess a grammatical sense of a simple phrase. By this reason, a task to construct a unified algorithm of word-formation in Turkic languages was put (Мусаев, Карабаева, Иманалиева, 2013).

Certainly, such algorithm (evidently, based on synharmonism) would involve variations.

Remark. Such variations arise in certain languages too. For example, in rare combinations Kyrgyz native speakers have difficulty in the following:

$$...r+l...=...rl...?\ =...rd...?$$

4. On denotations of domains of space in Turkic languages

The denotations of domains of space in the Kyrgyz language are follows:

üstü – upper-space; *astī* – before-and-lower-(observed)-space; *iç* – interior;

sīrt; tīş – exterior; *çek* – boundary-strip; *sol* – left-space; *oň* – right-space;

orto – middle-spot; *can* – near-space; *ara* – between-space; *ald (aldī)* – before-forward-space; *art (artī)* – behind-space; *karşī* – opposite-space.

These denotations in other Turkic languages are similar, sometimes due to rules of transition between languages. For instance, *(sol; oň)* (Kyrgyz, Kazakh) ~ *(sul; uň)* (Tatar) ~ *(sol; sağ)* (Turkish).

These terms can be used being inclined in various cases, for example

Üköktün solunda (locative case) – within left-space of the box;

Üköktün soluna (dative case) – to left-space of the box;

Üköktün solunan (ablative case) – from left-space of the box;

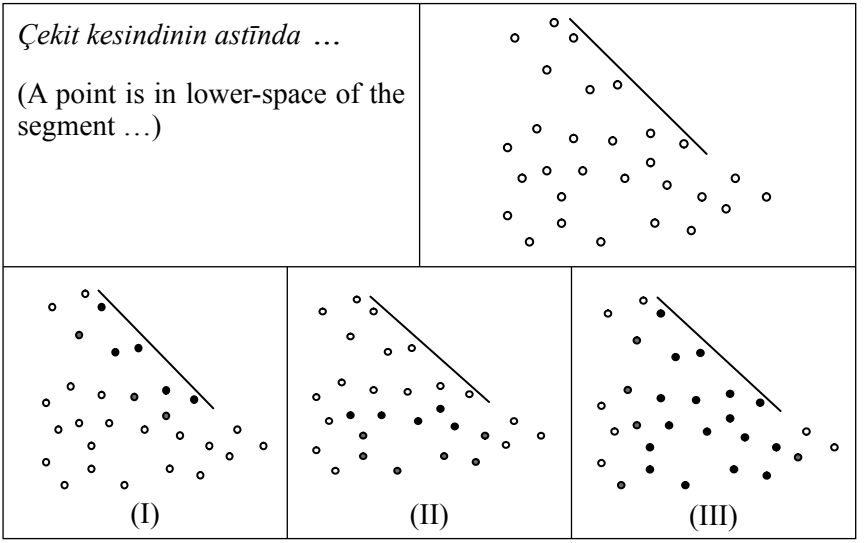
Üköktün solun kördüm (accusative case) – I saw left-space of the box.

Example of subdomain of domain:

Üköktün içinin solunda (possessive case, locative case) – within left-space of interior of the box.

We conducted the following experiments with Kyrgyz native speakers to specify these domains of space.

Native speakers were given sheets with an object and some dozens white points and texts of type “mark points meeting the statement “a point is in domain-space of the object” “. Results on the term “*canında”* (in the near space of) substituted L. Zadeh’s conclusions on fuzzy sets but results on other terms with respect to non-parallelepiped-like objects discovered the phenomenon of discretization of opinions of different people.



There arose three sufficiently different types of responds. Points marked by most of respondents are black; ones done by some respondents are grey.

These results can be explained as follows. Denote the segment as *S*. Respondents (I) and (III) marked points (x_1,y_1) meeting the condition $(\exists(x,y)\in S)(y_1<y)$ meanwhile respondents (II) marked points (x_1,y_1) meeting the condition $(\forall(x,y)\in S)\ (y_1<y)$. Also, respondents (I) deemed that "lower-space" meant "near-space" too.

5. Mathematical models of spatial notions in Turkic languages

A mathematical model of a notion consists of preliminaries (if the notion is not primary); description of media (objects presented as sets on display); permitted and prohibited relations between objects (overlapping, intersection, inclusion) and the user's actions (moving and transforming objects); command involving the notion with random generation of auxiliary words; temporal sequence of relations between objects to be done by the user in order to guess and fulfill the command.

For example, describe a model for *sol* and *oň*. It consists of three sequential tasks.

Because of the above-mentioned phenomenon, we take a square only for “object”.

Preliminary knowledge: *çarçī* (square), *caşīl* (green), *kök* (blue), *names* of five little Things T_1, T_2, T_3, T_4, T_5.

Denote domains on display:

$G:=\{2\le x\le 3;\ 1\le y\le 2\}$ *(painted with green);*

$B:=\{6\le x\le 7;\ 1\le y\le 2\}$ *(painted with blue);*

$G_L:=\{1\le x\le 2;\ 1\le y\le 2\}$; $G_R:=\{3\le x\le 4;\ 1\le y\le 2\}$;

$G_D:=\{2\le x\le 3;\ 0\le y\le 1\}$; $G_U:=\{2\le x\le 3;\ 2\le y\le 3\}$;

$B_L:=\{5\le x\le 6;\ 1\le y\le 2\}$, $B_R:=\{7\le x\le 8;\ 1\le y\le 2\}$,

$B_D:=\{6\le x\le 7;\ 0\le y\le 1\}$, $B_U:=\{6\le x\le 7;\ 2\le y\le 3\}$.

First task.

Initial situation: $G, B, T_1, T_2, T_3, T_4, T_5 \subset G_L$.

$j:=random(1,2,3,4,5)$. T_j is declared as movable.

Conditions: $T_j \cap G = \varnothing$; $T_j \cap B = \varnothing$.

Command: *(name of* T_j*)+Accusative_case; caşīl çarçīnīn solunan kök çarçīnīn soluna cīldīr!* (Move (name of T_j) from left-space of the green square to left-space of the blue square!)

Temporal sequence to check understanding of Command: $T_j \subset B_L$ ungrasp.

Second task. (Similar to First task for *oň*).

Third task (consolidation of knowledge).

Command: *(name of* T_j*)+Accusative_case; caşīl çarçīnīn solunan kök çarçīnīn oňuna cīldīr!* (Move (name of T_j) from left-space of the green square to right-space of the blue square!).

In such a way, other spatial notions are introduced consequently.

Acknowledgements

The authors are grateful to the National Commission on the State Language under the President of Kyrgyz Republic for encouragement of these investigations, to the Presidium of National Academy of Sciences of Kyrgyz Republic for sponsorship, to our coauthors for mutual work, to all students who participated in experiments.

REFERENCES

1. Asher, J. (1965) The strategy of total physical response: An application to learning Russian. International Review of Applied Linguistics, no. 3.

2. Winograd, T. (1972) Understanding Natural Language. Massachusetts Institute of Technology, New York.

3. 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.

4. Панков, П.С., Исмаилов, Б.Д., Асилбеков, А.Б., Карасев, В.Б., Парахин, В.А. (1991) Контрольно-обучающая программа по словоизменению в казахском языке на ПЭВМ. Научно-практические аспекты повышения качества подготовки учителей математики и информатики: тезисы докладов республиканской конференции. Алма-Ата. – Часть 2, с. 64–65.

5. Панков, П.С. (1992) Обучающая и контролирующая программа по словоизменению в кыргызском языке на ПЭВМ. Бишкек.

6. Pankov, P.S., Aidaraliyeva, J.Sh., Lopatkin, V.S. (1996) Active English on computer. Conference “Improving Content and Approach in the Teaching of English Language in the Context of Educational Reform”, Bishkek, pp. 25–27.

7. Pankov, P.S., Alimbay, E. (2005) Virtual Environment for Interactive Learning Languages. Human Language Technologies as a Challenge for Computer Science and Linguistics: Proceedings of 2nd Language and Technology Conference, Poznan, pp. 357–360.

8. Pankov, P.S., Dolmatova, P.S. (2007) Algorithmical Language for Computer-Based Presentation of Notions. 4th International Conference on Electronics and Computer. Suleyman Demirel University, Almaty, pp. 274–279.

9. Bayachorova, B.J., Pankov, P.S. (2009) Independent Computer Presentation of a Natural Language. Varia Informatica. Polish Information Processing Society, Lublin, pp. 73–84.

10. Pankov, P., Dolmatova, P. (2009) Software for Complex Examination on Natural Languages. Human Language Technologies as a Challenge for Computer Science and Linguistics: Proceedings of 4th Language and Technology Conference, Poznan, pp. 502–506.

11. Панков, П., Баячорова, Б., Жураев, М. Кыргыз тилин компьютерде чагылдыруу. Бишкек, 2010.

12. Pankov, P.S., Bayachorova, B.J., Juraev, M. (2012) Mathematical Models for Independent Computer Presentation of Turkic Languages. TWMS Journal of Pure and Applied Mathematics, Volume 3, No. 1, pp. 92–102.

13. Мусаев, С.Ж., Карабаева, С. Ж., Иманалиева, А.И. (2013) Проблемы и перспективы развития компьютерной лингвистики в Кыргыз-

стане, in: Proceedings of the I International Conference on Computer processing of Turkic Languages (TurkLang-2013), Astana, pp. 35–38.

14. Bayachorova, B., Pankov, P. (2013) Independent Computer Presentation of Abstract Notions of Natural Languages. Actual Problems of Computer Science, ECCC Foundation, Lublin, No. 1(3), pp. 64–69.

15. Karabaeva, S., Dolmatova, P. (2014) Mathematical and computer models of spatial relations in Kyrgyz language. Proceedings of V Congress of the Turkic World Mathematicians. Bishkek, pp. 175–178.

16. Karabaeva, S. (2016) Presentation of spatial-temporal relations in Kyrgyz language. Труды Международной конференции по компьютерной и когнитивной лингвистике TEL-2016. – Казань, с. 274–277.

17. Карабаева, С. (2016) Единый алгоритм словоизменения и представление пространства в кыргызском языке. – Saarbrűcken, Lap Lambert Academic Publishing.

УДК 004.91

THESAURUS ON ISLAM: DEVELOPMENT AND CURRENT STATE

N. Loukachevitch, B. Dobrov
Lomonosov Moscow State University, Moscow, Russia
louk_nat@mail.ru

The paper describes the principles and the current state of the created Islam Thesaurus.

The Russian Islam Thesaurus has been developed as a resource for automatic document processing of Internet-pages, news articles, or social network messages. It was prepared in the format of the RuThes thesaurus to be used in joint document processing.

The main stages of the Islam Thesaurus development include:

– to find an important Islam or Muslim-related concept analyzing the results of automatic term extraction, Islam dictionaries or Islam-related texts;

– to introduce the concept into the thesaurus providing it with a unique understandable name,

– to provide the concept with text entries that can express it in texts (synonyms). The synonyms should be equivalent relative to the concept relations. It is supposed that in a thesaurus for automatic document processing, rich synonymic rows should created, including different parts of speech and term variants, if they exist;

– to describe relations of the introduced concept to existing thesaurus concepts.

If a concept is usually expressed with an ambiguous word then one of the following forms could be chosen for its name: the ambiguous word with an additional label in parentheses or unambiguous multiword expression having the same sense. The procedure is similar to choosing descriptor names in traditional information-retrieval thesauri.

For example, word Sunnah mainly denotes the verbally transmitted record of the teachings, deeds and sayings, silent permissions (or disapprovals) of the Islamic prophet Muhammad. But in Russian and in English the same word can also mean an encouraged, commendable deed of a Muslim. Therefore, two concepts were introduced into the Islam Thesaurus, each with own set of text entries and relations.,

The Islam terminology contains many terms that are not understandable to other people, therefore many concepts contain some explanation in their names in form of additional labesl, for example, *JANAZAH (FUNERAL)*, *KHUTBAH (PREACHING), RIBA (USURY)*. These labels usually represent more general non-islamic concepts.

Currently, the Islam Thesaurus includes more than 5 thousand terms. It is assumed that the Islam Thesaurus will be compatible with the published version of the RuThes thesaurus, which will allow using this complex for processing a wide range of texts, including news reports, specialized websites, posts in social networks.

Keywords: thesaurus; islam; natural language processing.

ТЕЗАУРУС ПО ИСЛАМУ: РАЗРАБОТКА И ТЕКУЩЕЕ СОСТОЯНИЕ

Н.В. Лукашевич, Б.В. Добров
НИВЦ МГУ имени М.В. Ломоносова, Москва, Россия
louk_nat@mail.ru

В работе представлен Тезаурус по исламу, который содержит понятия ислама и связанные с ним понятия общественной жизни. Тезаурус содержит более пяти тысяч терминов. Тезаурус создан по модели тезауруса РуТез как ресурс для автоматической обработки текстов и поэтому может применяться для контент-анализа текстов, автоматической классификации текстов и др. Предполагается, что Тезаурус по исламу будет совместим с опубликованной версией тезауруса РуТез, что позволит применять этот комплекс для обработки широкого круга текстов, включая новостные сообщения, специализированные сайты, посты в социальных сетях.

Ключевые слова: тезаурус, ислам, автоматическая обработка текстов.

1. Введение

Для разнообразных приложений автоматической обработки текстов могут быть оказаться полезными так называемые тезаурусы, формализованные ресурсы для описания отношений между словами языка и/или терминами предметной области. Одним из известных тезаурусов является тезаурус WordNet (Fellbaum, 1998). Тезаурусы типа WordNet создаются и для многих других языков. Для русского языка имеется тезаурус РуТез (Лукашевич 2011, Loukachevitch et al., 2014), который также применялся в различных приложениях автоматической обработки текстов и информационного поиска.

Модели представления информации в тезаурусах WordNet и РуТез похожи. В настоящее время на основе тезауруса РуТез создан тезаурус типа WordNet для русского языка (RuWordNet)

(Loukachevitch et al, 2016). Но вместе с тем имеются и существенные отличия: РуТез больше приспособлен для представления различных словосочетаний, терминов предметной области. Он исходно содержит в себе не только лексику литературного русского языка, но и термины так называемой общественно-политической области (Loukachevitch, Dobrov, 2015).

В своей модели представления знаний РуТез учитывает традицию информационно-поисковых тезаурусов (ISO 2788-1986, Z39.19), описывающих терминологию предметных областей для нужд индексирования документов в информационно-поисковых системах. На основе модели представления знаний РуТез созданы ряд терминологических ресурсов в нескольких разных предметных областях, которые используются в задачах информационного поиска и автоматической обработки текстов, включая онтологию по естественным наукам и технологиям ОЕНТ (Добров и др., 2008), онтологию в области авиации Авиаонтология (Добров и др., 2004), тезаурус в банковской области (Nokel, Loukachevitch, 2016).

РуТез в рамках своего общественно-политического блока содержит значимый набор терминов, связанных с различными мировыми религиями, поскольку они представляют неотъемлемую часть жизни современного общества. В данной работе мы рассматриваем процедуру создания Тезауруса по исламу на основе модели представления знаний тезауруса РуТез. Такой тезаурус можно будет использовать в автоматической обработке текстов, совместно с тезаурусом РуТез. Создаваемый тезаурус должен охватывать широкий круг терминологии описывающий основные положения ислама, исламское право и обычаи, и др.

Что касается ресурсов, связанных с исламом, то можно найти серию работ, связанных с созданием онтологии знаний в исламе (Islamic knowledge ontology), создаваемой в Университете Малайзии (Saad et al. 2010–2016). В последней работе (Weaam and Saad, 2016) исследуются методы извлечения онтологии из текста Корана. В частности, извлекаются группы существительных, фильтруются с учетом их частотности в тексте, далее применяются шаблоны для извлечения отношений. В нашей работе Исламский тезаурус создается из большого количества разных ресурсов, и цель его не только включить важные понятия, упоминаемые в Коране и других религиозных исламских текстах, но и описать

взаимосвязи с другими религиями, конфликтные вопросы, обычаи мусульман (например, одежда), т.е. мы работаем со значительно большим социо-культурным контекстом и с разнообразными источниками.

2. Тезаурус РуТез

Тезаурус РуТез представляет собой лингвистическую онтологию, т.е. онтологию, понятия которой опираются на значения существующих в языке слов и выражений. Каждое понятие онтологии имеет уникальное и однозначное имя. Каждое понятие онтологии связано с набором слов и выражений, посредством которых данное понятие может выражаться в тексте – текстовые входы понятия.

Набор текстовых входов понятия может включать собственно синонимы, слова разных частей речи (так называемые дериваты), устойчивые словосочетания и другие типы выражений.

В тезаурусе РуТез имеется четыре основных типа отношений.

Первый тип отношений – **родовидовое отношение *ниже-выше***, представляет собой отношение класс-подкласс, обладает свойствами транзитивности и наследования.

Второй тип отношений – **отношение *часть-целое***. Используется не только для описания физических частей, но и для других внутренних сущностей понятия, таких как свойства или роли для ситуаций. Важным условием при установлении этого отношения является то, что понятия-части должны быть жестко связаны со своим целым, то есть каждый пример понятия-части должен в течение всего времени своего существования являться частью для понятия-целого, и не относиться к чему-либо другому. В этих условиях удается выполнить свойство транзитивности введенного таким образом отношения *часть-целое*, что очень важно для автоматического вывода в процессе автоматической обработки текстов.

Еще один тип отношения, называемого **несимметричной ассоциацией *асц*$_2$ – *асц*$_1$**, связывает два понятия, которые не могут быть связаны выше рассмотренными отношениями, но когда одно из которых не существовало бы без существования другого.

Последний тип отношений – **симметричная ассоциация *асц*** связывает, например, понятия, очень близкие по смыслу, но кото-

рые разработчики не решились соединить в одно понятие (предсинонимия).

Таким образом, система отношений тезауруса РуТез описывает наиболее существенные отношения понятий.

3. Создание Исламского тезауруса

Основными этапами создания Тезауруса в области ислама (далее Тезаурус) являются следующие:

- выявить важное понятие, имеющее отношение к исламу или мусульманам, включая понятия, соответствующие обычаям мусульман, конкретных святынь ислама, различные исламские организации, а также общие понятия религиозной жизни, находящие свою конкретизацию в исламе, близкие понятия других религий,
- ввести его в тезаурус, обозначив однозначным именем,
- снабдить понятие, разнообразными синонимичными текстовыми входами, которыми данное понятие может выражаться в тексте,
- описать отношения введенного понятия с уже существующими в тезаурус понятиями.

Рис. 1 показывает экранную форму, в которой введено понятие *Пророк Мухаммед* (слева вверху), форма снизу справа показывает

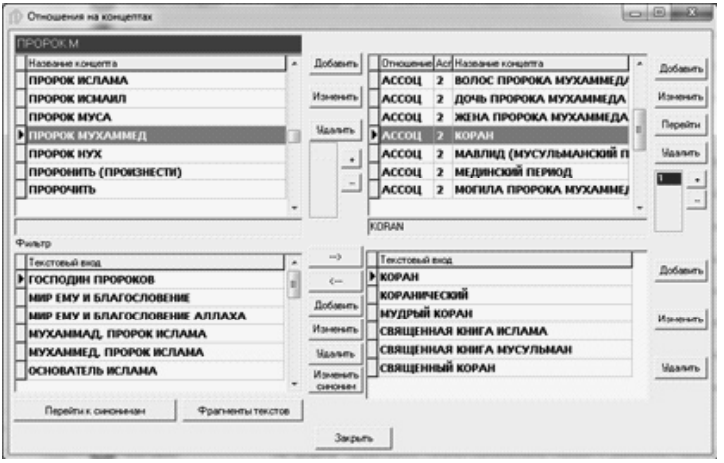


Рис 1. Экранная форма представления понятия Пророк Мухаммед в Тезаурусе ислама

текстовые выражения, которыми обычно упоминают *Пророка* в тексте. Сверху влево указаны отношение пророка *Мухаммеда* с другими понятиями. В частности, курсор стоит на понятии *Коран*. Отношение *ассоц*$_2$ означает, что существование *Корана* связано с существованием *Пророка Мухаммеда*. Форма снизу справа показывает текстовые выражения, которыми называется Коран в текстах.

3.1. Источники для создания Тезауруса

Источником новых понятий и текстовых выражений для ввода в Тезаурус служили:

- материалы исламских сайтов (http://rosmuslim.ru/?lang=ru, http://www.islamnews.ru/, http://golosislama.com/?all=1, http://islam-civil.ru/, http://islamreview.ru/, http://islam-news.ru/ и др.)
- Тезаурус религиоведения ИНИОН РАН
- Али-заде А. Исламский энциклопедический словарь. – М. : Ансар, 2007.
- анализ различных страниц интернета, социальных сетей, обсуждающих ислам.

На материале данных источников была создан текстовый корпус, из которого были извлечены отдельные слова и словосочетания совместно с частотностями их употребления в корпусе.

3.2. Принципы ввода понятий

Материалы этих сайтов анализировались как экспертным путем, так и автоматически: из них были извлечены слова и словосочетания, упорядочены по мере снижения частотности, и эти списки анализировались экспертами для ввода в тезаурус.

Если понятие было связано с многозначным словом, то могла быть выбрана одна из следующих форм выбора однозначного имени понятия:

- пояснение в скобках или
- многословный однозначный синоним.

Например, слово *сунна* имеет в исламе два значения:

Сунна 1. – мусульманское священное предание, излагающее примеры жизни исламского Пророка Мухаммада как образец и руководство для всей мусульманской общины (уммы) и каждого мусульманина

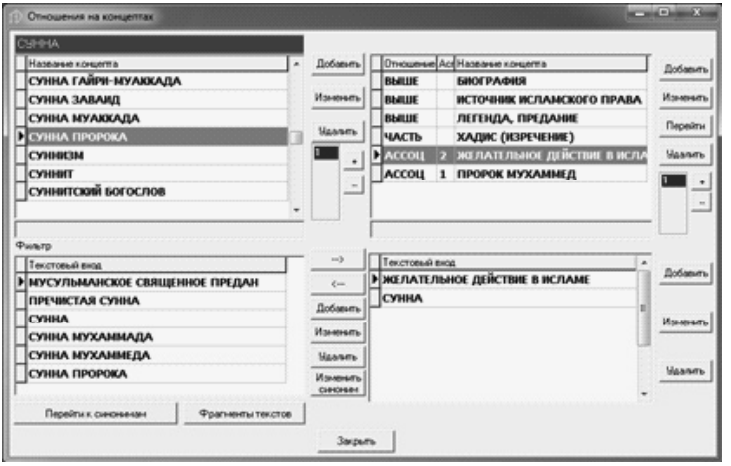


Рис 2. Представление значений слова *сунна*

Сунна 2 означает просто желательные, соответствующие **Сунне 1** поступки.

В результате были введены два понятия:

• - СУННА ПРОРОКА с текстовыми входами: *сунна, суннический, мусульманское священное предание, сунна пророка, сунна мухаммада, сунна мухаммеда, пречистая сунна* и

• - ЖЕЛАТЕЛЬНОЕ ДЕЙСТВИЕ В ИСЛАМЕ с текстовыми входами: *сунна, желательное действие в исламе*.

Таким образом, в данном случае были созданы имена понятий в форме однозначных словосочетаний.

Между собой эти два понятия связаны отношением зависимости. Понятие ЖЕЛАТЕЛЬНОЕ ДЕЙСТВИЕ В ИСЛАМЕ (по определению) зависит от понятия СУННА ПРОРОКА (рис. 2).

В качестве другого примера многозначного слова в исламе можно привести слово *хиджра*, которое имеет по крайней мере три значения. Для их обозначения были введены три понятия:

• ХИДЖРА ПРОРОКА МУХАММЕДА с текстовыми входами: *хиджра пророка Мухаммеда, хиджра, благословенная хиджра* (рис. 3),

• МУСУЛЬМАНСКИЙ КАЛЕНДАРЬ с текстовыми входами: *мусульманский календарь, исламский календарь, мусульманский лунный календарь, исламский лунный календарь, лунная хиджра,*

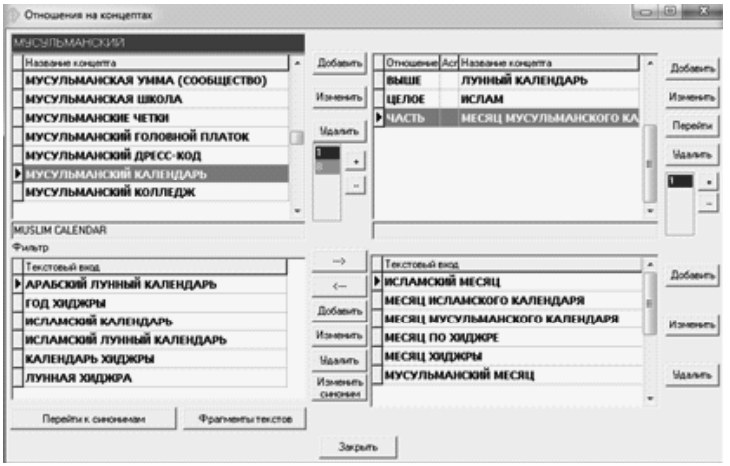


Рис. 3. Представление одного из значений слова *хиджра* как мусульманский календарь

арабский лунный календарь, хиджра, календарь хиджры, по хиджре, год хиджры,

• МУХАДЖИРСТВО (т.е. переселение мусульман из неисламской страны в исламскую страну) с текстовыми входами: *мухаджирство, хиджра, массовое переселение мусульман, совершить хиджру, совершать хиджру*.

Поскольку ислам содержит много терминов, которые не понятны другим людям, часть имен соответствующих понятий, формировалось с кратким пояснением непосредственно в имени понятия, например, ДЖАНАЗА (ПОХОРОННЫЙ ОБРЯД), ХУТБА (ПРОПОВЕДЬ), ДУА (МОЛИТВА), МАХР (ПОДАРОК), РИБА (РОСТОВЩИЧЕСТВО) и т.п. Видно, что краткое пояснение представляет собой просто родовое понятие, которое уже понятно не только мусульманам.

Для каждого понятия набираются разнообразные варианты его выражения в тексте, включая отдельные слова разных частей речи, а также словосочетания.

Например, понятию КОРАН сопоставлены следующие текстовые входы: *коран, коранический, священная книга ислама, священная книга мусульман, мудрый коран.*

Понятие ШИИТ имеет такие текстовые входы: последователь шиизма, приверженец шиизма, рафидит, рафидитка, шиит, шиитка.

Многие понятия могут включать текстовые входы как выраженные на общелитературном русском языке, так и специальный исламские термины. Например, текстовые входы для понятия МУСУЛЬМАНСКИЙ ОБРЯД ОБРЕЗАНИЯ: мусульманский обряд обрезания, обряд обрезания в исламе, обряд обрезания у мусульман, суннат, хатна, хитан.

3.3. Охват понятий Тезауруса

В результате работы в тезаурус были введены следующие типы понятий и конкретные сущности:

– основные понятия, догматы ислама (ХАДИС, ПОКЛОНЕНИЕ АЛЛАХУ, ИДЖТИХАД, ХИДЖРА, КИЯМАТ (СУДНЫЙ ДЕНЬ),НАФС (ЭГО);

– культовые здания в исламе (МЕЧЕТЬ АЛЬ-АКСА, МЕЧЕТЬ ПРОРОКА..);

– культовые предметы (ЧЕРНЫЙ КАМЕНЬ КААБЫ, ХРАМОВАЯ ГОРА, СТЕНА ПЛАЧА);

– действия мусульман в рамках поклонения (НАМАЗ, РАКАТ, ТАХАРАТ (ОМОВЕНИЕ), НАФИЛЬ (НАМАЗ), ДЖАНАЗА-НАМАЗ)

– различные ветви ислама (ХАНАФИТСКИЙ МАЗХАБ, МАЛИКИТСКИЙ МАЗХАБ) (рис. 4);

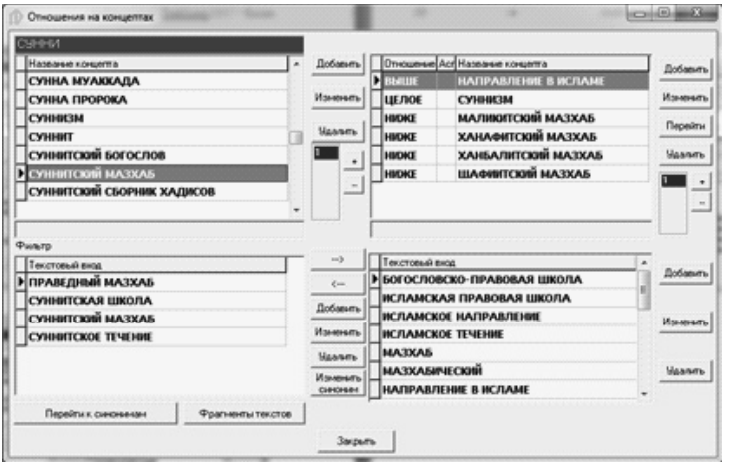


Рис 4. Представление течений в суннизме

– мусульманское право (ШАРИАТ, ШАРИАТСКИЙ СУД, ВАКУФНАЯ СОБСТВЕННОСТЬ);

– суры КОРАНА (СУРА АЛЬ-ФАТИХА, СУРА АЛЬ-БАКАРА);

– религиозная литература (СБОРНИК ХАДИСОВ, ...);

– известные духовные и военные лидеры ислама (АБУ БАКР АЛЬ-БАГДАДИ, МУЛЛА ОМАР, АБУ БАКР (ПРАВЕДНЫЙ ХАЛИФ), ГЮЛЕН, ФЕТХУЛЛАХ);

– мусульманский календарь и праздники (САФАР (МЕСЯЦ), ЗУЛЬ-ХИДЖА, РАБИ АС-САНИ, НОЧЬ РАГАИБ, НОЧЬ МИРАДЖ, ЛУННЫЙ КАЛЕНДАРЬ);

– запретные действия в исламе (ЗИНА (ПРЕЛЮБОДЕЯНИЕ), РИБА (РОСТОВЩИЧЕСТВО), ЗАПРЕТ НА УПОТРЕБЛЕНИЕ СВИНИНЫ..) (Рис. 5);

– исламские финансы (ИСЛАМСКИЙ БАНК, ИСЛАМСКОЕ ОКНО, МУШАРАКА (ПРОЕКТНОЕ ФИНАНСИРОВАНИЕ);

– известные духовные и военные лидеры ислама (АБУ БАКР АЛЬ-БАГДАДИ, МУЛЛА ОМАР, АБУ БАКР (ПРАВЕДНЫЙ ХАЛИФ), ГЮЛЕН, ФЕТХУЛЛАХ);

– мусульманский календарь и праздники (САФАР (МЕСЯЦ), ЗУЛЬ-ХИДЖА, РАБИ АС-САНИ, НОЧЬ РАГАИБ, НОЧЬ МИРАДЖ, ЛУННЫЙ КАЛЕНДАРЬ);

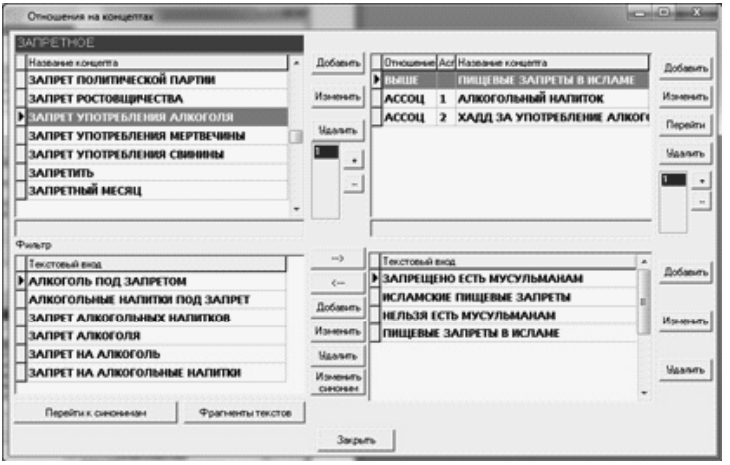


Рис. 5. Представление пищевых запретов в исламе

– религиозная атрибутика (ПОЛУМЕСЯЦ СО ЗВЕЗДОЙ);

– виды мусульманской одежды (НИКАБ, АБАЙЯ, БУРКИНИ);

– мусульманское образование (ИСЛАМСКИЙ УНИВЕРСИТЕТ);

– исламские организации и общины (СОВЕТ МУФТИЕВ РОССИИ, ОРГАНИЗАЦИЯ ИСЛАМСКОГО СОТРУДНИЧЕСТВА, ТАРИКАТ);

– обобщающие понятия, которые существуют в разных религиях (РЕЛИГИОЗНАЯ ОБЩИНА, РЕЛИГИОЗНЫЙ СУД, МОЛИТВЕННОЕ ПОМЕЩЕНИЕ); конфликт с другими религиями (КИБЕРДЖИХАД, КИБЕРХАЛИФАТ, РЕЛИГИОЗНАЯ ДИСКРИМИНАЦИЯ) (рис. 6);

– понятия и конкретные сущности, вовлеченные в текущие конфликты (БАШАР АСАД, СИРИЙСКИЙ КУРДИСТАН, СВОБОДНАЯ АРМИЯ СИРИИ, ПЕШМЕРГА, ШАБИХА (ВОЕНИЗИРОВАННОЕ ФОРМИРОВАНИЕ));

– различные нападения (ГРОМИТЬ КЛАДБИЩЕ, ОСКВЕРНЕНИЕ ХРАМА, ОСКВЕРНЕНИЕ РЕЛИГИОЗНЫХ ЧУВСТВ, ОСКВЕРНЕНИЕ МЕЧЕТИ);

– обряды (ОБРЯД ОБРЕЗАНИЯ, НАЗР КУРБАН (ЖЕРТВОПРИНОШЕНИЕ), ЖЕРТВЕННОЕ ЖИВОТНОЕ, ПАЛОМНИЧЕСТВО К МОГИЛАМ,ТАНЕЦ ДЕРВИША);

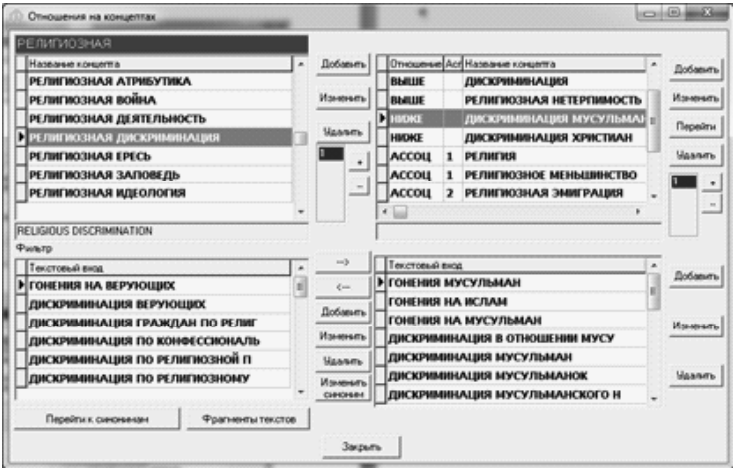


Рис. 6. Представление видов религиозной дискриминации

– политическое устройство стран с преобладанием ислама

– отношение к исламу (ИСЛАМОФОБИЯ, АНТИИСЛАМИЗМ);

– география регионов с преимущественным исламским населением (ЗЕНДЖАН (ПРОВИНЦИЯ), КАЗВИН (ПРОВИНЦИЯ)

4. Заключение

В работе представлен Тезаурус по исламу, который содержит понятия ислама и связанные с ним понятия общественной жизни. Тезаурус содержит более 5 тысяч терминов. Тезаурус создан по модели тезауруса РуТез как ресурс для автоматической обработки текстов и поэтому может применяться для контент-анализа текстов, автоматической классификации текстов и др.

Предполагается, что Тезаурус по исламу будет совместим с опубликованной версией тезауруса РуТез, что позволит применять этот комплекс для обработки широкого круга текстов, включая новостные сообщения, специализированные сайты, посты в социальных сетях. Эти тезаурусы уже применялись совместно в комбинации со статистическими тематическими моделями для контент-анализа новостного сайта (Loukachevitch et al., 2017).

Благодарность

Работа поддержана грантом РФФИ 16-29-09606.

ЛИТЕРАТУРА

1. Информационно-поисковый тезаурус ИНИОН. Религиоведение. Российская академия наука, Институт научной информации по общественным наукам, Москва, ИНИОН РАН, 2008.

2. Добров, Б. В., Лукашевич, Н. В., Невзорова, О. А., & Федунов, Б. Е. (2004). Методы и средства автоматизированного проектирования прикладной онтологии. Известия Российской академии наук. Теория и системы управления, (2), 58–68.

3. Добров Б.В., Лукашевич Н.В. (2008). Онтология по естественным наукам и технологиям ОЕНТ: структура, состав и современное состояние. Электронные библиотеки, Т. 11, №. 1.

4. Лукашевич, Н.В. (2011). Тезаурусы в задачах информационного поиска. Изд-во Московского университета.

5. Fellbaum, Ch.(1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

6. ISO 2788-1986. 1986. Guidelines for the establishment and development of monolingual thesauri.

7. Loukachevitch, N.V., Dobrov. B.V., Chetviorkin, I.I. (2014). RuThes-Lite, a publicly available version of thesaurus of Russian language RuThes. Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2014,. 340–350.

8. Loukachevitch, N., Dobrov, B. (2015) The Sociopolitical Thesaurus as a resource for automatic document processing in Russian. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 21.2, 237–262.

9. Loukachevitch, N.V., Lashevich, G., Gerasimova, A.A., Ivanov, V.V., Dobrov B.V. (2016). Creating Russian WordNet by Conversion, Proceedings of Conference on Computatilnal linguistics and Intellectual technologies Dialog-2016.

10. Loukachevitch N., Nokel M., Ivanov K. (2017). Combining Thesaurus Knowledge and Probabilistic Topic Models //arXiv preprint arXiv:1707.09816.

11. Nokel M., Loukachevitch N. (2016). Accounting ngrams and multiword terms can improve topic models. Proceedings of 12th Workshop on Multiword Expressions, ACL 2016, 44–49.

12. Saad, S., Salim, N., Zainal, H., & Noah, S. A. M. (2010). A framework for Islamic knowledge via ontology representation. In Information Retrieval & Knowledge Management,(CAMP), 2010 International Conference, IEEE, 310–314.

13. Saad S., Salim N., Zainuddin S. (2011). An early stage of knowledge acquisition based on Quranic text, Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on. IEEE, 130–136.

14. Weaam, T., Saad, S. (2016) Ontology population from quranic translation texts based on a combinationof linguistic patterns and association rules, Journal of Theoretical and Applied Information Technology,V. 86,. №. 2, 250.

15. Z39.19. (2005.) Guidelines for the Construction, Format and Management of Monolingual Thesauri. NISO.

УДК 81’32

AUTOMATED ANALYSIS OF NATURAL LANGUAGE QUESTION-ANSWER TEXTS IN THE E-TESTING SYSTEM

N. Prokopiev, D. Suleymanov
Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic Kazan, Russia
dvdt.slt@gmail.com

Automated knowledge control with usage of electronic testing is an important course of action in education in modern digital age. Usually an implementation of this kind of knowledge control involves tests with multiple choice questions, but according to analysis of works (Solovyev, 2011), (Klimov, 2013), (Cherepanova, 2013) it is evident that open-ended question testing with answers in natural language texts still remains weakly developed. This type of electronic testing would be especially useful in application to humanities and language disciplines education, because classic multiple choice tests results often don’t offer a full vision on student’s knowledge level. For the main natural language processing model in answer processor we suggest a model presented in (Suleymanov, 2000). An prototype of the system of electronic testing is presented. The system consists of four components: ontology assistant for creating and editing of domain knowledge bases, question generator providing automated approach to generation of questions, test constructor for building of branching probability-based tests, and examiner, the main component for examination and results analysis. The system is expected to be used for obtaining of experimental data for in-depth analysis and further development of used algorithms and models.

Keywords: natural language processing, e-learning, question-answer system.

АВТОМАТИЗИРОВАННЫЙ АНАЛИЗ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ВОПРОСНО-ОТВЕТНЫХ ТЕКСТОВ В СИСТЕМЕ ЭЛЕКТРОННОГО ТЕСТИРОВАНИЯ

Н. Прокопьев, Д.Ш. Сулейманов
Институт прикладной семиотики Академии наук Республики Татарстан, Казань, Россия
dvdt.slt@gmail.com

Автоматизированный контроль знаний обучающихся с использованием электронного тестирования является важным прикладным направлением в образовании в современную информационную эру. Обычно для реализации такого контроля знаний используются вопросы с вариантами ответа, однако все еще малоизученным, согласно анализу работ (Соловьев, 2011),

(Климов, 2013), (Черепанова, 2013) в этой научной области, остается направление автоматизации проверки вопросов с открытым, естественно-языковым ответом. Такого рода тестирование имеет свое применение, в особенности в гуманитарных и языковых учебных дисциплинах, где результаты тестов по вопросам с вариантами ответа зачастую не дают полной картины об уровне знаний обучающегося. В качестве основной модели при реализации модуля разбора ответов предлагается использовать теоретическую модель, представленную в работе (Сулейманов, 2000). В статье представлен прототип системы электронного тестирования. Система состоит из 4 компонентов: редактор онтологии для создания и редактирования баз знаний предметной области, генератор вопросов, предоставляющий автоматизацию создания вопросов, редактор тестов для создания вероятностных ветвящихся электронных тестов, и экзаменатор, главный компонент для прохождения тестирования и анализа результатов. Систему предполагается использовать для получения экспериментальных данных, глубокого их анализа и дальнейшего развития используемых алгоритмов и моделей.

Ключевые слова: обработка естественного языка, электронное образование, вопросно-ответные системы.

1. Описание модели разбора ответа

Основным методологическим принципом модели является утверждение о том, что заданный вопрос естественным образом ограничивает контекст ответа, как по множеству вариантов ответа, так и по структуре. Из этого принципа следуют принципы реализации, заключающиеся в возможности выделить из ответа семантические структурные единицы, назовем их концептулы, и задать контекстную грамматику правильного ответа, как цепочку концептул, и соответственно, свести задачу семантического анализа к классической задаче синтаксического разбора. Таким образом, реализуется прагматически-ориентированный подход к анализу естественного языка, то есть реализация алгоритма разбора естественного языка, не универсального, а направленного на решение конкретной задачи в условиях вопросно-ответного контекста. Структура модуля разбора представлена на рис. 1.

Лексический процессор получает на вход ответ тестируемого и модель соответствия для заданного вопроса. Модель соответствия представляет собой перечисление соответствий между используемыми в шаблоне ответа концептулами (структурными единицами) и конкретными словами или словосочетаниями из естественного языка, которые ожидаются в ответе.

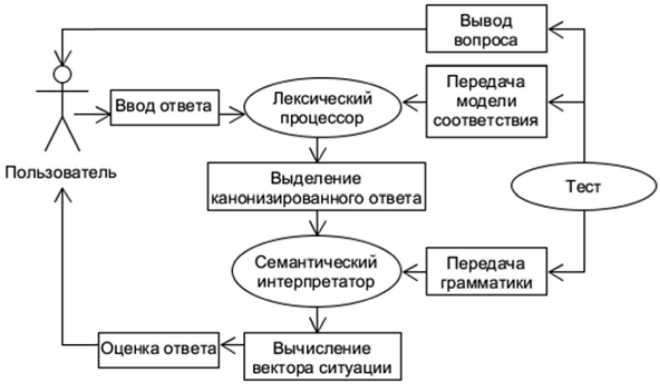


Рис. 1. Структура модуля анализа ответов

Для представления ответа в виде цепочки концептул лексический процессор производит токенизацию ответа согласно полученной модели соответствия. На выходе он формирует канонизированное представление ответа, содержащее представление ответа в виде цепочки концептул, массив запрещенных лексем и массив неопределенных лексем.

На вход семантического интерпретатора поступает канонизированное представление и грамматика ответа на заданный вопрос. Грамматика ответа представляет собой синтаксическое дерево, узлами которого являются концептулы, перечисленные в модели соответствия данного вопроса. Для каждого типа вопросов задается своя грамматика. Интерпретатор проверяет соответствие цепочки концептул грамматике правильного ответа и на выходе формирует полный вектор ситуации. Проверка соответствия ответа грамматике происходит путем попытки обхода данного синтаксического дерева по узлам согласно канонизированному представлению. Если обход завершается на конечном узле дерева, то считается, что ответ соответствует грамматике. При обходе игнорируются неопределенные концептулы. Вектор ситуации, формируемый на выходе, содержит данные о правильности ответа, о соответствии его контексту вопроса, о длине ответа, об использовании запрещенных концептул, о модальности ответа. С помощью данного вектора и производится дальнейшая оценка ответа.

2. Структура системы

Для работы модуля анализа естественно-языковых ответов необходимы некоторые метаданные о вопросах, включающие грамматики правильных ответов, модели соответствия и базы знаний. Соответственно, необходимы компоненты, позволяющие вносить эти метаданные. Естественно, внесение множества метаданных для каждого вопроса является трудоемкой и требует автоматизации. Потому было решено разработать систему с возможностью генерации вопросов со всеми необходимыми метаданными. Предлагаются следующие компоненты системы.

2.1. Редактор онтологии

Данный компонент предназначен для создания и заполнения баз знаний предметной области. Предполагается, что эксперт заполняет базу знаний в соответствии с некоторым учебным материалом. Под предметной областью в данном случае подразумевается замкнутая понятийная область знаний, к примеру, учебник по школьному предмету, дисциплина высшей школы и т.п.

Каждая база знаний представлена в виде реляционной базы данных, хранящейся в файле формата SQLITE. В реляционную структуру погружена онтология предметной области, позволяю-

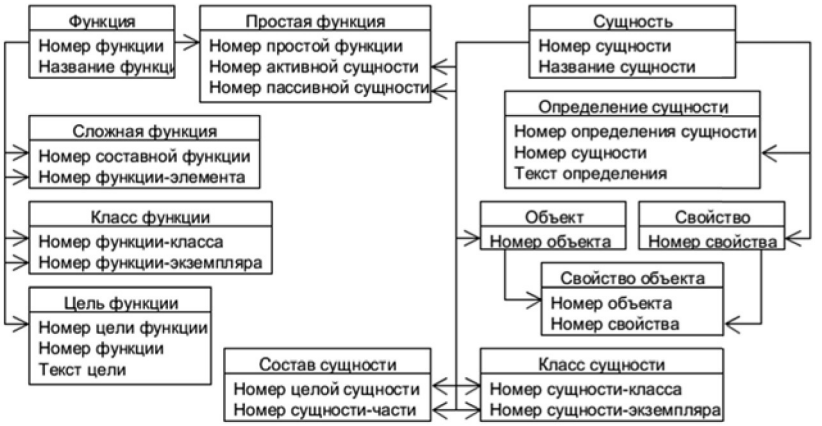


Рис. 2. Структура базы знаний

щая хранить в базе данных словарь понятий из некоторой предметной области (например, из учебной дисциплины) и отношения между понятиями. Представление онтологии предметной области в виде реляционной базы данных основывается на подходе, предложенном в (Казаков, Манцивода, 2011).

Структура одной базы знаний представлена на рис. 2. Таблицы «Сущность», «Объект», «Свойство», «Функция», «Определение сущности», «Цель функции» соответствуют словарям, остальные таблицы соответствуют отношениям между записями из словарей. Такая структура достаточно гибка для обеспечения генерации множества типов вопросов, несмотря на функциональные ограничения, накладываемые на онтологическую схему погружением в реляционную структуру.

2.2. Генератор вопросов

Данный компонент реализует автоматическую генерацию вопросов по заполненной базе знаний. Общий алгоритм данного генератора описан в (Аюпов, Невзорова, Прокопьев, Сулейманов, 2014). Однако, в развитие данной работы, в системе не предполагается генерация по фиксированным шаблонам, заполняемым экспертом, так как предлагаемые в системе типы вопросов не являются часто изменяющимися. Вместо этого предлагается генерировать вопросы при помощи встроенных подпрограмм, различных для каждого типа вопросов. Такой подход повышает гибкость генератора и поддерживает большее количество типов вопросов.

В общем случае при генерации вопроса производится SQL запрос к базе знаний, один или несколько, по некоторому отношению для получения эталонного ответа. Далее если в итоге эталонный ответ получен, происходит замена одного из элементов отношения на случайное вопросное слово из заданного в подпрограмме массива вопросных слов, и приведение полученного текста вопроса к грамматически правильному виду. Таким образом, получается новый вопрос.

С вопросом связываются метаданные для ответа: грамматика правильного ответа в виде синтаксического дерева с концептулами в узлах и модель соответствия, в которой каждой концептуле из грамматики сопоставляется множество нормализованных слов или словосочетаний для конкретного вопроса.

2.3. Конструктор тестов

Данный компонент предназначен для составителей тестов, в нем предоставляется удобный интерфейс для конструирования ветвящихся тестов с вероятностным характером появления вопросов.

2.4. Экзаменатор

Основной компонент для прохождения тестирования, содержащий модуль анализа ответов, оценки, накопления и комплексного анализа результатов.

На основе собранной статистики результатов вычисляется сложность вопросов и составляется карта знаний тестируемого, представляющая собой сопоставление предметной области и некоторой оценки знания, исходящей из 100% уровня освоения предметной области, что эквивалентно получению максимальной оценки за все ответы на встреченные в тестах вопросы из данной области. Эти данные используются, в том числе, для опциональной адаптации теста к способностям тестируемого.

Адаптация заключается в изменении сложности теста путем манипуляций с вероятностями появления вопросов, более простых или более сложных для тестируемого в соответствии с его картой знаний. Модуль адаптации теста включен в данный компонент и реализован с использованием генетического алгоритма.

3. Заключение

Предложенная система включает возможности по обширной автоматизации процессов электронного тестирования, а также предполагает дальнейшее развитие и применение ее в качестве источника экспериментальных данных для улучшения модуля анализа естественного языка, пополнения его новыми типами вопросов, уточнения грамматик. Для данной цели предполагается использование системы в режиме электронного образовательного ресурса.

ЛИТЕРАТУРА

1. Соловьев А. А. (2011). Синтаксические и семантические модели и алгоритмы в задаче вопрос-ответного поиска. Материалы всероссийской научной конференции «Электронные библиотеки: перспек-

тивные методы и технологии, электронные коллекции» – RCDL’2011, Воронеж.

2. Климов А.В. (2013). Разработка методов семантического анализа текстов при тестировании знаний человека. Материалы 51-ой международной студенческой конференции: «Студент и научно-технический прогресс», Новосибирск.

3. Черепанова Ю. Ю. (2013). Контроль знаний с ответами на естественном языке. Восточно-европейский журнал передовых технологий 4/2 (40).

4. Сулейманов Д. Ш. (2000). Системы и информационные технологии обработки естественно-языковых текстов на основе прагматически-ориентированных лингвистических моделей. Диссертация на соискание ученой степени доктора технических наук, Казань.

5. Казаков И. А., Манцивода А. В. (2011). Базы данных как онтологии. Известия Иркутского государственного университета, серия «Математика», Т. 4, №1.

6. Аюпов М. М., Невзорова О. А., Прокопьев Н. А., Сулейманов Д. Ш. (2014). Семантические технологии генерации учебных вопросов. Четырнадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2014: Труды конференции. Т.3. – Казань: Изд-во РИЦ “Школа”.

УДК 81’32

CHARACTER-BASED DEEP LEARNING MODELS FOR TOKEN AND SENTENCE SEGMENTATION

A. Toleu[1,2], G. Tolegen[1,2], A. Makazhanov[1],
[1]National Laboratory Astana, 53 Kabanbay Batyr ave., Astana, 010000, Kazakhstan
[2]Tsinghua University, Department of Computer Science and Technology, Beijing, 100084, China
aibek.makazhanov@nu.edu.kz

In this work we address the problems of sentence segmentation and tokenization. Informally the task of sentence segmentation involves splitting a given text into units that satisfy a certain definition (or a number of definitions) of a sentence. Similarly, tokenization has as its goal splitting a text into chunks that for a certain task constitute basic units of operation, e.g. words, digits, punctuation marks and other symbols for part of speech tagging. As seen from the definition, tokenization is an absolute prerequisite for virtually every natural language processing (NLP) task. Many of so called downstream NLP applications with higher level of sophistication, e.g. machine translation, additionally require sentence segmentation. Thus both of the problems that we address are the very basic steps in NLP and, as such, are widely regarded as solved problems. Indeed there is a large body of work devoted to these problems, and there is a number of popular, highly accurate off the shelf solutions for them. Nevertheless, the problems of sentence segmentation and tokenization persist, and in practice one often faces certain difficulties whenever confronted with raw text that needs to be tokenized and/or split into sentences. This happens because existing approaches, if they are unsupervised, rely heavily on hand-crafted rules and lexicons, or, if they are supervised, rely on extraction of hand-engineered features. Such systems are not easy to maintain and adapt to new domains and languages because for those one may need to revise the rules and feature definitions.

In order to address the aforementioned challenges, we develop character-based deep learning models which require neither rule nor feature engineering. The only resource required is a training set, where each character is labeled with an IOB (Inside Outside Beginning) tag. Such training sets are easily attainable from existing tokenized and sentence-segmented corpora, or, in absence of those, have to be created (but the same is true for rules, lexicons, and hand-crafted features). The IOB-like annotation allows us to solve both tokenization and sentence segmentation problems simultaneously casting them as a single sequence-labeling task, where each character has to be tagged with one of four tags: beginning of a sentence (S), beginning of a token (T), inside of a token (I)

and outside of a token (O). To this end we design three models based on artificial neural networks: (i) a fully connected feed forward network; (ii) long short term memory (LSTM) network; (iii) bi-directional version of LSTM. The proposed models utilize character embeddings, i.e. represent characters as vectors in a multidimensional continuous space.

We evaluate our approach on three typologically distant languages, namely English, Italian, and Kazakh. In terms of evaluation metrics we use standard precision, recall, and F-measure scores, as well as combined error rate for sentence and token boundary detection. We use two state of the art supervised systems as baselines, and show that our models consistently outperform both of them in terms of error rate.

Keywords: Token and Sentence Segmentation; Neural Networks; Deep Learning.

СИМВОЛЬНЫЕ МОДЕЛИ ГЛУБИННОГО ОБУЧЕНИЯ ДЛЯ ГРАФЕМАТИЧЕСКОГО АНАЛИЗА

***А. Толеу*[1,2], *Г. Толеген*[1,2], *А. Макажанов*[1]**
[1]Национальная Лаборатория Астана, пр. Кабанбай батыра 53, Астана, 010000, Казахстан
[2]Университет Цинхуа, Факультет Компьютерных Наук и Технологий, Пекин, 100084, КНР
aibek.makazhanov@nu.edu.kz

В настоящей работе мы рассматриваем задачу графематического анализа, а именно проблемы сегментации текста на предложения и токены. Сегментация текста по предложениям рассматривается как задача нахождения отрывков текста, удовлетворяющих одному или нескольким определениям предложения. Сегментация на токены (токенизация) – задача разбиения текста на операционные единицы, т.е. слова, цифры, знаки препинания и пр. Токенизация является базовой задачей обработки естественного языка (ОЕЯ). Большинство прикладных задач ОЕЯ, отличающихся относительной сложностью, например машинный перевод, нуждаются в сегментации входного текста по предложениям. Таким образом, обе рассматриваемые нами задачи являются основополагающими для ОЕЯ, и, как следствие, считаются в достаточной степени решенными. Действительно, опубликовано немало исследований по данной тематике, и существуют готовые решения широкого применения с хорошей точностью. Тем не менее, проблемы графематического анализа в большинстве случаев остаются открытыми, и на практике с ними приходится сталкиваться каждый раз, когда появляется необходимость в работе с необработанным текстом, т.е. не разбитым на предложения и токены. Это происходит потому, что существующие под-

ходы основаны либо на словарях и правилах (необучаемые), либо на извлечении вручную заданных признаков (обучаемые). Такие подходы тяжело адаптировать к новым языкам/жанрам, так как это требует переопределение словарей, правил и признаков.

Для снятия вышеупомянутых ограничений мы разработали символьные модели глубинного обучения, которые не нуждаются в определении правил или признаков. Единственное в чем есть необходимость – это обучающая выборка, в которой каждый символ помечен IOB меткой. Подобные обучающие выборки легко получить из имеющихся сегментированных и токенизированных корпусов. В случае отсутствия последних обучающую выборку придется создавать вручную, как в прочем, и словари и правила для других методов. Использование IOB разметки позволяет решать обе задачи одновременно, как одну задачу разметки последовательности, цель которой присвоить каждому символу одну из четырех меток: начало предложения (S), начало токена (T), тело токена (I), или пробел (O). Для решения данной задачи мы разработали три модели, основанные на искусственных нейронных сетях: (1) поступательная сеть; (2) LSTM сеть; (3) двунаправленная LSTM сеть. Разработанные модели используют символьные вложения, т.е. представления символов в виде векторов в многомерном пространстве.

Мы оцениваем наш подход на трех типологически отдаленных языках: английском, итальянском и казахском, используя стандартные метрики точности, покрытия, F-меры и процента ошибки. Для сравнения мы используем две широко распространённые системы графематического анализа, и показываем, что обе уступают нашим моделям по метрике процента ошибки.

Ключевые слова: Графематический анализ; нейронные сети; глубинное обучение.

1. Introduction

Let us begin by a quick recap of definitions. Sentence segmentation, aka sentence boundary detection, is a problem of segmenting a text into sentences for further processing; and tokenization is a problem of segmenting a text into chunks that for a certain task constitute basic units of operation (e.g. words, digits, etc.). At a first glance the problems seem trivial; after all, most written languages use special symbols to terminate sentences and whitespaces to delimit words. This is however not always the case.

First, although for many languages sentence final punctuation consists of a period (dot), a question and an exclamation mark, some languages use different sets of symbols (Brown, 2017). Second, regard-

less of symbols used as delimiters in any given language, chances are that those symbols have other functions as well, e.g. periods (dots) may be used in abbreviations, initials or in numbers as decimal points. Third, sentence and token definitions depend on the task at hand. For instance, while sentence segmentation may not be needed and a simple whitespace tokenization may be enough for a bag of word-based document classification, for parsing one may need to consider multiple sentence utterances in direct speech as a part of a host sentence (sentences in a sentence) and count clitics (syntactic words usually delimited with hyphens and apostrophes, but not whitespaces) as separate tokens. Thus, to solve sentence and token segmentation problems one cannot blindly segment texts at the occurrences of certain symbols, and has to resort to a more sophisticated approach.

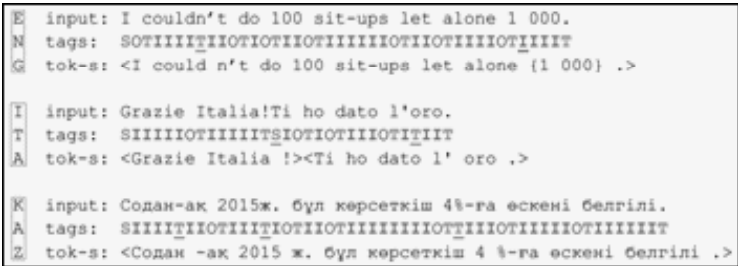


Fig. 1. An Example of an IOB-labeled text in English, Italian, and Kazakh

In this work we cast the token and sentence segmentation (TSS) problems as a single sequence labeling task and propose artificial neural network-based solutions, namely three character-based deep learning models. Unlike much of the previous work, our approach requires neither rule nor feature engineering. The only resource required is a training set, where each character is labeled with an IOB (Inside Outside Beginning) tag. Performing TSS jointly and using IOB-like format is not, in itself, a novelty, Evang et al. (2013) have implemented this approach in their CRF-based system called Elephant. However, unlike Elephant, our models make use of character embeddings, i.e. map characters into continuous vector space, and make no use of predefined features. Experiments show that our models achieve top performance for Kazakh language, for which TSS evaluation has never been carried out before. In order to show that the proposed models can achieve competitive results we compare them to a popular TSS system

Punkt (Kiss and Strunk, 2006) and the aforementioned Elephant system, which is considered the state-of-the-art. For this experiment we use publically available data sets for English and Italian languages.

2. Related Work

Existing systems for token and sentence boundary detection are based on hand written rules, unsupervised and supervised learning approaches. Rule-based systems (Grefenstette, 1999; Jurafsky and Martin, 2008; Dridan and Oepen, 2012) utilize hand-written rules, fixed lists of abbreviations and other lexical items to detect sentence boundaries. As a result such approaches are hard to maintain and not easy to adapt to new languages (Silla Jr. and Kaestner, 2004) or domains.

Unsupervised learning systems do not require specific hand-coded regular expressions and annotated training data. Mikheev (2002) presented an unsupervised approach for sentence boundary detection, proper name identification and abbreviation detection. The proposed system achieved respective error rates of 1.41% and 0.65% on WSJ and Brown corpora. The author concluded that the most crucial factor for sentence segmentation was detection of abbreviations and proper names. A similar system called Punkt was proposed by Kiss and Strunk (2006). The approach here has two detection stages: abbreviation detection and token-based classification. This system reached high accuracy, rivaling handcrafted rule-based and unsupervised systems. Compared with Mikheev's system, Punkt's error rates on WSJ and Brown corpora were 1.65% and 1.02%, respectively.

Supervised learning approaches utilize hand-engineered features, such as POS tags, tokens neighboring potential sentence boundaries, abbreviation lists, letter case (lowercase, uppercase), etc. These systems utilized maximum entropy models (Reynar and Ratnaparkhi, 1997) and conditional random fields (Fares et al., 2013). Many works have shown that conditional random field (CRF) is the most popular model for sequence labeling tasks (Lafferty et al., 2003; Tolegen et al., 2016).

Evang et al. (2013) presented a CRF-based TSS system – Elephant that uses a single character as a basic unit of operation. The system uses several features, such as Unicode categories, Unicode character codes, and combination of the two, as well as the 10 most active outputs of learned hidden states of a deep learning model as one feature category. Unlike our approach, Elephant uses the discrete features

rather than distributed embedding features. Numerous works on deep learning for NLP have shown the advantage of embeddings that tend to capture meaningful information and reduce task-specific features engineering.

3. Method

3.1 IOB labeling

In order to jointly learning one model for two tasks, we adopted IOB tagging scheme to identify the boundaries of the tokens and sentences. An example is given in Fig. 1. The tags S and T denote the beginnings of sentence and token boundaries respectively. Inside of a token is labeled I, and outside as O. Passages included in "<" and ">" denote segmented sentences. In the given example whenever tokens and sentence boundaries are not preceded by an outside character (O) they are underlined.

3.2 A general neural network

We introduce a general neural network model (Collobert et al., 2011) for token and sentence boundaries detection. The model is usually characterized by three specialized layers: (i) a character look-up table layer that extracts a window of character's embeddings from a character parameter matrix; (i) a general hidden layer; (iii) one output layer that is used to compute normalize scores for labels. The model architecture is shown on Fig. 2 and in what follows we refer to this model as NN.

Character look-up table. Let C be the list of characters derived from training data, d be the dimension of character embeddings, $Q \in R^{d \times |C|}$ be the matrix of character's embeddings. Suppose that a string s is made up of a sequence of characters $[c_1,\ldots,c_l]$, where l is the length of string. Then the character-level representation of s of is given by the matrix Q, where the j-th column of matrix Q corresponds to the character embedding for C_j. We use a sliding window approach to get a fixed sized w (a hyper-parameter) window of character embeddings around current character. Each character in the window is first passed through the look-up operation which produces a matrix of character embeddings that can be viewed as a w×d-dimensional vector x by concatenating each column vector, which can be fed into the next layers.

Hidden layer. The embedding of characters $x \in R^{w \times d}$ is extracted from the look-up table and is fed into a hidden layer which performs non-linear transformation followed by an element-wise activation function σ such as *tanh*, and the computation of this layer is:

$$h = \sigma(W_1 x + b_1) \quad (1)$$

where $W_1 \in R^{H_1 \times wd}$ is the parameter, $b_1 \in R^{H_1 \times 1}$ is a bias term, $h \in R^{H_1}$ is hidden units, H_1 is dimension of hidden layer.

The output layer is finally added on the top of the hidden layer for scoring boundary labels:

$$Score\ (x,T,\theta) = softmax(W_2\ h + b_2) \quad (2)$$

where $Score(x,y,\theta) \in R^{|T| \times 1}$ is a score of labels that computed by neural network with parameters $\theta = \{Q, W_1, b_1, W_2, b_2\}$, |T| is the number of tags. The parameters of models are initialized to small random numbers and automatically trained the by back-propagation algorithm.

3.3 Bi-directional LSTMs

Recently, LSTM neural networks have shown great promise in many NLP tasks (Greff et al., 2015; Ling et al., 2015; Toleu et al.,

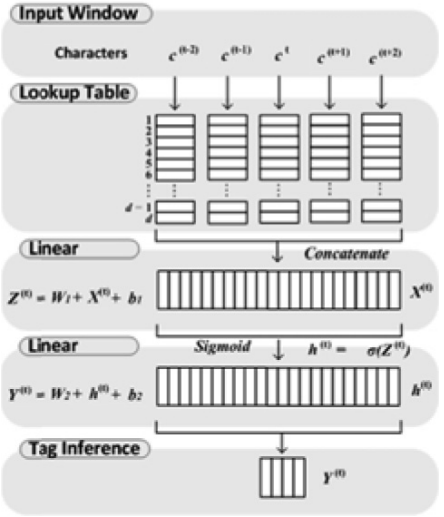


Fig. 2. Model architecture

2017) including language modelling, part-of-speech tagging etc. The architecture of LSTM consists of a set of recurrently connected states that can be viewed as memory blocks. Each block contains certain self-connected memory cells and three gates: input, output and forget gate. The gates provide continuous analogues of write, read and reset operation for the cells.

In order to examine the effectiveness of LSTM network for TSS, we use a model to predict each boundary label using LSTM. The architecture of our LSTM-based network is a variant that was described by Graves and Schmidhuber (2005), and is frequently cited in the literature.

Given a string made up of a sequence of characters , we encode each character into a vector representation then feed into our LSTM-based models, computing the forward hidden state and the backward hidden state. Both hidden states are concatenated into a single vector and fed into the output layer. In what follows we refer to this model as bi-LSTM, the model only uses the forward hidden states as LSTM. The architecture of the model is shown on Fig. 3.

4. Experiments

4.1. Data sets

The experiments were conducted on three datasets: (i) Kazakh texts from Kazakh corpus (Makhambetov et al., 2013) and UD treebank (Makazhanov et al., 2015); (ii) English newswire texts taken from the Groningen Meaning Bank, GMB (Basile et al., 2012); (iii) Italian texts from the PAIS`A corpus (Borghetti et al., 2011). Each dataset was split into three parts: a training set, a validation set, which is used for early stopping to select the best model and for optimizing the hyper-parameters, and a test set used for the final evaluation. Kazakh data needed additional processing as it was not IOB-labeled. We have performed an automatic IOB labeling based on existing token and sentence segmentations. Table 1 provides statistics on the domains of the texts and data quantities in terms of numbers of sentences and tokens.

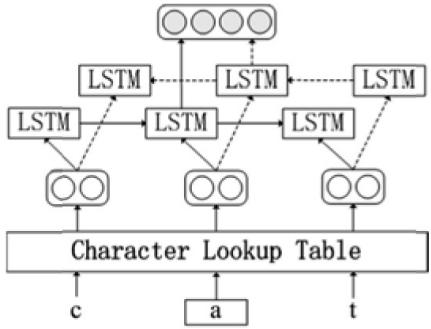


Fig. 3. Architecture of bi-LSTM-based model

Table 1. **Characteristics of the data sets**

Language	Domain	# sentences	# tokens
Kazakh	web/various	4 360	96,760
English	newswire	2 886	64,443
Italian	web/various	42 674	869,095

4.2. Model setup

We implement all neural network models using Java programming language and use the same hyper-parameters in all of three models: 35 for character level embeddings with random initialization, 9 for window size, 100 hidden states. We run 300 epochs on training and development sets, and select one model that is optimized on evaluation over the development set. The selected model is applied to the test set for the final evaluation. We used the CoNLL evaluation script to report, accuracy, precision, recall and F-measure over the token and sentence boundary labels.

4.3. Results

As it can be seen from Table 2, a general neural network model (NN) achieves a perfect 100 on all metrics, and clearly outperforms LSTM-based models in the task of sentence boundary detection for English language. One possible explanation is that the model NN has a window (the size is 9) to capture some corresponding characters and

predict a label to the centered one character, in this case, the prediction is made by conditioning on the left and right 4 characters. As our preliminary experiments showed that taking a smaller or larger window size, it harms the model performance on the sentence boundary label, but for token boundary, it does not have a significant effect.

Table 2. **Evaluation results for English**

	Sentence segmentation			Tokenization		
Models	Precision	Recall	F-measure	Precision	Recall	F-measure
NN	**100**	**100**	**100**	99.92	99.82	99.87
LSTM	99.34	99.34	99.34	99.94	**99.86**	**99.90**
bi-LSTM	99.67	99.34	99.50	**99.95**	**99.86**	**99.90**

On the other hand, the LSTM-based models achieve marginal improvement of the NN model in tokenization. In general all of the three models achieve near perfect results on the English data set.

Table 3. **Evaluation results for Italian**

	Sentence segmentation			Tokenization		
Models	Precision	Recall	F-measure	Precision	Recall	F-measure
NN	**99.28**	96.32	97.78	99.63	99.78	99.70
LSTM	99.00	96.27	97.62	99.52	99.71	99.61
bi-LSTM	99.25	**96.76**	**97.99**	**99.74**	**99.86**	**99.80**

As shown in Table 1, the size of the Italian data set is more than ten times larger than that of English and eight times larger than that of Kazakh. It is interesting to see the performance of neural network models for token and sentence boundary detections given larger training data. As evident from Table 3, the bi-LSTM model benefited the most from the abundance of data and was second to the NN model only in terms of precision of sentence segmentation. In general for the Italian language sentence segmentation turned out to be less accurate compared to English, but tokenization is still at the acceptable 99.8% in terms of F-measure.

From Table 4 one can observe that for Kazakh language the NN model detects sentence boundaries more accurately even though the other models use the same context window size (from the preliminary experiments, we observed that all of the LSTM-based models gave lower results without using a context window). This model has the highest recall in the tokenization task. As we have learned from the experiment on Italian, LSTM-based models are more sensitive to the size of training data, and thus maybe performing lower on a relatively small data set. In general all of the models exhibit a significantly lower performance on Kazakh data set. This can be explained by the fact that a large portion of this data set came from the Web (cf. Table 1), a notoriously noisy source. While the Italian data set contains certain amount of Web texts as well, this data set as a whole is much larger than the Kazakh one. Thus, we speculate that the fact that the data set was noisy and small may have hindered the performance of the models.

Table 4. **Evaluation results for Kazakh**

	Sentence segmentation			Tokenization		
Models	Precision	Recall	F-measure	Precision	Recall	F-measure
NN	**92.70**	**99.44**	**95.95**	99.74	**99.44**	99.59
LSTM	92.43	97.95	95.11	99.58	99.43	99.50
bi-LSTM	92.20	99.25	95.60	**99.82**	99.40	**99.61**

Table 5. **Comparison with other systems**

	English		Italian	
Models	Sentence (F-measure)	Sent. + Tok. (error rate)	Sentence (F-measure)	Sent. + Tok. (error rate)
Punkt	98.51	–	98.34	–
Elephant	**100**	0.27	**99.51**	0.76
NN	**100**	0.05	97.78	0.12
LSTM	**100**	**0.03**	97.62	0.13
bi-LSTM	**100**	**0.03**	97.99	**0.07**

In order to assess the performance of our models relative to existing systems we compare the performance of our models to the results reported by Evang et al. (2013) for their system Elephant and for another popular TSS system, Punkt (Kiss and Strunk, 2006). That is to say, we do not actually run and evaluate those systems. Instead we run our systems on the data that were used in the original experiment (Evang et al., 2013) and compare the results to the ones reported in that original experiment. The comparison is carried out in terms of F-measure of sentence boundary detection and combined (sentence and token segmentation) error rate. The results of the comparison are given Table 5.

As it can be seen, for English all of the models achieve a perfect F-measure of 100% on the sentence segmentation task, except Punkt that performs at 98.51%. When it comes to the combined TSS error rate our LSTM-based models achieve the lowest score of 0.03, improving 9 times over the state-of-the-art system, Elephant. When it comes to sentence boundary detection for Italian, however, our models are outperformed by both of the baseline systems. Here Elephant achieves a very strong F-measure of 99.51%, Punkt yields 98.34%, and the best of our models, bi-LSTM, performs at a decent 97.99%. Nevertheless, as it was the case with English, in terms of error rates for both token and sentence segmentation, our models perform much better, yielding the scores of 0.12, 0.13, 0.07 for NN, LSTM and bi-LSTM (without using any external features) respectively. Here the best performing model, bi-LSTM, improves almost 11 times over Elephant, whose error rate was 0.76. These results indicate that character-based deep learning models are better at modeling token boundary detection and also give very competitive results for sentence segmentation.

5. Conclusion

We have presented character-based deep learning models for joint token and sentence boundary detection. The main advantage of our approach is that it does not require any manual rule and feature engineering, and as such, is easy to maintain and adapt to new languages/domains. We have carried out both an absolute and comparative evaluation of our models on three languages (Kazakh, English and Italian). Our experiments showed that the proposed models achieve competitive results when compared to the state-of-the-art systems.

Acknowledgements

This work has been supported by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan under the targeted program O.0743 (0115PK02473), and by the Nazarbayev University under the research grant №129-2017/022-2017.

REFERENCES

1. Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In LREC 2012, pages 3196–3200, Istanbul, Turkey.

2. Claudia Borghetti, Sara Castagnoli, and Marco Brunello. 2011. I testi del web: una proposta di classificazione sulla base del corpus PAISA`. InM. Cerruti, E. Corino, and C. Onesti, editors, Formale e informale. La variazione diregistro nella comunicazione elettronica, pages 147–170. Carocci, Roma.

3. Brown T. Julian. Punctuation. https://www.britannica.com/topic/punctuation. Last accessed 09.09.2017.

4. Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, pages 2493–2537.

5. Rebecca Dridan and Stephan Oepen. 2012. Tokenization: Returning to a long solved problem – a survey, contrastive experiment, recommendations, and toolkit –. In ACL2012 (Volume 2: Short Papers), pages 378–382, Jeju Island, Korea. Association for Computational Linguistics.

6. Kilian Evang, Valerio Basile, Grzegorz Chrupała, and Johan Bos. 2013. Elephant: Sequence Labeling for Word and Sentence Segmentation. In EMNLP 2013, pages 1422–1426, Washington, USA.

7. Murhaf Fares, Stephan Oepen, and Zhang Yi. 2013. Machine learning for high-quality tokenization – replicating variable tokenization schemes. In CICLING 2013, pages 231–244.

8. Graves, A, Liwicki, M, Fernandez, S, Bertolami, R, Bunke, H, and Schmidhuber, J. A Novel Connectionist System for Improved Unconstrained Handwriting Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(5), 2009.

9. Klaus Greff, Rupesh Kumar Srivastava, Jan Koutn´ık, Bas R. Steunebrink, and Jurgen Schmidhuber. ¨ 2015. LSTM: A Search Space Odyssey. CoRR abs/1503.04069.

10. Gregory Grefenstette. 1999. Tokenization. In Hans van Halteren, editor, Syntactic Wordclass Tagging, pages 117–133. Kluwer Academic Publishers, Dordrecht.

11. Daniel Jurafsky and James H. Martin. 2008. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, 2nd edition.

12. Kiss, Tibor, and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. Computational Linguistics 32.4, pages 485–525

13. John D. Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In ICML ’01, pages 282–289.

14. Wang Ling, Chris Dyer, Alan W. Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Lus Marujo, and Tiago Lus. 2015. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In EMNLP2015, pages 1520–1530.

15. Makazhanov A., Sultangazina A., Makhambetov O., Yessenbayev Z. 2015. Syntactic Annotation of Kazakh: Following the Universal Dependencies Guidelines. A report. TurkLang 2015, pages 338–350.

16. Makhambetov O., Makazhanov A., Yessenbayev Z., Matkarimov B., Sabyrgaliyev I., Sharafudinov A. 2013. Assembling the Kazakh Language Corpus. In EMNLP2013, pages 1022–1031.

17. Andrei Mikheev. 2002. Periods, capitalized words, etc. Computational Linguistics, 28(3):289–318.

18. Carlos N. Silla Jr. and Celso A. A. Kaestner. 2004. An analysis of sentence boundary detection systems for English and Portuguese documents. In CICLing 204, pages 135–141.

19. Toleu, Alymzhan and Tolegen, Gulmira and Makazhanov, Aibek. 2017. Character-Aware Neural Morphological Disambiguation. In Proceedings of ACL 2017, pages 666–671,Vancouver, Canada.

20. Gulmira Tolegen, Alymzhan Toleu, and Zheng Xiaoqing. 2016. Named Entity Recognition for Kazakh Using Conditional Random Fields. In Proceedings of TurkLang 2016, pages 122–129.

21. Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In Proceedings of ANLP97, pages 16–19.

СЕКЦИЯ 2 | **РЕЧЕВЫЕ ТЕХНОЛОГИИ**

УДК 004.934.2

REGARDING THE IMPACT OF KAZAKH PHONETIC TRANSCRIPTION ON THE PERFORMANCE OF AUTOMATIC SPEECH RECOGNITION SYSTEMS

M. Karabalayeva[1], Zh. Yessenbayev[1], Zh. Kozhirbayev[1,2]
[1]National Laboratory Astana, Nazarbayev University, 53 Kabanbay batyr Ave., Astana 010000, Kazakhstan
[2]Faculty of Information Technologies, L.N. Gumilyov Eurasian National University, 2 Satpayev Str., Astana 010008, Kazakhstan
muslima.karabalayeva@nu.edu.kz

Over the past decades automatic speech recognition has made remarkable advances, in both theoretical and practical aspects. Evolution of research in this field has been proceeding from the recognition of individual sounds and phonemes to the recognition of continuous and mixed speech, including tasks of automatic transcription of broadcast news and telephone conversations. Despite the high performance of continuous speech recognition systems, which makes up to 95%, the performance of phoneme recognition systems remains below 85%. However, phoneme recognition is widely used in a number of applications, such as spoken term detection, language identification, speaker identification and others.

The paper presents the results of the experiments on continuous Kazakh speech recognition using different phoneme sets and alternative phonetic transcriptions. This study was instigated by the fact that in modern Kazakh linguistics there is no common agreement about the phonetic system of the Kazakh language, while the list of phonemes and their number noticeably vary in different textbooks. Therefore, we aimed our experiments to study the impact of the phonetic system of the language, its orthoepic rules and the corresponding phonetic transcriptions on the performance of the phoneme recognition systems, which are the initial stage in the general systems of continuous speech recognition.

The following 6 systems of phonetic transcription have been considered and tested in our study. The first one is a project of the new Kazakh alphabet and a set of spelling rules proposed by Prof. A. Sharipbay. The second system is a set of orthoepic rules for the actual Kazakh Cyrillic alphabet, introduced by Kazakh linguists – the authors of the Kazakh "Orthoepical Dictionary". The third one of the systems considered is a phonetic system and a set of empirical transcription rules used by the authors of this work in their studies. The fourth variant is based on the actual Kazakh Cyrillic alphabet without taking into account any orthoepic rules, i.e. a transcription system in which one letter corresponds to one phoneme. The remaining two systems are variations or combinations of these systems mentioned above.

In total, three series of experiments were conducted: word-based recognition and two series of phone-based recognition. The latter two differ in test sets. Word-based experiments did not reveal any special differences in the recognition performance among the systems studied, which is due to the strong impact of the language model on the decoding process. On the contrary, phone-based experiments showed that: 1) the existing orthoepic rules are not fully adequate to the actual sounding of Kazakh speech; 2) the existing phonetic system of the Kazakh language can be optimized by removing some phonemes. The speech recognition system is implemented using the Kaldi platform.

Despite the fact that the present work is a preliminary study, on the whole, the presented experimental results make it possible to evaluate the adequacy of considered phonetic transcriptions to the actual sounding of Kazakh speech, and can be of particular interest in view of the forthcoming transformation of the Kazakh writing system.

Keywords: automatic speech recognition; phoneme recognition; phonetic transcription; Kazakh speech.

О ВЛИЯНИИ ФОНЕТИЧЕСКОЙ ТРАНСКРИПЦИИ КАЗАХСКОГО ЯЗЫКА НА КАЧЕСТВО АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

М. Х. Карабалаева, Ж. А. Есенбаев[1], Ж. М. Кожирбаев[1,2]

[1]Национальная Лаборатория Астана, Назарбаев Университет, Астана, Казахстан

[2]Факультет информационных технологий, Евразийский университет имени Л.Н. Гумилева, Астана, Казахстан

muslima.karabalayeva@nu.edu.kz

Статья описывает результаты экспериментов по машинному распознаванию казахской речи с использованием разных систем фонетической транскрипции казахского языка. Эксперименты проведены на небольшом

речевом корпусе казахских предложений, озвученных дикторами в студийных условиях.

Целью экспериментов является сравнительный анализ нескольких систем фонетической транскрипции казахского языка, основанных на различном количестве фонем и использующих различные наборы орфоэпических правил. Результаты описанных экспериментов позволяют сделать выводы об адекватности зафиксированных орфоэпических правил реальному звучанию казахской речи.

Полученные результаты представляют особый интерес в связи с актуальной задачей реформирования казахской письменности, которое неизбежно затронет и фонетику, и орфоэпию казахского языка.

Ключевые слова: автоматическое распознавание речи; пофонемное распознавание; фонетическая транскрипция; казахская речь.

Over the past decades automatic speech recognition has made remarkable advances, in both theoretical and practical aspects. Evolution of research in this field has been proceeding from the recognition of individual sounds and phonemes to the recognition of continuous and mixed speech, including tasks of automatic transcription of broadcast news and telephone conversations. Despite the high performance of continuous speech recognition systems, which makes up to 95%, the performance of phoneme recognition systems remains below 85% (Lopes, Perdigão, 2011; Graves, Mohamed, Hinton, 2013). This is due to the fact that modern speech recognition systems are based on language models that are able to overcome the errors of the first stage of recognition, which is phonetic transcription. However, such systems have certain limitations imposed on the system's lexicon. Therefore, with no detraction from the merits of such systems, let us note that the task of recognizing phonemes by itself still remains relevant. This task has an important application value for speech recognition systems, which are constructed solely for the solution of this problem and which do not have those vocabulary restrictions mentioned above. Besides these continuous speech recognition systems, phoneme recognition is also widely used in a number of applications, such as spoken term detection (Schwarz, 2008), language identification (Matejka, 2009), speaker identification and others (Furui, 2005).

The present work is aimed to perform a comparative analysis of a few phoneme recognition systems for continuous Kazakh speech, which are based on different phoneme sets and different orthoepic rules, which, in their turn, generate alternative phonetic transcriptions necessary for constructing speech recognition systems.

This study was instigated by the fact that in modern Kazakh linguistics there is no common agreement about the phonetic system of the Kazakh language, while the list of phonemes and their number noticeably vary in different textbooks (Sharipbay, 2017, p. 20–22). Therefore, within the framework of our study, we considered several phonetic transcription systems proposed by different authors.

One of these is a project of the new Kazakh alphabet and a set of spelling rules proposed by Prof. A. Sharipbay in an attempt to introduce an optimal phonetic system as part of the transformation of the Kazakh writing system (Sharipbay, 2017). Another system is a set of orthoepic rules for the actual Kazakh Cyrillic alphabet, introduced by Kazakh linguists – the authors of the Kazakh "Orthoepical Dictionary" (Aitbaiuly et al., 2007). Yet another one of the systems considered is a phonetic system and a set of empirical transcription rules used by the authors of this work in their studies. There also was tested a variant on the basis of the actual Kazakh Cyrillic alphabet without taking into account any orthoepic rules, i.e. a transcription system in which one letter corresponds to one phoneme. The remaining two systems are variations or combinations of these systems mentioned above.

The experiments were conducted using high-quality speech signals recorded in unified studio conditions. This allowed us to avoid system errors, which could arise due to the diversity of acoustic conditions and the dissimilar quality of the sound recording equipment. The speech recognition system is deployed on the Kaldi platform (Povey et al., 2011), using a cluster for distributed computing managed by the batch-queuing system Univa Grid Engine, formerly the Sun Grid Engine (Univa Grid Engine, 2017).

The results of the described experiments make it possible to draw conclusions about the adequacy of the considered orthoepic rules to the actual sounding of Kazakh speech. The obtained results are of particular interest in view of the actual problem of the transformation of the Kazakh writing system, which will inevitably affect both phonetics and orthoepy of the Kazakh language.

1. Description of the speech corpus

The experiments have been conducted on two compact-sized speech corpora of the Kazakh sentences *kazspeech* (Makhambetov et al., 2013) and *enuspeech* (Tech. Report, 2014), recorded from native

speakers in a sound recording studio. The total duration of these audio files in the two corpora makes up 1832 minutes of speech (more than 30 hours). The whole set was divided into a training set, a development set, and a test set in a ratio of 81% / 10% / 9%, regarding the duration of audio files. At the same time, the ratio of male and female voices in each of the three sets has been kept approximately equal. The characteristics of the speech corpus are shown in Table 1. The last column contains the values of the OOV (out-of-vocabulary) rate, which is computed as the ratio of the number of unknown words in a set (i.e. words not found in the system's lexicon) to the total number of words in the set.

Table 1. **Characteristics of the speech corpus**

Set	Number of sentences	Duration (*hr*)	Number of speakers	Male speakers	Female speakers	OOV Rate
Train set	15937	24.8	179	76	103	0 %
Dev set	1875	2.92	21	9	12	19.67 %
Test set	1875	2.82	21	9	12	18.75 %

Each uttered sentence in the corpus has its corresponding orthographic transcription, which is the text of the sentence written in the actual Kazakh Cyrillic alphabet.

2. Description of the phonetic transcription systems

Each transcription rule is a single text line representing a rule of substitution. A rule of substitution consists of three parts: the substituted text, the equal sign, and the new text. For example:

```
Һ=Х
ЬЕ=ЙЕ
Ь=
```

Here it is possible to use regular expressions (Regular expression operations, 2017) and syntax of 'sed' scripts ('sed' manual, 2017). For example:

```
ГЫ$=ГІ
\(.\+\)Б$=\1П
```

If several transcription alternatives are available for the text being substituted, then the rule of substitution contains two or more equal signs. For example:

СБ=СП=ЗБ
ТБ=ТП=ДБ

A set of transcription rules is applied line by line to the input text, which is an orthographic transcription of the uttered sentences – a sequence of letters, while the output is its phonetic transcription – the sequence of phonemes. In this procedure, the order of the rules in the list matters, since earlier rules are applied earlier. And if you change the order of the rules in the list, the output transcription may also change.

The following 6 systems of phonetic transcription have been considered and tested in our study.

1.1. Grapheme = phoneme

This system of phonetic transcription is the most primitive, and it implies that the spelling of Kazakh words exactly matches their pronounce.

The phoneme list contains 42 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ё, Ж, З, И, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, Ұ, Ү, Ф, Х, Һ, Ц, Ч, Ш, Щ, Ъ, Ы, І, Ь, Э, Ю, Я. This list coincides with the 42 letters of the actual Kazakh Cyrillic alphabet. Let us note that although Ъ (the hard sign) and Ь (the soft sign) are not formally phonemes, in our experiments they were modeled as separate phonetic units.

There are no transcription rules in this system.

1.2. Grapheme = phoneme, without Ё, Ъ, Ь

This system is similar to the previous one, but excludes three symbols: Ё, Ъ and Ь.

The letter Ё in the Kazakh language is found only in words adopted from Russian. In Russian, the usage of this letter in written texts is not mandatory. In the overwhelming majority of cases, Ё is substituted by Е in written texts, and this is exactly what happened to the orthographic transcriptions in our speech corpus. As a result, in the texts of the training set, the letter Ё occurs only three times (in words *шахтёрлер*, *актёрлердің* and *Гёте*), whereas in all other cases it is most likely written as Е.

The symbols Ъ and Ь are excluded from the system, since they do not denote any separate phoneme.

The phoneme list contains 39 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ж, З, И, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, Ұ, Ү, Ф, Х, Һ, Ц, Ч, Ш, Щ,

Ы, І, Э, Ю, Я. This list coincides with the letters of the actual Kazakh Cyrillic alphabet, excluding symbols Ё, Ъ, Ь.

The transcription rules are:

Ё=Е

Ъ=

Ь=

This list contains only three transcription rules, meaning that the letter Ё should everywhere be replaced by Е, and the letters Ъ and Ь should be removed from the phonetic transcription: *шахтёрлер* [*шахтерлер*], *Ельцин* [*елцин*], *съезд* [*сезд*].

Here is a few sample transcribed words:

СУБЪЕКТІЛЕРІ → СУБЕКТІЛЕРІ

АНСАМБЛЬДІҢ → АНСАМБЛДІҢ

ФЕЛЬДФЕБЕЛЬ → ФЕЛДФЕБЕЛ

Sample source sentence:

ОСЫНДАЙ КҮДІКТІ ОПЕРАЦИЯЛАРДЫ ҚАРЖЫ МОНИТОРИНГІ СУБЪЕКТІЛЕРІ ОЛАР АНЫҚТАЛҒАН СӘТТЕН БАСТАП КОМИТЕТКЕ ТАПСЫРАДЫ

Sample transcribed sentence:

ОСЫНДАЙ КҮДІКТІ ОПЕРАЦИЯЛАРДЫ ҚАРЖЫ МОНИТОРИНГІ СУБЕКТІЛЕРІ ОЛАР АНЫҚТАЛҒАН СӘТТЕН БАСТАП КОМИТЕТКЕ ТАПСЫРАДЫ

1.3. Basic transcription

We gave the name of “basic transcription” to our own simple phonetic transcription system, which, firstly, takes into account the nascence of diphthongs, and, secondly, reduces the number of phonemes by joining rare phonemes into one class with other more frequent phonemes.

As used herein, the term “basic transcription” implies that this transcription system includes, in our opinion, the minimum necessary list of common transcription rules, applicable to the vast majority of words of the Kazakh language, both native and loan. And, conversely, this basic transcription does not include rules that describe nuances and complicated orthoepic cases. Also, the basic transcription does not consider cases when one word might have several alternative phonetic transcriptions.

For instance, if the letter Е is found at the beginning of the word, then it is always pronounced as the diphthong ЙЕ (*ел* [*йел*], *емес* [*йемес*]). In the same way, it is pronounced as the diphthong ЙЕ when found after a vowel, the hard sign or the soft sign (*ниет* [*нийет*], *пьеса* [*пйеса*]).

Likewise, the letter Ю, when found at the beginning of the word, after a vowel, the hard sign or the soft sign, is pronounced as ЙУ or ЙҮ (*Юрмала* [*йурмала*], *көбею* [*көбейү*]). When found after consonants, it gets close to the sound Ү (*жюри* [*жүри*], *эволюция* [*эволүция*]).

In the similar way, the letter Я at the beginning of the word, after vowels or the soft sign can be transcribed by the diphthong ЙА or ЙӘ (*қияр* [*қыйар*], *паэлья* [*паэлйә*]). After consonants, it gets close to the sound Ә (*лямбда* [*ләмбда*]).

Thus, we can remove the symbols Ю and Я from the phoneme list. Rarely used phoneme Һ can be joined into one class with the phoneme Х. Also, we can join the phoneme Щ into one class with the phoneme Ш, because Щ is the soft allophone of Ш.

In the basic transcription we introduced an additional symbol W to denote the short consonant У, which does not make a syllable and which can be considered as part of some diphthongs. For instance, it can be found between two vowels (*ауа* [*awa*]), at the beginning of words (*уақыт* [*wақыт*], *уәде* [*wәде*]), and also it often occurs at the beginning of words when the first letter is Ө or Ө (*он* [*wон*], *өйткені* [*wөйткені*]).

The phoneme list contains 36 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ж, З, И, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, W, Ұ, Ү, Ф, Х, Ц, Ч, Ш, Ы, І, Э.

The transcription rule list contains 139 rules. If necessary, the complete list of all transcription rules mentioned in the present paper can be provided upon request.

Sample transcribed words:

СУБЪЕКТІЛЕРІ → СУБЙЕКТІЛЕРІ

ОЙНАУҒА → WОЙНАWҒА

ОКТЯБРЬСК → WОКТӘБРСК

Sample transcribed sentence:

WОСЫНДАЙ КҮДІКТІ WОПЕРАЦИЙАЛАРДЫ ҚАРЖЫ МОНИТОРИНГІ СУБЙЕКТІЛЕРІ WОЛАР АНЫҚТАЛҒАН СӘТТЕН БАСТАП КОМИТЕТКЕ ТАПСЫРАДЫ

1.4. Transcription based on the orthoepic dictionary

This system of phonetic transcription was based on the rules described in Chapter "Orthoepic rules or word sounding principals" of the Kazakh "Orthoepical dictionary" (Aitbaiuly et al., 2007, p. 770).

Unfortunately, it was not possible to formalize and apply in our system a complete list of all the rules described in the dictionary, since some of these rules take into account the origin of words. That is, orthoepic rules may differ depending on whether they are applied to original Kazakh words or words adopted from other languages (Russian, Arabic, Persian, etc.).

However, we tried to formalize and apply all the other rules described in the dictionary, those being the same (or having a priority use case) for all words, both native Kazakh and loanwords.

The phoneme list contains 39 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ж, З, И, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, Ұ, Ү, Ф, Х, Һ, Ц, Ч, Ш, Щ, Ы, І, Э, Ю, Я. This list coincides with the letters of the actual Kazakh Cyrillic alphabet, excluding symbols Ё, Ъ, Ь.

The transcription rule list contains 252 rules.

Sample transcribed words:

ӨКІНУГЕ → УӨКҮНҮГЕ

ОКТЯБРЬСК → УОКТЯБІРСК

ШАЙБАНИ → ШӘЙБАНЫЙ

Sample transcribed sentence:

УОСҰНДАЙ КҮДҮКТІ УОПЕРАЦЫЙАЛАРДЫ ҚАРЖЫ МОНЫЙТОРІЙНГІ СУБЙЕКТІЛЕРІ УОЛАР АНЫҚТАЛҒАН СӘТТЕН БАСТАП КОМІЙТЕТКЕ ТАПСЫРАДЫ

1.5. Combined transcription rules

This system of phonetic transcription combines the two previous ones: at first we apply the rules according to the orthoepical dictionary (system 4), then we apply the rules of the basic transcription (system 3). This allows us, in addition to the Kazakh orthoepic rules, to take into consideration diphthongs (in those cases not considered in the orthoepical dictionary) and to reduce the number of phonemes to 36.

The phoneme list contains 36 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ж, З, И, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, W, Ұ, Ү, Ф, Х, Ц, Ч, Ш, Ы, І, Э.

The transcription rule list contains 380 rules.

Sample transcribed words:

ОКТЯБРЬСК → WОКТӘБІРСК

АҢҚИЫП → АҢҒЫЙЫП

ЖАПЫРАҚ → ЖАПЫРАҚ, ЖАПРАҚ

Sample transcribed sentence:

WОСҰНДАЙ КҮДҮКТІ WОПЕРАЦЫЙАЛАРДЫ ҚАРЖЫ МОНЫЙТОРІЙНГІ СУБЙЕКТІЛЕРІ WОЛАР АНЫҚТАЛҒАН СӘТТЕН БАСТАП КОМІЙТЕТКЕ ТАПСЫРАДЫ

1.6. Transcription proposed by Prof. A. Sharipbay

This system of phonetic transcription was based on the rules proposed and described in the monograph of Professor Altynbek Sharipbay (Sharipbay, 2017).

The automatic online transcriptor "Transcription of Kazakh Cyrillic writing into Latin" (Online Transcriptor, 2017) was used to generate phonetic transcriptions on the basis of these rules.

The phoneme list contains 31 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ж, З, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, Ұ, Ү, Ф, Х, Ш, Ы, І.

The number of transcription rules is unknown.

Sample transcribed words:

ЦИРКУЛЯЦИЯСЫ → СІРКҮЛӘТСЫЙӘСЫ

ДУЭТ → ДҮЕТ

КОНЪЮНКТУРА → КОНЫЙҮНКТҰРА

Sample transcribed sentence:

ОСЫНДАЙ КҮДІКТІ ОПЕРАТСЫЙӘЛАРДЫ ҚАРЖЫ МОНІТОРІҢГІ СҰБЫЙЕКТІЛЕРІ ОЛАР АНЫҚТАЛҒАН СӘТТЕН БАСТАП КОМІТЕТКЕ ТАПСЫРАДЫ

3. The word-based experiment

We carried out a series of experiments on acoustic modeling and speech recognition over the described speech corpus (see Section 1) using the described phonetic transcription systems (see Section 2). The experiments were carried out on the Kaldi platform, using a cluster for distributed computing and a batch-queuing system.

At the first stage, word-based experiments were conducted, where the recognizable speech units were the words of the Kazakh language. Each uttered sentence was interpreted as a sequence of words separated by spaces.

The lexicon and the trigram language model (Federico, Bertoldi, Cettolo, 2008; The IRSTLM Toolkit, 2017) had been constructed from texts of the training set only, so the OOV rate was high (about 20%, see Table 1).

The experiments were carried out according to the standard Kaldi ASR algorithm (Povey et al., 2011; Yessenbayev, Karabalayeva, 2016), which was a sequential training of a monophone model, then several triphone models, and finally, an SGMM model.

To estimate the performance of a recognition system we use the value of WER (word error rate), which is computed as the ratio of erroneously recognized words to the total number of recognized words. WER is a common metric of the performance of experimental speech recognition models. The lower WER is, the better system performance is.

The results of the word-based experiment are shown in Table 2. The best results were obtained for the systems without transcription (systems 1, 2) and with the basic transcription (system 3). Here it can be noted that changing the phonetic transcription system does not lead to significant changes in the system performance on the test set (dispersion is about 0.5% on an absolute scale, about 1.1% relative to the lowest WER). As mentioned above, this is due to the fact that the language model alleviates and overcomes the errors that arise in the acoustic model of the system. Also, we can observe a big difference between the WER values on the training set and the test set. It is due to the high OOV rate in the test set, which again confirms the essential dependence of speech recognition systems on the lexicon and the language model.

Table 2. **The results of the word-based experiment, WER**

N	Phonetic transcription system	Train, %	Dev, %	Test, %
1	Grapheme = phoneme	0.50	40.28	36.88
2	Grapheme = phoneme, without Ё, Ъ, Ь	0.52	40.19	36.79
3	Basic transcription	0.54	40.21	36.93
4	Transcription based on the orthoepic dictionary	0.56	40.60	37.06
5	Combined transcription rules	0.59	40.71	37.21
6	Transcription proposed by Prof. A. Sharipbay	0.57	40.93	37.04

4. The phone-based experiment

At the next stage, independent phone-based experiments were conducted, where the recognizable speech units were phonemes. Each uttered sentence was interpreted as a sequence of phonemes separated by spaces.

As previously, the lexicon and the language model had been constructed from texts of the training set only. The experiments were carried out according to the standard Kaldi ASR algorithm.

The results of the phone-based experiment are shown in Table 3. As previously, the best results are obtained for the first three systems, with the leader being system 1. You can see that here, unlike the previous experiment, the WER values on the test and training sets are comparable, i.e. the dependence on the lexicon and the language model of the system is minimized. This factor makes the results of this experiment much more significant than the results of the previous word-based experiment (see Section 3). With regard to phonetic transcription systems, the error dispersion in the test set is notably larger than in the word-based experiment (about 2.9% on an absolute scale, about 17% relative to the lowest WER). The order of the error is also comparable with similar systems for English (Lopes, Perdigão, 2011; Graves, Mohamed, Hinton, 2013).

Table 3. **The results of the phone-based experiment, WER**

N	Phonetic transcription system	Train, %	Dev, %	Test, %
1	Grapheme = phoneme	12.15	18.46	16.78
2	Grapheme = phoneme, without Ё, Ъ, Ь	13.40	20.05	18.00
3	Basic transcription	13.35	19.48	17.94
4	Transcription based on the orthoepic dictionary	13.77	19.98	18.33
5	Combined transcription rules	14.67	21.63	19.43
6	Transcription proposed by Prof. A. Sharipbay	14.93	21.42	19.66

5. The test manual annotation experiment

Based on the experimental results presented in the previous section, we can already draw certain conclusions about the performance of the described systems of Kazakh phonetic transcription. However,

it is not entirely correct to compare these systems directly with each other, because they use different sets of phonemes, and therefore their test sets differ.

In order to compare the performance of the described phonetic transcription systems, we have chosen *a minimal phonetic alphabet* based on the phonetic system proposed by Prof. A. Sharipbay (Sharipbay, 2017, p. 104-105), since this alphabet contains the smallest number of phonemes among all considered phonetic systems.

For phonetic systems with a greater number of phonemes than in the minimal phonetic alphabet, we have defined the following rules of substitution in order to get rid of redundant phonemes in recognized texts:

```
Ц → ТС
Ч → ТШ
Э → Е
Һ → Х
Щ → Ш
Ю → Y
Я → Ә
И → І
Ё → Е
Ъ →             (remove)
Ь →             (remove)
```

In addition, in the transcription systems (2) and (4) that contain the symbol W denoting the short consonant У, we have applied two more rules of substitution, in the order indicated:

```
У → Ұ
W → У
```

Using this minimal phonetic alphabet, we have annotated a small phonetically representative speech corpus *test_manual*, as an independent test set. This corpus contains 83 Kazakh sentences, uttered by three speakers (two males, one female). These 83 sentences consist of 534 words and 3220 letters. All the letters of the Kazakh Cyrillic alphabet are found in the corpus at least 11 times, except for the letters Ё (0 times), Щ (8 times), Ъ (6 times). All sentences of the corpus were transcribed manually, by careful listening to the uttered sentences and writing down their phonetic transcription.

The phoneme list contains 31 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ж, З, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, Ұ, Y, Ф, Х, Ш, Ы, І.

In this list, the symbol У is used only to denote the short consonant sound as part of diphthongs (just like the W symbol in the basic transcription system (2), see Section 2). Whereas the vowel У, which makes a syllable and can be found in loanwords, is transcribed either by the symbol Ұ or the symbol Y – depending on the hardness or softness of neighboring phonemes.

The symbol И is excluded from the system, since И is considered as an allophone of the phoneme I.

Besides that, we have excluded the symbol Һ (transcribed as X); the symbol Ц (transcribed as TC or C); the symbol Ч (transcribed as TШ); the symbol Щ (transcribed as Ш); the symbol Э (transcribed as E); the symbol Ю (transcribed as ЙҰ, or ЙY, or Y); the symbol Я (transcribed as ЙА, or ЙӘ, or Ә). The symbol Ё does not occur in the set.

Sample source sentences:

ОРЫСТЫҢ ТАМАША САТИРИК ЖАЗУШЫ МИХАИЛ МИХАЙЛОВИЧ ЗОЩЕНКО

СУББВЕНЦИЯ КӨЛЕМІ ЕКІ МИЛЛИАРД ТЕҢГЕ

РАДИЩЕВ ПЕТЕРБУРГТЕН МӘСКЕУГЕ САЯХАТ

Sample transcribed sentences:

УОРЫСТЫҢ ТАМАША САТІРІГ ЖАЗҰШЫ МІХАЙЫЛ МІХАЙЛАВІШ ЗОШЕНКА

СҰБВЕНСІЙӘ КӨЛЕМІ ЙЕКІ МІЛІАРТ ТЕҢГЕ

РАДЫЙШЕФ ПЕТӨРБУРКТЕН МӘСКЕУГЕ САЙАХАТ

The results of the test manual annotation experiment are shown in Table 4. The best result was shown by the system with basic transcription (3), and the second best is the system without transcription (1). At the same time, the inferiority of the system based on the orthoepic dictionary (4) and the system without Ё, Ъ, Ь (2) compared with the system 1 is not essential.

Table 4. **The results of the test manual annotation experiment, WER**

N	Phonetic transcription system	Test, %
1	Grapheme = phoneme	23.38
2	Grapheme = phoneme, without Ё, Ъ, Ь	23.68
3	Basic transcription	23.05
4	Transcription based on the orthoepic dictionary	23.48
5	Combined transcription rules	24.50
6	Transcription proposed by Prof. A. Sharipbay	25.03

6. Conclusions

The paper presents the results of the experiments on Kazakh speech recognition using different phoneme sets and alternative phonetic transcriptions. Based on the results obtained, it can be concluded that a system with basic transcription, which takes diphthongs into account and excludes rare loan phonemes, provides the most adequate model of real Kazakh speech. On the contrary, systems using transcriptions based on more complex orthoepic rules showed an inferior recognition performance, which, presumably, was due to some inconsistency of these rules with the actual sounding of Kazakh speech. Indeed, during the time that the actual Kazakh Cyrillic alphabet and the phonetic system on its basis are being used, native speakers have been able to adapt to foreign loan phonemes. Now they are able to utter them as required (e.g. in Russian), in an almost authentic manner. This conclusion is also supported by quite good recognition results obtained by the systems without using any orthoepic rules (in the "Grapheme = phoneme" experiments).

It can also be observed that reducing the number of phonemes to 36 (in the basic transcription) is absolutely harmless and even leads to an increase in the recognition performance. This seems to be due to that phonemes excluded from the phonetic system can be legitimately considered either as diphthongs or as allophones of other phonemes, more frequent in speech. At the same time, in the process of automatic recognition, we do not observe any negative confusion effects that could appear in consequence of classifying such phonemes to the wrong classes. On the contrary, such negative effects get reduced, and the recognition performance gets somewhat increased. This testifies to the correctness of the substitution rules that we use to remove the symbols Ё, Ю, Я, Һ, Щ, Ъ, Ь.

Further reducing the number of phonemes to 31 (in the minimal phonetic alphabet) leads to a decrease in the recognition performance. However, this can be explained by the complexity of the system of orthoepic rules, rather than by shortening the phoneme list. Let us also note that a certain portion of errors in the last experiment (with the manually annotated corpus) may result from the rough mapping of phonetic systems with a larger number of phonemes into the system with the minimal phonetic alphabet.

This work is a preliminary study. A more detailed study will require extensive analysis of the contribution of each single phoneme into the performance of continuous speech recognition. Only on the basis of such an analysis we shall be able to elaborate and introduce an adequate phonetic system with appropriate transcription rules. That is going to become the subject of our future work.

Acknowledgement

This work has been conducted under the targeted program O.0743 (0115PK02473) of the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan.

REFERENCES

1. Lopes, C., Perdigão F. (2011). Phoneme recognition on the TIMIT database. Speech Technologies, 2011, pp. 285–302.

2. Graves, A., Mohamed, A.r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, 2013, pp. 6645–6649.

3. Schwarz, P. (2008). Phoneme recognition based on long temporal context. PhD thesis, Brno University of Technology, Faculty of Information Technology, 2008.

4. Matejka, P. (2009). Phonotactic and Acoustic Language Recognition. PhD thesis, Brno University of Technology, Faculty of Electrical Engineering and Communication, 2009.

5. Furui, S. (2005). 50 Years of Progress in Speech and Speaker Recognition Research. Proceedings of ECTI Transactions on Computer and Information Technology, vol. 1, no. 2, 2005.

6. Sharipbay, A. (2017). Kazakh zhazuin latyn alipbiyine auistyru maseleleri. Monograph. Astana: L.N. Gumilyov Eurasian National University, 2017. – 136 p. – ISBN: 978-601-301-989-5.

7. Aitbaiuly, O., et al. (2007). Orthoepical dictionary. Ed. Aitbaiuly O., et al. – Almaty: Arys, 2007. – 800 p. – ISBN: 9965-17-473-3.

8. Povey, Daniel and Ghoshal, Arnab and Boulianne, Gilles and Burget, Lukas and Glembek, Ondrej and Goel, Nagendra and Hannemann, Mirko and Motlicek, Petr and Qian, Yanmin and Schwarz, Petr and Silovsky, Jan and Stemmer, Georg and Vesely, Karel. (2011). The Kaldi Speech Recognition Toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, 2011.

9. Univa Grid Engine, official site. (2017). URL: www.univa.com [Access date: 15.07.2017].

10. Makhambetov, O., Makazhanov, A., Yessenbayev, Zh., Matkarimov, B., Sabyrgaliyev, I., and Sharafudinov, A. (2013). Assembling the Kazakh Language Corpus. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA, October. Association for Computational Linguistics, pp. 1022–1031.

11. Technical Report (final). (2014). Sozdaniye akusticheskogo korpusa kazahskogo yazyka i utochneniye ego foneticheskogo stroya, predstavleniye kazahskih fonem v mezhdunarodnom foneticheskom alfavite. L.N. Gumilyov Eurasian National University; PI Sharipbay A.A.; exec.: Yessenbayev Zh.A. [et al.]. – Astana, 2014. – 68 p. – Reg No. 0112PK02344.

12. Regular expression operations, The Python Standard Library. (2017). URL: https://docs.python.org/2/library/re.html [Access date: 15.07.2017].

13. ‘sed’ manual. – GNU. (2017). URL: https://www.gnu.org/software/sed/manual/sed.txt [Access date: 15.07.2017].

14. Online Transcriptor. (2017). Transcription of Kazakh Cyrillic writing into Latin. URL: http://e-zerde.kz/latin/index22.php [Access date: 15.07.2017].

15. Yessenbayev, Zh., Karabalayeva, M. (2016) A baseline system for Kazakh broadcast news transcription. In the Proc. of the V International Scientific-Practical Conference on Informatization of Society. Astana, Kazakhstan, 2016, pp. 48–50.

16. Federico, Marcello and Bertoldi, Nicola and Cettolo, Mauro. (2008). IRSTLM: an open source toolkit for handling large scale language models. In the Proc. of Interspeech 2008, pp. 1618–1621.

17. The IRST Language Modeling (IRSTLM) Toolkit. (2017). URL: http://hlt-mt.fbk.eu/technologies/irstlm [Access date: 15.07.2017].

УДК 004.522

DEVELOPMENT OF A KYRGYZ LANGUAGE SPEECH SYNTHESIZER: A DEMONSTRATION OF THE OSSIAN FRONTEND AND THE MERLIN NEURAL NETWORK SPEECH SYNTHESIS SYSTEM

J. Meyer
Department of Linguistics University of Arizona Tucson, Arizona, 85721
joshua.richard.meyer@gmail.com

This paper describes the rapid and efficient development of a Kyrgyz language speech synthesizer using open source technologies created by The Centre for Speech Technology Research at the University of Edinburgh. The synthesizer was trained on recordings from an open license audio book recorded by the Kyrgyz Language Audiobooks initiative from the Bizdin Muras group.

Additionally, I give specific recommendations for using lightly-supervised acoustic modelling with the Kaldi ASR toolkit to increase the labeled dataset used for training a new voice.

I. Introduction

The creation of new Text-To-Speech (TTS) technologies typically require either (1) expert, linguistic knowledge of the language or (2) a sizeable collection of audio with transcriptions. Open source software exists for either of these two approaches, for example eSpeak NG allows linguists with some programming knowledge to create a synthesizer by coding orthography → pronunciation rules.[1]

Practically speaking, the linguist need not be an expert programmer to create a useable voice quickly with eSpeak NG. However, a major drawback of such rule-based speech synthesizers is that they inevitably sound robotic. For certain applications this is not an issue, and even beneficial (screen readers like NVDA can more easily speed up this kind of speech).[2] However, most users prefer human-like voices.

The second approach allows the researcher / developer to create a natural, human-sounding voice by training a statistical model to generate an approximation of the provided voice. While the end product of this approach sounds best, it requires access to some collection of speech recordings (ie. speech corpus) as well as more expert programming knowledge. In this paper I demonstrate that we can partially avoid

[1] eSpeak NG GitHub repo: https://github.com/espeak-ng/espeak-ng
[2] NVDA GitHub repo: https://github.com/nvaccess/nvda

the corpus collection problem by generating transcripts. Additionally, I show that by using certain user-friendly open source toolkits, we can effectively circumvent the need of much of the traditional, technical expertise needed in creating TTS (ie. signal processing and finite state machine technology). These toolkits hide much of the technical problems *under the hood*, if the developer wishes, she can easily find the source code which she would like to modify.

The current paper demonstrates the development of a Kyrgyz language synthesizer using the open-source toolkits Ossian and Merlin [1] which are designed to be userfriendly and allow the non-programming-expert to create a good voice quickly.[1][2]

Furthermore, I propose a method to create accurate alignments at the utterance level for an entire audiobook, by training a speech recognizer on only a small subsection of that same book. This is an idea from semi-supervised, or lightly-supervised learning [2]. Typically, audiobooks come in audio files which are too large to use directly for training a synthesizer. By splitting large files on silence and then generating transcriptions, we can use more audio and spend less time creating hand-aligned segments.

II. Pipeline

The project can be condensed into these main steps:

1) find audiobook (human)
2) split audiobook on silence (sox)
3) hand-align sample of audiobook (human)
4) train speech recognizer on sample of audiobook (Kaldi)
5) use trained recognizer to generate transcripts for more audiobook (Kaldi)
6) use text from audiobook to train TTS frontend (Ossian)
7) train acoustic and duration model for DNN TTS (Merlin)
8) synthesize new speech (Merlin + Ossian)

III. Data

The audio book used to train this system was spoken by a female native Kyrgyz speaker from the north of the country and recorded by the Bizdin Muras Group.[5] The audio files made available to the author

[1] Ossian GitHub repo: https://github.com/CSTR-Edinburgh/Ossian

[2] Merlin GitHub repo: https://github.com/CSTR-Edinburgh/merlin [5]Bizdin Muras Group website: http://bizdin.kg/

were formatted as stereo MPEG 1.0 Layer III with a sampling rate of 44 kHz and a 160 kbps bit rate. The recordings were available in 30 minute segments.

The original recordings were then split on silence longer than .55 seconds, and the resulting utterance-sized files were then hand-aligned to text by the author. By using .55 seconds of silence as the delimiter, the audio was split on what mostly corresponded to complete sentences in the text file. After splitting the audio on silence, the first 1.5 hours were hand-aligned to text by the author. The following describes the proposed lightly supervised method.

IV. Speech recognizer: kaldi

A. Training

All speech recognition training and decoding may be performed with the Kaldi toolkit. [3][1].

The original source should be first down-sampled to 16kHz and the two original channels should be merged (to convert stereo to mono). The features should be 13-dimensional MFCCs. Delta and Delta Delta features should also be extracted, for a total of 13 MFCCs + 13 Δ + 13 $\Delta\Delta$ = 39 features. Cepstral mean and variance normalization then may be applied.

Initially, monophones are trained from a flat start. Then, context dependent triphones are trained using the alignments given by the monophones. The context-dependent triphones are then used to generate alignments for the audio, and those alignments are used to train the final DNN. Since in TTS we're working typically with a single speaker, there should be no speaker adaptation applied.

The language model incorporated into the decoding graph can be a tri-gram backoff model trained on the text of the audiobook itself. This means the decoder will be highly biased to predict strings from the audiobook, which is good for the current application.

B. Decoding

Additional N hours of the audio book (which has been split on silence) can then be decoded with the DNN ASR system trained on the first, hand-aligned section of the audiobook. The generated transcrip-

[1] Kaldi GitHub repo:https://github.com/kaldi-asr/kaldi

tions for these N hours can be treated as correct and used in the training of the speech synthesizer. Hypothetically as many hours are possible can be decoded and used in this same way.

V. TTS frontend: Ossian

Training in Ossian was performed with the standard naive_01_nn recipe.

The main function of Ossian is to convert raw text into linguistic features which are believed to correlate to phonetic information (including prosody). In the naive_01_nn case, we are extracting features in a very general way, with the main processing being standard tokenizing by splitting text on whitespace and punctuation. Each token is automatically classified as word, space, or punctuation by using information from the UTF-8 character set. This is performed by a Python module called RegexTokeniser.

In this naive recipe, we use Ossian to train word vectors as a correlate to POS tagging. These vectors have a context size of 250 tokens, and are appended to their tokens during training. This is performed by a Python module called VSMTagger.

The next step in naive frontend processing is to break a word down into its component phonemes. Ossian helps us do this for a new language where we don't have a letter-to-sound (LTS) rule set. We can tell Ossian to just break a word down into its letters, and use the letters themselves as a stand-in for phonemes. This is performed by a Python module called NaivePhonetiser.

Ossian will also make a prediction about where pauses occur. As a first informed guess in training, Ossian will first hypothesize that pauses occur at punctuation marks in text. This initial punctuation-pause prediction is used to bootstrap a more refined model later on in training. Once we have a more refined model, we use its prediction to train a Python module called SKLDecisionTreePausePredictor.

Lastly, (linguistic) phrase breaks are predicted to be at silence markers. This is performed by a Python module called PhraseMaker.

Ossian represents each utterance internally in a hierarchical structure (saved as an XML file for convenience). This hierarchical structure is able to efficiently save information relating to phrases, words, and phones (and hypothetically any other relationship we can think of).

Next, we take these hierarchical structures and flatten them into strings. The manner of encoding these trees as character strings is done in such a way as to retain hierarchical as well as neighboring information. For example, phrase information is appended to the phone as well as the two phones to the left and two phones to the right of the central phone in question. This allows us to model influence of co-articulation as well as higher-level influence.

The level of detail in these representations is up to the researcher, and can incorporate sophisticated linguistic knowledge. For our purposes in this paper, we want to show that creating this Kyrgyz synthesizer without any explicit linguistic knowledge can lead to already good results.

With the help of Ossian, we are able to quickly take raw text and generate potentially linguistic features, without any pronunciation dictionary. This is a major advantage of Ossian in particular when working with low-resource languages. In particular, if the writing system of the language is fairly transparent (that is, the letters correspond strongly to different sounds) we can create a fairly high-quality linguistic representation in little time. o

VI. TTS core regression: Merlin

Merlin takes as training data input (1) the linguistic text features generated by Ossian as well as (2) the audio features from training utterances. The audio features will have been extracted and engineered via the WORLD Vocoder, but we will not go into depth here on how that is performed. (WORLD [4], D4C edition [5])[1] Suffice it to say that the raw audio contains too much information in addition to noise. We need to compress the signal to retain as much useful information as possible and lose as much noise as possible.

Neural nets can only work with numerical information. That is, they take vectors of real numbers as input, and they return vectors of real numbers as output. However, the output provided by Ossian (and other such frontends) is a character string which contains hierarchical and contextual information about the phone. So we need to first convert this string into a real vector. This is achieved with a combination of one-hot encoding and real-number values.

[1] WORLD GitHub repohttps://github.com/mmorise/World

TTS can be naturally viewed as a sequence-to-sequence problem, where we are trying to map a sequence of phonemes (derived from input characters) onto a sequence of acoustic features. If we have one item in our sequence per phoneme, and one frame of audio features per item in our sequence of output audio, we will have an obvious mismatch in lengths of our text input and audio output. The text input will be much shorter, which makes sense because each phoneme will map onto multiple frames of audio. In training, this is not a major problem, because we can use the same forced-alignment techniques well known in standard HMM speech recognition. However, when we need to synthesize speech we need to decide how many audio frames are generated by each input phoneme. In other words, we need to predict the duration of a given phone, and we accomplish that in Merlin with a duration predictor. The duration predictor is a neural net which takes as input the same linguistic features derived by Ossian for a given phone, and predicts its length in frames.

Merlin in fact will train two separate DNNs which are used in speech synthesis. The first model is the acoustic predictor, and is trained to predict the acoustic quality of sounds. That is, this DNN predicts the audio features from text. The second DNN model is a duration predictor, and predicts from the text how long sounds will last on the output.

For the current project, the 1.5 hours of hand-labeled data were used to train both DNNs in Merlin. The proposed lightly supervised method is a recommendation for future work.

VII. Conclusion

This paper demonstrates how minimal resources can be leveraged in addition to open-source speech technology toolkits to successfully train a synthetic voice for a new language. At this point I would like to continue this line of work by:

1) implement the proposed lightly-supervised technique to augment the training data

2) evaluating the final voice with ratings from Kyrgyz speakers

3) experiment with more sophisticated morpheme vectors in lieu of POS tagging

4) experiment with speech recognizer to find minimum requirements for producing good alignments (larger ngram LM and less hand-alignment)

Acknowledgment

I would like to thank the researchers at CSTR for welcoming me to Edinburgh and helping me understand their fabulous toolkits; in particular Oliver Watts, Simon King, and Srikanth Ronanki.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (DGE-1746060). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Wu, Zhizheng, Oliver Watts, and Simon King: Merlin: An open source neural network speech synthesis system. Proc. SSW, Sunnyvale, USA (2016).

[2] Lamel L., Gauvain J. L., Adda G.: Investigating lightly supervised acoustic model training. Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on. IEEE, 2001. . 1. . 477–480.

[3] Povey, Daniel, et al: The Kaldi speech recognition toolkit." IEEE 2011 workshop on automatic speech recognition and understanding. No. EPFLCONF-192584. IEEE Signal Processing Society, 2011.

[4] M. Morise, F. Yokomori, and K. Ozawa: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 18771884, 2016.

[5] M. Morise: D4C, a band-aperiodicity estimator for high-quality speech synthesis. Speech Communication, vol. 84, pp. 57–65, Nov. 2016.

УДК 004.934

APPLICATION OF THE FAST CONTINUOUS WAVELET TRANSFORM FOR SENTENCE RECOGNITION

V. Semenov, A. Shurbin
Federal State Educational Budget Institution of Higher Professional Education «The Ulianov Chuvash State University»
syundyukovo@yandex.ru, shurti@mail.ru

The fast Fourier transform is used to calculate the continuous wavelet transform. The MHAT wavelet constructed on the basis of the second derivative of the Gaussian function is used. The wavelet spectrum is divided into segments of fixed duration. To determine the boundaries between the vowels and consonant phonemes, the energy of the segments of the wavelet spectrum of the sentence is calculated. The energy of the segments is calculated using the Fourier spectrum of the wavelet spectrum of the sentence. The analysis results show that the energy of segments of vowel phonemes is allocated in the form of maximum peaks, and the energy of consonants is always lower than the vowel energy. The boundary between the vowels and consonant phonemes is determined to within 2-3 segments. For each phoneme standards of phonemes are formed based on the wavelet spectrum of phonemes. These standards are used to recognize sentences. Since the various sentences have a certain number of vowel letters, they can be preliminarily distinguished by the number of peaks in the wavelet spectrum.

Keywords: continuous wavelet transform, fast Fourier transform, speech recognition, sentence.

ПРИМЕНЕНИЕ БЫСТРОГО НЕПРЕРЫВНОГО ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ ДЛЯ РАСПОЗНАВАНИЯ ПРЕДЛОЖЕНИЙ

В. И. Семенов, А. К. Шурбин
ФГБОУ ВПО «Чувашский государственный университет им. И. Н. Ульянова», Чебоксары
syundyukovo@yandex.ru, shurti@mail.ru

Для вычисления непрерывного вейвлет-преобразования применяется быстрое преобразование Фурье . Используется *MHAT-вейвлет*, который конструируется на основе второй производной функции Гаусса. Вейвлет-спектр разбиваются на сегменты фиксированной длительности. Для определения границ между гласными и согласными фонемами вычисляется энергия сегментов вейвлет-спектра предложения. Энергия сегментов вы-

числяются с применением Фурье спектра вейвлет-спектра предложения. Результаты анализа показывают, что энергия сегментов гласных фонем выделяется в виде максимальных пиков, а энергия согласных букв всегда ниже, чем энергия гласных. Граница между гласными и согласными фонемам определяется с точностью до 2–3 сегментов. Для каждой фонемы формируются эталоны фонем на основе вейвлет-спектра фонем. Эти эталоны используются для распознавания предложения. Так как различные предложения имеют определенное количество гласных букв, то их можно предварительно различать по количеству пиков вейвлет-спектра

Ключевые слова: непрерывное вейвлет-преобразование, быстрое Фурье преобразование, распознавание речи, предложение.

При произношении одного и того же предложения несколько раз, длительность предложения в каждом случае будет разной, так как длительность, громкость и темп речи изменяются в широких пределах. При этом различные звуки речи растягиваются или сжимаются не одинаково. Например, гласные изменяются значительно сильнее, чем согласные, при увеличении длительности произношения слова. Одной из основных трудностей при распознавании является неопределенная временная организация речевого сигнала. Из непрерывного речевого потока довольно непросто выделить какие-либо речевые единицы, то есть сегментация речи является одной из основных задач системы распознавания слитной речи. Для того, чтобы сравнить с эталоном, надо путем деформации оси времени совместить участки, соответствующие одним и тем же звукам. Для нелинейного согласования речи используются методы динамического программирования (алгоритм динамического искажения времени) и марковского моделирования. В настоящей работе, как для сегментации речи, так и для формирования эталонов фонем используется непрерывное быстрое вейвлет-преобразование.

Для вычисления вейвлет-спектра речевого сигнала используется формула непрерывного вейвлет-преобразования

$$W(a,b)=\frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} S(t)\ \psi\left(\frac{t-b}{a}\right)\ dt.$$

Вейвлет–спектр $W(a,b)$ (масштабно-временной спектр) в отличие от Фурье-спектра является функцией двух аргументов. Первый аргумент a (временной масштаб) аналогичен перио-

ду осцилляций, второй аргумент b – смещению сигнала по оси времени. Малые значения a соответствуют высоким частотам, большие значения a – низким частотам. Непрерывное вейвлет-преобразование выполняется, как правило, прямым численным интегрированием [1], что для больших временных последовательностей занимает большое время. Для вычисления непрерывного вейвлет-преобразования нами разработан алгоритм, который позволяет в 15000 раз увеличить быстродействие для 32768 отсчетов, по сравнению прямым численным интегрированием. Такое увеличение быстродействия достигнуто за счет применения быстрого преобразования Фурье (БПФ). [2,3,4,5,6]. В качестве «материнской» функции нами выбрана симметричная, четная, гладкая функция $MHAT(t)$, называемая «мексиканская шляпа», которая конструируется на основе второй производной функции Гаусса.

Для того, чтобы совместить участки, соответствующие одним и тем же звукам определяются границы между гласными и согласными фонемами. Вейвлет-спектр разбиваются на сегменты фиксированной длительности (n = 128). Общее число сегментов 256. Для определения границ между гласными и согласными фонемами вычисляется энергия сегментов функций $W(1,b)$, $W(2,b)$, $W(20,b)$ и исследуемого слова $S(t)$. В каждом сегменте вычисляются коэффициенты Фурье:

$$d(n)=\frac{1}{M}\sum_{k=0}^{M-1} W(a,k)\cos(\frac{2\pi nk}{M}),$$

$$e(n)=\frac{1}{M}\sum_{k=0}^{M-1} W(a,k)\sin(\frac{2\pi nk}{M}).$$

По формуле:

$$F(i)=d^2(i)+e^2(i)$$

вычисляется Фурье-спектр функций $W(1,b)$, $W(2,b)$, $W(20,b)$ и $S(t)$. Энергия сегментов вычисляются по формуле:

$$E=\sum_{i=1}^{n} F(i).$$

Обозначим энергию сегментов вейвлет-преобразования (ВП) $W(1,b)$, $W(2, b)$ и исследуемого слова $S(t)$ функциями $E1(n)$, $E2(n)$ и

$E3(n)$ соответственно, где n меняется от 1 до 256. Результаты анализа показывают, что энергия сегментов гласных букв в $W(1,b)$, $W(2,b)$ выделяется в виде максимальных пиков, а энергия согласных букв всегда ниже, чем энергия гласных. Энергия сегментов шипящих букв в $E1(n)$ выделяется в виде максимальных пиков, в $E2(n)$ и $E3(n)$ – в виде минимумов. Чтобы определить местоположение фонем в предложении, вычисляется ВП функцией $E1(n)$, $E2(n)$ и $E3(n)$ с масштабным коэффициентом $a = 4$. Для нахождения местоположения гласных букв нормируются энергии $E2(n)$, $E3(n)$, находится их сумма и выполняется ВП $W4(4,b)$. Таким образом, если слово содержит одну гласную букву, то выделяется один положительный максимум, если две гласные буквы – два положительных максимума, и т.д. Каждое слово имеет определенную структуру. Граница между гласными и согласными буквами или между гласными и шипящими определяется с точностью до 2–3 сегментов. При произношении предложения число гласных увеличивается пропорционально числу слов, то есть максимумов и минимумов будет больше [4]. Дополнительную информацию о расположения гласных и согласных звуков речи в предложении можно извлечь, исследуя зависимость энергии сегментов вейвлет-спектра от масштабного коэффициента *a.*Масштабный коэффициент a меняется от 1 до 50 с шагом 1. 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.

Для формирования эталонов фонем применяется следующий алгоритм. Вычисляются вейвлет-коэффициенты $W(1,b)$, $W(2,b)$, $W(3,b)$, $W(22,b)$ и $W(50,b)$ предложений, где b меняется от 1 до 32768. Полученные вейвлет-коэффициенты (функции) $W(1,b)$, $W(2,b)$, $W(3,b)$, $W(22,b)$ и $W(50,b)$ разбиваются на сегменты фиксированной длительности ($n = 128$), что соответствует 16 мс при частоте дискретизации 8000 гц. Мерой сходства (различия) явля-

ется евклидово расстояние между эталонными признаками фонем в предложений и признаками сегментов произнесенного в данный момент речевого сигнала:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}.$$

Для каждого предложения на этапе обучения в базу данных записываются несколько вариантов одного и того же предложения с соответствующими признаками для которых евклидово расстояние минимальное. При каждом произношении предыдущие эталонные признаки складываются с новыми признаками и делятся на два. Таким образом, для каждого предложения и его варианта путем многократного произношения формируется усредненные эталонные признаки. При распознавании вычисленные значения расстояний для каждого предложения сравниваются между собой. Предложение, у которого евклидово расстояние минимальное, выводится на экран монитора. Так как различные предложения имеют определенное количество гласных букв, то их можно предварительно различать по количеству пиков вейвлет-спектра $W2(4,b)$ и сравнивать предложения с одинаковым количеством пиков.

ЛИТЕРАТУРА

1. Семенов В.И., Желтов П.В. Патент на изобретение № 2403628 РФ, МПК G10L 15/10. Способ распознавания ключевых слов в слитной речи. Опубл. 10.11.2010 Бюл. №31.

2. Желтов П.В., Семенов В.И. Распознавание речи на основе вейвлет-преобразования. // Чуваш.гос.ун-т.-Чебоксары, 2008.-16с.-Деп. в ВИНИТИ РАН 29.02.08, №174-В2008.

3. Желтов П.В., Желтов В.П., Семенов. В.И. Математическая модель распознавания слитной речи. //Вестник ЧГУ. – 2012. – № 3. С. 210–212.

4. Желтов П.В., Семенов. В.И., Трофимова А.И., Шурбин А.К. Алгоритмы идентификации фонем и формирования слова в системах распознавания речи на основе вейвлет-преобразования. // Вестник ЧГУ. – 2014. – № 2. С. 98–102.

5. Желтов П.В., Желтов В.П., Семенов. В.И. Применение вейвлет-преобразования при сегментации речи. // – 2015. – № 1. URL: ww.science-education.ru/121-19205.

УДК 004.522

RECENT RESULTS IN SPEECH RECOGNITION FOR THE TATAR LANGUAGE

A. Khusainov[1], A. Khusainova[2]
[1] Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic Kazan, Russia
[2] Kazan Federal University. Kazan, Russia
khusainov.aidar@gmail.com

This paper presents a comparative study of several different systems for speech recognition for the Tatar language, including systems for very large and unlimited vocabularies. The conventional way of building speech recognition systems is to obtain required acoustic models, a pronunciation dictionary, a language model, and use some of the decoders. The situation can be worse whenever you have to recognize the speech of an under-resourced language. In that case some (or all) of the required resources and algorithms may not exist. Therefore, we have to determine an acoustic alphabet, to record and annotate speech corpora, to build models with different existing approaches and to evaluate the recognition quality of combinations of the system's parts.

We present our recent results in creating continuous speech recognition systems for the Tatar language: we describe the word-based approach to create very large vocabulary speech recognition system and sub-word based approach – for the case of unlimited vocabulary recognition. All the compared systems use a corpus-based approach, so recent results in speech and text corpora creation are also shown. The recognition systems differ in acoustic modelling algorithms, basic acoustic units, and language modelling techniques. The DNN-based system with word-based language model shows the best recognition result obtained on the test part of speech corpus.

Keywords: speech recognition; acoustic modelling; language modelling; Tatar language.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ В ОБЛАСТИ АНАЛИЗА СЛИТНОЙ ТАТАРСКОЙ РЕЧИ

А.Ф. Хусаинов[1], А. Хусаинова[2]
[1] Институт прикладной семиотики Академии наук Республики Татарстан, Казань, Россия
[2] Казанский федеральный университет, Казань, Россия
khusainov.aidar@gmail.com

В статье представлено сравнительное исследование нескольких различных систем распознавания речи для татарского языка, включая системы с

очень большим и неограниченным словарем. Для создания систем распознавания речи необходимо построение акустических моделей, словаря транскрипций, языковой модели и использование декодера. В случае малоресурсного языка некоторые (или все) требуемые ресурсы и алгоритмы могут отсутствовать, что делает необходимым решение задач определения акустического алфавита, записи и аннотации речевого корпуса, построения модели на основе различных существующих подходов и оценки качества распознавания.

В данной работе представлены последние результаты создания систем распознавания слитной речи для татарского языка: описана система пословного распознавания с очень большим словарем и послоговая модель для распознавания без ограничения лексикона. Все построенные системы распознавания используют корпусный подход, поэтому приводится статистика по созданным речевых и текстовым корпусам. По результатам проведенных экспериментов система на основе глубоких нейронных сетей с языковой моделью на основе слов показала наилучший результат распознавания

Ключевые слова: распознавание речи, акуситческие модели, языковая модель, татарский язык.

1. Основная часть

1.1. Введение

Для построения системы распознавания речи требуется наличие акустических моделей звуков языка, словаря с фонетическими транскрипциям, языковая модель и реализация алгоритма декодера. Для класса малоресурсных языков задача усложняется отсутствием некоторых (или всех) необходимых материалов и алгоритмов.

В данной статье мы представляем последние результаты, достигнутые в институте прикладной семиотики АН РТ в области распознавания слитной речи на татарском языке: представлены пословная система распознавания речи для очень большого словаря, а также система, использующая части слов в качестве базовых элементов (sub-word based recognition) для распознавания речи с использованием практически неограниченной лексики.

Татарский язык является вторым по количеству носителей в России; на 2016 год число носителей было определено на уровне 4.2 миллионов человек в России и около 5.2 миллионов во всем мире (*Lewis, Simons, Fennig, 2016*). Используемый с 1939 года кириллический алфавит содержит 39 букв. На основе имеющейся

системы классификации языков (*Berment, 2004; Krauwer, 2003*) в 2013 году татарский язык был отнесен к классу малоресурсных языков (*Khusainov, 2014*). Однако последние результаты в областях машинного перевода (*Salimzyanov, Washington, Tyers, 2013; Yandex Translate; Suleymnov, Gatiatullin, Gilmullin, 2012*), автоматического анализа и синтеза речи (*Khusainov, Khusainova, 2016; Khusainov, Suleymanov, 2013*) могут со временем изменить результаты данной классификации.

1.2. Обучающие данные

Размер и качество обучающих данных играет решающую роль в современных системах распознавания речи. Для обучения системы распознавания татарской речи был создан многодикторный корпус звучащей речи. Корпус состоит из читаемой речи носителей языка (преимущественно казанского диалекта). Для построения языковой модели мы использовали татарский национальный корпус «Туган тел» (*Suleymanov, Nevzorova, Khakimov, 2013*).

1.2.1. Речевой корпус

Большинство современных систем анализа речи использует для обучения речевых корпусы длительностью в сотни и тысячи часов. Такой объем данных позволяет добиться построения систем, устойчивых к изменению особенностей произношения дикторов, их пола, возраста, а также к изменению характера внешних шумов.

Запись и аннотирование многодикторного речевого корпуса татарского языка продолжаются, на данный момент корпус состоит из двух основных частей. Первая часть, «Базовая», было направлена на произнесение всех татарских фонем большим количеством дикторов. Исходя из этой цели, каждый диктор читал примерно 2х минутных фрагмент специально подготовленного текста, состоящего из 11 предложений (из литературных источников), 13 отдельных слов и 7 новостных предложений. Каждый из таких текстовых фрагментов был подготовлен таким образом, чтобы содержать все фонемы татарского языка и максимально разнообразный контекст произнесения фонем (контекст – предыдущая и последующая фонемы). «Базовая» часть используется на первых этапах алгоритмов для построения начальных моделей монофонов (см. п.1.3.5).

Вторая, «Основная», часть корпуса состоит из произнесения

случайно выбранных текстовых фрагментов. Запись каждого диктора имеет продолжительность 20-30 минут в зависимости от скорости чтения. Источником предложений служил татарский национальный корпус «Туган тел», в котором были расшифрованы аббревиатуры, числа и даты.

«Базовая» и «Основная» части корпуса были вручную аннотированы. На данный момент корпус состоит из аудиофайлов, соответствующих им текстов, фонетических транскрипций и дополнительных метаданных: пол, возраст, родной язык диктора, экспертная оценка качества произношения по 5-балльной шкале.

Следующими шагами в работе над корпусом планируется доведение его общей продолжительности до 100 часов читаемой речи и начало аннотирования эфирной речи архивов телевизионных и радиопередач.

Основные характеристики речевого корпуса приведены в Табл. 1.

Таблица 1. **Характеристики многодикторного корпуса звучащей татарской речи**

Характеристика	Значение
Количество дикторов	441
Продолжительность	78:51:52
Количество дикторов в «Базовой» части	251
Продолжительность записей в «Базовой» части	8:12:16
Средняя продолжительность записей «Базовой» части	1:58
Количество дикторов в «Основной» части	190
Продолжительность записей в «Основной» части	70:39:00
Средняя продолжительность записей в «Основной» части	0:22:18

1.2.2. Текстовый корпус

Языковые модели были обучены на текстах из татарского национального корпуса. Были разработаны алгоритмы фильтрации, которые устраняли дублирование предложений, переводили текст к нижнему регистру, расшифровывали даты, аббревиатуры и числа.

Основные характеристики текстового корпуса после обработки приведены в Табл. 2.

Таблица 2. **Характеристики текстового корпуса**

Характеристика	Значение
Количество файлов	217 294
Количество слов	69 810 033
Количество слогов	186 014 478 (2,66 / слово)
Количество букв	434 636 548 (6,23 / слово)

1.3. Описание архитектуры систем распознавания речи

1.3.1. Общее описание

Для проведения экспериментов были построены системы распознавания речи, отличающиеся базовыми акустическими единицами, размером обучающих данных, алгоритмами построения моделей и декодирования. Все эксперименты были проведены на базе инструментария Kaldi (*Povey, Ghoshal, Boulianne, et al., 2011*).

Мы использовали два вида акустических единиц: монофоны и трифоны. Записи из «Базовой» части речевого корпуса были использованы на начальном этапе построения акустических моделей.

Базовыми признаками речи были выбраны векторы кепстральных коэффициентов (MFCCs), к которым были добавлены «дельта» и «дельта-дельта» коэффициенты, вместе формирующие вектор размерности 39. В более продвинутых системах (Tri2, Tri3, Tri4, NN) мы дополнительно использовали алгоритмы трансформации признаков LDA/MLLT, а также техники адаптации к дикторам SAT и fMLLR (*Rath, Povey, Vesely, Cernocky, 2013*).

Языковые модели были построены на основе слов и слогов. Послоговые языковые модели использовали подход деления редких слов на слоги для уменьшения числа внесловарных элементов (OOV, out-of-vocabulary). Для обоих подходов были построены полные и усеченные (pruned) 3-граммные и 4-граммные модели.

Общий перечень построенных систем распознавания речи представлен в Табл. 3, где, например, система MonoSW является sub-word версией системы Mono.

Таблица 3. **Характеристики текстового корпуса**

Система	Акустический элемент	Обучающие аудио данные	Признаки	Языковая модель
Mono, MonoSW	монофон	отдельные слова	MFCCs	Усеченная 3-граммная
Tri1, Tri1SW	трифон	отдельные слова	+ delta, delta-delta	+ 3-граммная
Tri2, Tri2SW	трифон	«Базовая» часть	+ LDA / MLLT	Как указано выше
Tri3, Tri3SW	трифон	«Базовая» часть	+ fMLLR	Как указано выше
Tri4, Tri4SW	трифон	Весь обучающий подкорпус	Как указано выше	+ 4-граммная
NN, NNSW	трифон	Весь обучающий подкорпус	Как указано выше	Как указано выше

Система распознавания речи NN построена на основе нейросетевого подхода. Данная нейросеть использует функцию активации нейронов p-norm и вычисляет апостериорные вероятности для контекстно-зависимых состояний (*Zhang, Trmal, Povey, Khudanpur, 2014*). Нейросетевая система обучена на корпусе с пофонемной разметкой, проведенной с помощью системы Tri4.

Все системы Mono, Tri1-Tri4 и NN используют словарь из 200 тысяч наиболее частотных слов, в то время как SW-системы работают со словарем из 50 тысяч частотных слов и 10 тысяч самых частотных слогов.

1.3.2. Акустические модели

В экспериментах для данной работы мы построили акустические модели для записей хорошего качества: 16 бит, 16 кГц. Они могут быть использованы для распознавания записей, сделанных в офисных помещениях, перед домашним ПК, в условиях небольшого окружающего шума.

Монофонные и трифонные акустические модели построены для 32 звучащих фонем (non-silence) и 2 фонем тишины (silence phones). Контекстная зависимость каждой фонемы реализова-

на соседними фонемами. Таким образом, каждая контекстно-зависимая фонема (CD, context-dependent phoneme) представляется тройкой a-b+c, где b – имя центральной фонемы, a и c – имена левого и правого фонемного контекста.

Фонемный алфавит для татарского языка представлен следующими 34 фонемами: a, ae, b, ch, d, dzh, e, f, g, h, i, j, k, kh, l, m, n, ng, o, oe, p, r, s, sh, t, ts, u, ue, v, y, z, zh, фонемами для тишины sil и короткой паузы sp (short-pause).

1.3.3. Языковые модели

Задача построения языковых моделей возникает при реализации большого количества приложений, начиная от автоматической проверки правописания до систем машинного перевода. Во всех случаях от языковой модели требуется оценить вероятность последовательности слов заданного языка на основе закономерностей, выявленных на обучающем текстовом корпусе.

Языковая модель для распознавания татарской речи была построена с помощью инструмента SRILM (Speech Technology and Research (STAR) Laboratory) (*Stolcke, 2002*). Данный инструмент предоставляет функционал создания n-граммных моделей, умеет интерполировать несколько моделей и оценивать качество построенных моделей. Стандартный способ использования SRILM состоит из нескольких основных шагов:

1. Выполнение скрипта “ngram-count” для расчета количества n-грамм.
1. Выполнение скрипта “ngram-count” для построения языковой модели на основе результатов расчетов 1го шага. Дополнительно может применяться алгоритм сглаживания.

(a) Выполнение скрипта “ngram” с опцией –prune для усечения модели.

2. Определение качества модели с помощью функции “ngram”, вызванной с параметром “ppl”.

Татарский язык принадлежит к агглютинативным языкам, что характеризуется его богатой морфологией. В случае системы пословного распознавания речи единственным способом покрыть весь лексикон – использовать словарь очень большого размера. Эксперимент показал, что словарь из 20 тысяч самых частотных слов даёт значение внесловарных слов (OOV rate), равное 17%, 50 тысяч слов – 10%; даже словарь, состоящий из 200 тысяч са-

мых частотных слов языка даёт 4.4% OOV на тестовом подкорпусе. При использовании послоговой модели удаётся достичь 0% OOV уже при 60 тысячном словаре, состоящем из 50 тысяч самых частотных слов и 10 тысяч слогов. Качество построенных языковых моделей приведено в Табл. 4.

Table 4. **Характеристики языковых моделей**

Языковая модель	Память	Perplexity	OOV
Пословные модели			
Усеченная 3-граммная	31 МБ	1041.9	4.4%
3-граммная	170 МБ	315.9	4.4%
4-граммная	204 МБ	278.9	4.4%
Послоговая модель			
Усеченная 3-граммная	29 МБ	242.8	0%
3-граммная	218 МБ	65.7	0%
4-граммная	360 МБ	45.0	0%

1.3.4. Характеристики оценки качества распознавания

Наиболее распространенной метрикой при оценке качества распознавания является характеристика пословной ошибки WER (word error rate). Агглютинативная природа татарского языка приводит к ситуации, когда данная величина может быть не всегда информативной: один неправильно распознанный в слове аффикс будет учтён как целиком неправильно распознанное слово. Например, слово «калтырадым» из тестового корпуса было распознано как «калтырады» и статистика WER показывает ошибку полностью неправильно распознанного слова. Для предоставления дополнительного источника информации о качестве распознавания мы рассчитали дополнительный показатель послоговой ошибки – SER (syllable error rate).

Одно из важных приложений системы распознавания слитной речи – приложение диктовки сообщений. В данном приложении пользователи оценивают качество работы системы по количеству корректировок символов, которые им необходимо сделать для получения правильного результата. Исходя из этого, также была рассчитана величина побуквенной ошибки – CER (character error rate).

1.4. Результаты

В Табл. 5 представлена оценка качество работы созданных систем на 5-часовом тестовом подкорпусе.

Таблица 5. **Результаты оценки качества работы систем распознавания речи**

Система	Языковая модель	WER		SER		CER	
		Слова	Слоги	Слова	Слоги	Слова	Слоги
Mono	Усеченная 3-граммная	52,06	–	39,65	–	28,70	–
Tri1	Усеченная 3-граммная	28,80	24,08	18,32	13,98	12,54	8,18
Tri1	3-граммная	22,59	18,42	14,09	10,84	9,78	6,44
Tri2	Усеченная 3-граммная	24,14	21,20	13,95	11,46	8,69	6,40
Tri2	3-граммная	19,08	16,17	10,86	9,11	6,91	5,82
Tri3	Усеченная 3-граммная	21,16	18,67	11,35	9,74	6,67	5,33
Tri3	3-граммная	17,21	14,90	9,04	7,81	5,37	4,91
Tri4	Усеченная 3-граммная	18,57	19,70	9,29	10,08	5,24	5,54
Tri4	3-граммная	15,19	16,09	7,46	8,29	4,18	4,59
Tri4	4-граммная	15,10	15,71	7,41	8,05	4,15	4,44
NN	Усеченная 3-граммная	16,47	17,17	8,27	8,13	4,94	4,29
NN	3-граммная	12,99	13,25	6,44	6,37	3,86	3,41
NN	4-граммная	12,89	**12,79**	6,38	**6,14**	3,83	**3,29**

Анализ значений WER для пословных систем показал, что главная составляющая ошибок – ошибки замены слов. Их количество от 5 до 10 раз превышает количество ошибок типа «вставка» и «пропуска». Например, для системы NN с 4-граммной языковой моделью распределение ошибок было следующим: 496 ошибок-«вставок», 395 «пропусков», 2362 «замен». Одна из причин – большое количество внесловарных слов, вторая – богатая морфология татарского языка. Влияние второй причины может

быть заметно также по разнице значений SER и WER: количество ошибок в распознавании слогов примерно в два раза меньше соответствующего значения для слов.

Послоговые системы распознавания показали качество распознавания, сравнимое с пословными системами. При этом для послоговых систем отсутствует проблема OOV-слов, однако определить на основе акустических данных более короткие фрагменты слова оказывается сложнее.

1.5. Заключение

В данной работе представлены результаты работы созданных систем распознавания татарской речи для очень большого словаря. Описываются характеристики первого созданного многодикторного корпуса звучащей татарской речи, использованного для обучения акустических моделей (монофонов и трифонов). Татарский национальный корпус был использован для построения пословных и послоговых языковых моделей. Наилучший результат распознавания слитной татарской речи (87% точность) был получен нейросетевой системой с послоговой языковой моделью.

ЛИТЕРАТУРА

1. Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2016. Ethnologue: Languages of the World, Nineteenth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

2. V. Berment, “Me´thodes pour informatiser des langues et des groups de langues peu dote´es”, Ph.D. Thesis, J. Fourier University, Grenoble I, 2004.

3. S. Krauwer, “The basic language resource kit (BLARK) as the first milestone for the language resources roadmap”, In Proc. of International Workshop Speech and Computer SPEECOM, Moscow, Russia, 2003, P. 8–15.

4. A. Khusainov, “Tekhnologiya avtomatizatsii sozdaniya I otsenki kachestva programmnikh sredstv analiza rechi c uchetom osobennostey maloresursnykh yazikov”, Ph.D. Thesis, Kazan, 2014, 162 p.

5. I. Salimzyanov, J. Washington, and F. Tyers, “A free/open-source Kazakh-Tatar machine translation system”, Proc. of the Machine Translation Summit XIV, Nice, France, 2013.

6. Yandex Translate. Online version: https://translate.yandex.com/translator/Russian-Tatar.

7. D. Suleymnov, A. Gatiatullin, R. Gilmullin, “Lexicograficheskaya baza dannykh dlya system mashinnogo perevoda blizkorodstvennykh yazykov”, In Proc. of Third International Conference «Informatizatciya obschestva», Astana, Kazakhstan, 2012, P. 585–587.

8. A. Khusainov, A. Khusainova, “Speech human-machine interface for the Tatar language”, Artificial Intelligence and Natural Language Conference, Helsinki: FRUCT Oy, 2016. P. 60–65.

9. A. F. Khusainov, D. Sh. Suleymanov, “Language Identification System for the Tatar Language”, Speech and Computer, Lecture Notes in Computer Science. 2013. Volume 8113. P. 203–210

10. 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.

11. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al, “The kaldi speech recognition toolkit”, In Proc. ASRU, 2011, P. 1–4.

12. S. P. Rath, D. Povey, K. Vesely, and J. H. Cernocky, “Improved feature processing for deep neural networks,” In Proc. InterSpeech, 2013.

13. X. Zhang, J. Trmal, D. Povey, S. Khudanpur, “Improving deep neural network acoustic models using generalized maxout networks”, in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, May 2014, P. 215–219.

14. A. Stolcke, “SRILM – An Extensible Language Modeling Toolkit”, Proc. Intl. Conf. on Spoken Language Processing, vol. 2, Denver, 2002, P. 901–904.

СЕКЦИЯ 3 | **UNITURK И КОРПУСНЫЕ ТЕХНОЛОГИИ**

УДК 81’33; 811.512.145

PRELIMINARY TEXT PROCESSING FOR THE COLLECTION OF TATAR NATIONAL CORPUS

M. Ayupov[1], A. Galieva[2]
[1,2]Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic Kazan, Russia
[1]Kazan Federal University. Kazan, Russia
[1]madehur@mail.ru, [2]amgalieva@gmail.com

Tatar National Corpus is a linguistic resource of the modern literary Tatar language. The Corpus contains texts of different styles and genres (fiction, media texts, official documents, educational and scientific literature, etc.). The Corpus has a system of grammatical annotation that is oriented at presenting all the existing grammatical word-forms.

Practical significance of any linguistic corpus is proportional to its volume and representativeness of linguistic data, so tasks on texts collecting, cleaning and reducing to the single format are of current importance. One of ways of replenishing corpus collection is scanning printed texts and further recognizing them.

Standard printed products are heterogeneous in the form of presenting the actual material. In addition to the main text, they may contain various extra-text elements - tables and their names, formulas, drawings, captions, footers, page numbers, etc., all these elements are to be removed from the texts for corpus collection.

When scanning and processing texts a number of difficulties arises, and these difficulties should be solved automatically; because manual processing requires a lot of time and human resources. Main tasks are the following: cleaning the texts from extra-text elements, word-wraps processing and checking the texts for possible typos and spelling mistakes.

Deleting extra-text elements is based on distinctive features of these elements, including their arrangement, size and type of font, parameters of varying, and special software for this task is developed. The word-wrap property allows

long words to be able to be broken and wrap onto the next line, nevertheless such word-wraps pollute linguistic data. Use of Tatar morphological analyzer enables to increase efficiency of word-wraps processing. So special analysis procedures that are implemented in special software are based on basic features of extra-text elements.

Keywords: Tatar National Corpus, morphological analyzer, the Tatar language, text processing.

ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ТЕКСТОВ ДЛЯ КОЛЛЕКЦИИ ТАТАРСКОГО НАЦИОНАЛЬНОГО КОРПУСА

М.М. Аюпов[1], А.М. Галиева[2]
[1,2]Институт прикладной семиотики Академии наук Республики Татарстан, Казань, Россия
[1]Казанский федеральный университет, Казань, Россия
[1]madehur@mail.ru, [2]amgalieva@gmail.com

Татарский национальный корпус «Туган тел» является лингвистическим ресурсом современного литературного татарского языка и содержит тексты различных жанров и стилей. Практическая значимость лингвистического корпуса пропорциональна его объему и репрезентативности языкового материала, поэтому важными являются задачи по сбору текстового материала для пополнения корпусной коллекции, очистке текстов и приведению их к единому формату. Одним из способов пополнения корпуса является сканирование татарских текстов на бумажных носителях с дальнейшим их распознаванием.

При сканировании и распознавании текстов возникает ряд трудностей, которые необходимо решать в автоматическом режиме, так как ручная обработка занимает много времени и человеческих ресурсов. Основными задачами при этом являются следующие: очистка текстов от нетекстовых элементов, обработка переносов, проверка на возможное наличие опечаток и орфографических ошибок. Для обработки массива отсканированных текстов для корпусной коллекции применяются специальные процедуры анализа, реализованные в специальном программном инструментарии.

Для более эффективной обработки переносов использовать морфологический анализатор татарского языка. При разработке программного инструментария для очистки текстов мы основывались на важнейших характеристиках самих внетекстовых элементов, том числе на параметры их варьирования.

Ключевые слова: Татарский национальный корпус, морфологический анализатор, татарский язык, обработка текстов.

1. Введение

Татарский национальный корпус «Туган тел» [1], текстовая коллекция которого постоянно расширяется, является лингвистическим ресурсом современного литературного татарского языка, который позволяет эффективно решать многие лингвистические задачи, в частности, получение новых данных о грамматической структуре и лексическом составе языка, помощь в разработке новых словарей и т.д. Так, материалы корпуса находят применение при разработке Русско-татарского общественно-политического тезауруса [2] и татарско-русского словаря коллокаций [3].

Практическая значимость лингвистического корпуса пропорциональна его объему и репрезентативности языкового материала, соответственно, важной задачей является сбор текстового материала для пополнения корпуса, очистка текстов и приведение их к единому формату. Одним из способов пополнения корпуса является сканирование татарских текстов на бумажных носителях с дальнейшим их распознаванием.

При сканировании и распознавании текстов возникает ряд трудностей, которые необходимо решать в автоматическом режиме, так как ручная обработка занимает много времени и человеческих ресурсов. Основными задачами при этом являются следующие:

– очистка от нетекстовых элементов,

– обработка переносов,

– проверка на возможное наличие опечаток и орфографических ошибок.

Способам разрешения этих трудностей, обусловленных особенностями исходных оригинальных текстов, посвящена настоящая статья.

2. Общая информация о корпусе

Татарский национальный корпус «Туган тел» является лингвистическим ресурсом современного литературного татарского языка и имеет текущий объем 116 миллионов словоупотреблений. Корпус содержит тексты различных жанров (художественная литература, тексты СМИ, тексты официальных документов, учебная литература, научные публикации и др.). Каждый документ имеет

метаописание (автор, выходные данные, жанры, части, главы и др.) [4].

Тексты, включенные в корпус, снабжены морфологической разметкой (представлена информация о части речи основы словоформы и наборе ее грамматических характеристик). Морфологическая разметка текстов корпуса выполняется автоматически с использованием модуля двухуровневого морфологического анализа татарского языка, реализованного в программном инструментарии HFST (Helsinki Finite-State Transducer Technology) [5].

Существенной особенностью лингвистических корпусов является система разметки, от характера и степени разработанности которой во многом зависят возможности, предоставляемые для пользователя. Система морфологической разметки корпуса татарского языка в первую очередь ориентирована на представление всех реально существующих грамматических форм слов, не всегда отражаемых в описательных исследованиях по татарской грамматике либо имеющих различные альтернативные трактовки.

Для формального представления татарской агглютинативной морфологии используется модель, в которой словоформа строится на основе последовательного присоединения к основе регулярных словообразовательных и словоизменительных аффиксов. Например, имя существительное имеет следующую регулярную конструкцию: <основа> <множественность> <притяжательность> <падежность> <модальность> [4].

В татарском языке каждое грамматическое значение, как правило, выражается отдельным аффиксом, аффиксы являются регулярными и в пределах контекста однозначными.

При разработке системы корпусных обозначений для грамматических категорий татарского языка разработчиками были изучены системы обозначений в словарях и справочниках разного типа, в грамматиках тюркских языков, системы грамматической аннотации в имеющихся корпусах, работы по общей морфологии и другие исследования.

Для полного описания морфологической модели литературного татарского языка в настоящее время используется около 60 морфологических тегов.

Так как в настоящее время в корпусе не снята грамматическая омонимия, часть слов получает альтернативные грамматические разборы.

В настоящее время продолжает работа по снятию в корпусе грамматической омонимии. Различные нарушения регулярности морфологии татарского языка, часть которых вызвана большим количеством заимствований с разной степенью освоения и несовершенством современной татарской орфографии, приводят к затруднениям при автоматической обработке, так как на части языкового материала многие морфотактические правила не работают (о подходе к вопросу о снятии грамматической омонимии см. [6]).

Поисковая система корпуса позволяет реализовать поиск по:

– лемме (лексеме);

– словоформе;

– заданному набору морфологических параметров.

Поисковая система татарского корпуса поддерживает поиск минус-слов, поиск по части слова, поиск с использованием логических формул; таким образом, пользователь может задавать сложные запросы, требуемые спецификой своего научного исследования.

3. Очистка текста от нетекстовых элементов

Стандартная печатная продукция является неоднородной по формам представления фактического материала. Кроме основного текста, в нем могут содержаться различные нетекстовые элементы – таблицы и их названия, формулы, рисунки, подрисуночные подписи, колонтитулы, номера страниц и т.п. – то есть элементы текста, которые не должны содержаться в версиях текста для корпусной коллекции. При распознавании отсканированных текстов приходится решать задачу удаления таких нетекстовых элементов.

Для реализации этой задачи можно было бы сохранить документ в текстовом формате, что является быстром способом удаления отдельных видов нетекстовых элементов (например, рисунков и таблиц). Но в этом случаи усложнилась бы задача удаления подрисуночных надписей, колонтитулов и номеров страниц, так как они бы стали элементами текста, а таблицы бы преобразовались в последовательность абзацев, потеряв при этом свою функциональность. Поэтому было решено разработать и реализовать специальное программное решение для удаления нетекстовых элементов.

При разработке программного инструментария мы основывались на важнейших характеристиках внетекстовых элементов, том числе на параметры их варьирования. Например, в том случае, если номера страниц не входят в верхний колонтитул и идут внизу страницы, они представляют собой возрастающие на единицу числа в отдельной строке, которые алгоритмически легко находятся и удаляются. А если номера страниц входят в верхний колонтитул, то они удаляются после применения алгоритма для колонтитулов.

После распознания текста верхние колонтитулы располагаются в отдельной строке и имеют другой размер шрифта (обычно меньший), чем основной текст. Кроме того, для колонтитулов значим параметр четности страниц: все колонтитулы на четных страницах имеют одинаковый вид, что справедливо и для нечетных страниц. Колонтитулы могут отличаться только в случаи включения в них номеров страниц. При этом отличие будет только в начале или в конце колонтитула, где будет идти число, на единицу отличающееся от числа, отвечающего за номер страницы в колонтитуле предыдущей или последующей страницы.

Рисунки обычно сопровождаются подрисуночными надписями, которые идут после рисунка и располагаются между рисунком и основным текстом. Еще одной отличительной особенностью подрисуночных надписей является размер шрифта, отличающийся от размера основного текста, а в большинстве случаев и другой тип шрифта.

Таблицы обрабатываются как рисунки, отличие только в том, что пояснительная запись к таблице идет не после таблицы, а до нее.

Таким образом, удаление внетекстовых элементов основывается на важнейших особенностях самих внетекстовых элементов, с учетом их расположения, размера и типа шрифта, параметров варьирования.

4. Обработка переносов

Так как расстановка переносов позволяет улучшить верстку текста, не допуская зияющих пустот, она широко используется в книжном деле. После сканирования и распознавания текста каждый перенос слова превращается в начало слова, дефис, знак кон-

ца строки и перенесенную часть слова. Необходимо учитывать, что разделенная знаком переноса единица может быть словом, который пишется слитно или словом, который пишется через дефис. Причем слова, которые пишутся через дефис, составляют относительно небольшую часть, в среднем около 5% от общего числа переносов. Можно было бы все эти переносы рассмотреть как слова, пишущийся слитно. Но так как объем татарского корпуса со временем будет увеличиваться, количество ошибочных слов в этом случаи будет расти. Поэтому было решено для более эффективной обработки переносов использовать морфологический анализатор татарского языка [7].

Алгоритм обработки состоит из следующих этапов. На первом шаге предполагается, что слово должно писаться слитно и поэтому из переноса удаляется дефис и знак конца строки; полученная единица передается в морфологический анализатор. Морфологический анализатор пытается разложить данную единицу на корень и цепочку аффиксов. Если это удается, то считается, что слово должно писаться слитно. В противном случае в морфологический анализатор передается единица, написанная через дефис. Если на выходе получаем грамматически правильную словоформу, значит слово пишется через дефис.

Например, возьмем перенос во второй строке на рис. 1. Сначала передаем в морфологический анализатор слитное слово *балачагалар,* на выходе получаем это же слово со знаком вопроса, что означает: данное слово не найдено в базе данных, то есть написано ошибочно. На следующем шаге передаем слово, написанное через дефис: *бала-чагалар*. Морфологический анализатор вернет это слово с примечанием, что оно представляет собой основу плюс аффикс множественности, то есть написание правильное. Следовательно, после обработки, должно писаться через дефис.

Урамда очраган кешеләр Утташ камга олылап баш иделәр. Язгы суда йомычка агызып, чыр-чу килгән балачагалар, Утташ камның бизәлгән, чылтыр- чылтыр килгән атын күреп, баштанаяк ап-ак киемен, каны качкан кырыс йөзен күреп, уеннарыннан туктап калдылар. Кечкенәләр, кинәт котлары алынып, җыларга тотындылар, өлкәнрәкләре, бу безне тотып ала күрмәсен дигән сыман, тизрәк читкә тайпылырга ашыктылар. Усал кам узып киткәч кешеләр артларына борылып карадылар. Куркып кына балалар яңадан уеннарына тотындылар, ярым пышылдап, ерактагы камны үртәргә тотындылар:

Рис. 1. Пример отсканированного текста

Рассмотрим перенос из восьмой строки на рис. 1. На первом шаге передаем в морфологический анализатор слитное слово *артларына*. На выходе получим: основа *арт-* и аффиксы *-лар*, *-ын* и *-а*, то есть такое слово существует и оно написано правильно. Значит это слово, после обработки, в тексте напишется слитно.

Встречаются случаи, когда в начале или в конце разделенного переносом слова уже имеется дефис. В этом случаи в слове автоматически удаляются дефис, который служил для переноса слова и знак конца строки.

Реализованный специальный программный инструментарий позволяет правильно обработать все переносы за исключением случаев неправильного распознавания слов.

5. Проверка на возможное наличие ошибок распознавания и орфографических ошибок

Количество ошибок распознавания зависит от таких факторов, как полиграфическое качество исходника, размер и контрастность текста, сложность взаимного размещения элементов на странице и т.п.

Для проверки татарских текстов на ошибки распознавания и орфографические ошибки также используется морфологический анализатор татарского языка. Слова, которые морфологический анализатор не может разложить на корневую основу и морфемы считаются потенциально ошибочными и выделяются цветом для дальнейшего ручного исправления.

6. Заключение

Таким образом, для обработки массива отсканированных текстов для корпусной коллекции применяются специальные процедуры анализа, реализованные в специальном программном инструментарии. Распознавание и удаление нетекстовых элементов основывается на важнейших особенностях самих нетекстовых элементов, с учетом их типа, особенностей расположения, размера и типа шрифта, а также параметров варьирования нетекстовых элементов. Обработка переносов и проверка текстов на возможное наличие ошибок распознавания, опечаток и орфографических ошибок производится с использованием морфологического анализатора татарского языка.

ЛИТЕРАТУРА

1. Татарский национальный корпус «Туган тел». Электронный адрес: http://tugantel.tatar/

2. Галиева А.М., Кириллович А.В., Лукашевич Н.В., Невзорова О.А., Сулейманов Д.Ш. Создание русско-татарского тезауруса по общественно-политической тематике: общие принципы и аспекты реализации // Научно-техническая информация. серия 2: Информационные процессы и системы. – 2017. – № 2 – С. 20–28.

3. Galieva A., Vavilova Z., Gafarova V. Developing Tatar Corpus-Based Dictionaries for Educational Purposes *// INTED2017 Proceedings, pp. 9014-9022. 11th International Technology, Education and Development Conference Valencia, Spain. 6-8 March, 2017.*

4. Сулейманов Д.Ш., Невзорова О.А., Галиева А.М., Гатиатуллин А.Р., Гильмуллин Р.А., Хакимов Б.Э. Размеченный корпус татарского языка «Туган тел»: аспекты реализации // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. – Казань: Изд-во «Фэн» Академии наук РТ, 2014. – С. 88–93.

5. Helsinki Finite-State Transducer Technology URL: http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/

6. Хакимов Б.Э., Гильмуллин Р.А., Гатауллин Р.Р. Разрешение грамматической многозначности в корпусе татарского языка / Б.Э. Хакимов, // Ученые записки Казанского университета. Сер. Гуманит. Науки. – 2014. – Т. 156, кн. 5. – С. 236–244.

7. Gatiatullin A., Ayupov M. Modifications of morphological analysis programs for the problems of multilingual search // Proceedings of the International Conference «Turkic Languages Processing» TurkLang-2015 – Kazan, 2015. – Pp. 120–126.

УДК 81

FROM KYRGYZ INTERNET TEXTS TO AN XML FULL-FORM ANNOTATED LEXICON: A SIMPLE SEMI-AUTOMATIC PIPELINE

L. Boizou, D. Mambetkazieva
Vytautas Magnus University, Kaunas, Lithuania
loic.boizou@vdu.lt

With the intention to foster the development of new free resources for Kyrgyz, the present paper describes a simple semi-automatic pipeline that generates a full-form lexicon out of a corpus made from texts freely available on the internet. All components were developed as short Haskell programs. The corpus, which comprises about 1.6 million words and 170 texts, includes various genres, such as literary works (novels, short stories, plays), laws and other regulatory texts, news, institutional websites of companies, universities, government structures, Wikipedia articles. Its design (genre proportion, text length) is far from optimal, but, our aim is to get a significant lexical coverage of standard written Kyrgyz, more than a faithful representation of the language.

A word list was automatically extracted from the corpus and items with non-Kyrgyz characters were filtered out. The resulting list of about 130,000 distinct word forms was analysed as morphemic sequences with a set of grammatical values. This morphological analysis relies on a simple finite-state machine which describes Kyrgyz word structure. This FSM is stored in a simple text file, where each line is a transition. Each transition represents a single morpheme surface form, except the morphemic sequence possession + case, which exhibits some irregularities and is represented as one transition for the sake of simplicity. This analyser works like a morphological guesser: it provides a list of all formally plausible morphemic segmentations for each word form. There is no information about existing stems and no disambiguation is performed at this stage.

The FSM-based morphological analysis is followed by several formal validations aimed at reducing the number of ambiguous morphological interpretations. The program checks basic parameters related to the stem structure, such as the minimal length (at least two letters) or the property of its final segment (two letter words should end in a consonant, except *де* and *же*, final consonant clusters should be retained only if there is no alternative analysis). Then, the output file is manually corrected, by adding the plus sign before each correct interpretation. If there is no correct interpretation, the corrector can directly provide the right morphological interpretation. Besides, the corrector can add a minus sign before any incorrect extracted stem (be the error in the form or in the part of speech)

and after each correction session, the list is processed again to remove interpretations with marked wrong stems. Given the lack of collaborators, this correction is currently carried out without a double-check or multi-annotator team.

Finally, all corrected word forms are included in an automatically generated XML lexicon, which follows the TEI P5 guidelines. The grammatical information is converted according to the Universal Dependency tagset. Such widely used standards make data easily reusable by other researchers and compatible with other tools.

Keywords: Kyrgyz language; morphology; finite-state machine; corpus-based lexicon; Universal dependency.

ОТ КЫРГЫЗСКИХ ТЕКСТОВ ИЗ ИНТЕРНЕТА ДО АННОТИРОВАННОМУ XML-ЛЕКСИКОНУ СЛОВОФОРМ: ОПИСАНИЕ НЕСЛОЖНОГО ПОЛУАВТОМАТИЧЕСКОГО КОНВЕЙЕРА

Л. Буазу, Д. Мамбетказиева
Университет Витаутаса Великого, Каунас, Литва
loic.boizou@vdu.lt

В целях содействия развитию новых свободных ресурсов для кыргызского языка в настоящей статье описывается простой полуавтоматический конвейер, который создает лексикон словоформ на основе корпуса, созданного из свободно распространяемых в Интернете текстов. Все компоненты были разработаны как короткие программы Haskell.

Корпус, который состоит приблизительно из 1,6 миллиона слов и 170 текстов, включает в себя различные жанры, такие как литературные произведения (романы, повести, пьесы), законы и другие нормативные тексты, новости, институциональные сайты компаний, университетов, правительственных структур, статьи в кыргызской Википедии. Его структура (жанровая пропорция, длина текстов) неоптимальная, но наша цель - получить значительный лексический охват стандартного письменного кыргызского языка, более чем точное представление языка.

Список слов был автоматически извлечен из корпуса, а элементы с не кыргызскими символами были отфильтрованы. Полученный список из примерно 130 000 различных форм слов был проанализирован как морфемные последовательности с грамматическими значениями. Этот морфологический анализ осуществлен простым конечным автоматом, который описывает структуру кыргызского слова. Этот конечный автомат хранится

в простом текстовом файле, где каждая строка является переходом. Каждый переход представляет собой единую морфемную форму поверхности, за исключением часто неправильной морфемной последовательностьи принадлежность + падеж, которая представляется как один переход для простоты. Эта программа работает как морфологический анализатор незнакомых слов (guesser). Он предоставляет список всех формально правдоподобных морфемных сегментов для каждой формы слова. Информация о существующих основах отсутствует, и на данном этапе не проводится устранения морфологической неоднозначности.

После морфологического анализа следуют несколько формальных валидаций, направленных на уменьшение числа неоднозначных морфологических интерпретаций. Программа проверяет базовые параметры, связанные со структурой основы, такие как минимальная длина (по крайней мере две буквы) и свойство ее конечного сегмента (двухбуквенные слова должны заканчиваться согласным, за исключением де-и же, окончательные консонантные кластеры должны быть сохранены только если альтернативного анализа нет). Затем выходной файл вручную корректируется, добавляя знак «плюс» перед каждой правильной интерпретацией. Если нет правильной интерпретации, корректор может непосредственно предоставить правильную морфологическую интерпретацию. Кроме того, корректор может добавить знак «минус» перед любой некорректной извлеченной основой (ошибка может быть в форме или в части речи), и после каждого сеанса коррекции список обрабатывается снова, чтобы удалить интерпретации с отмеченными неправильными основами. Учитывая отсутствие сотрудников, эта коррекция в настоящее время выполняется без проверки команды аннотаторов.

Наконец, все скорректированные текстовые формы включены в автоматически созданный XML-лексикон, который следует руководящим принципам TEI P5. Грамматическая информация преобразуется в соответствии со стандартом Universal Dependency (универсальная зависимость). Такие широко используемые стандарты делают данные легко и повторно используемыми другими исследователями, а также совместимыми с другими инструментами.

Ключевые слова: кыргызский язык; морфология; конечный автомат; лексика основанная на корпусе; Universal Dependency.

1. Introduction

Despite its official status in Kyrgyzstan, the Kyrgyz language still lacks NLP resources, especially open ones. For example, Sketch Engine provides a Turkic web corpora (Baisa, Suchomel, 2012) with a significant Kyrgyz part (about 20 million words), but this valuable resource is not freely available. Tamgasoft developed an online morphological analyser (http://tamgasoft.kg/morfo/), but it does not allow analysing full texts, and its lexical coverage seems low. Other resources such as spell checkers (https://www.spellchecker.net/kyrgyzstan_spell_checker.html), script converters between Cyrillic, Latin and Arabic (http://www.transliteration.kpr.eu/ky/, http://translit.kerben.org/) and online bilingual dictionaries are not suitable for automatic processing either.

The present paper briefly presents an attempt to generate semi-automatically a full-form lexicon of standard written Kyrgyz based on texts freely available on the internet. Such a lexicon could be used for developing future resources, especially if it relies on widely accepted standards. The different components, which are developed in Haskell, constitute a light pipeline from the corpus data to the XML lexicon. The main component of this pipeline is a morphological analyser based on a finite-state machine, which functions as a guesser.

The first part of this paper describes the corpus and the extraction of the initial list of word forms. The morphological analysis, which mainly relies on a simple finite-state machine with few extra processing and filtering tasks, is presented in the second part. The third part deals with the process of manual correction of the results and the choices related to the final lexicon, its structure and the encoding of grammatical values.

2. Lexical list acquisition

This section presents how the initial list of word forms was prepared out of a corpus and describes the corpus itself.

2.1. The corpus

In order to create a lexicon based on textual materials and rather than on existing dictionaries, a corpus was collected 5 years ago from different internet sources. The whole corpus includes about 170

texts and 1.6 million words. The biggest part (about half of the corpus) is made up of literary texts of various genres, novels, novellas (*повесть*) and plays, which were taken from the website http://www.literatura.kg. Texts are original works in Kyrgyz with a few exceptions translated from Russian (*Кассандра тамгасы* by Aitmatov, *Иван Денисовичтин бир күнү* by Solzhenitsyn). Some texts (about 1/5 of the corpus) come from a variety of news websites, e.g. *Erkin Too*, *Super Info*, *Azattyk* and *BBC Kyrgyz*. Other texts (about 1/3 of the corpus) were taken from the websites of government institutions (Kyrgyz Presidency, Kyrgyz National Bank), universities (Kyrgyz National University, Manas University…), companies (Kumtör, Megafon) and from the Kyrgyz Wikipedia. Texts from social media, which are often closer to the spoken language and rich in non-standard elements, were purposely not included in the corpus.

This corpus is far from optimal, the amount of news texts is too small and some large texts have excessive influence on the corpus. The overweight of Manas epics (about 75,000 words for both parts) and of the documents from the Kyrgyz National Bank (about 120,000 running words) is obvious: a quick survey of the word list shows that *банк* and *Манас* appear respectively as the 47th and 63rd most frequent words. Nevertheless, a fair representativeness is more an issue for creating frequency-based dictionaries or for quantifying language phenomena. In order to get a rough list of words with a sufficient coverage of the general vocabulary, the corpus size and genre diversity are more important. We are aware that the size of the current corpus is minimal, but the diversity seems acceptable.

2.2. The word list

The list of word forms was extracted from the corpus with a simple Haskell function and sorted in alphabetic order. This function filtered out strings with non-Kyrgyz characters. Most of these strings are Latin words, incorrect hybrid strings that mixed Cyrillic and Latin symbols with similar glyphs (ex. *a*, *e*, *o*, *c*) and foreign words from Kyrgyz Wikipedia articles.

The resulting list contains slightly less than 130,000 distinct word forms (word forms that start with a capital letter are treated as distinct from the one with a minuscule), with a small percentage of Russian words. In Kyrgyz, a huge number of words were borrowed from Rus-

sian without orthographic adaptation (that is, the spelling is the same in Russian and Kyrgyz). Thus, it is not always possible to determine if a base form in a word list is a loanword from Russian used in Kyrgyz (be it considered as an acceptable loanword in standard Kyrgyz or a barbarism) or a Russian word (in a fragment in Russian). The presence of Kyrgyz suffixes clearly indicates that the word form is a loanword, e.g. *ядрого* (*ядро* "nucleus" loanword from Russian with Kyrgyz dative suffix *-го*) "to the nucleus", while the presence of Russian inflection (except for the form used as lemma) shows that the word form is a genuine Russian word. Besides, it should be mentioned that only nouns (in nominative singular) are really problematic, e.g. *ядро* "nucleus", while verbs and adjectives borrowed from Russian are regularly integrated with suffixes, e.g *ядролук* "nuclear" (vs. Russian *ядерный*). Without social media texts, which were not included in the corpus, Kyrgyz written texts are relatively free from Russian words considered as barbarisms.

The extracted list of word forms constitutes the input for the core component of the pipeline that performs the morphological analysis.

3. Morphological analysis

The morphological analysis is implemented in three successive steps: (1) the pre-processing of the input data, (2) the morphological analysis, and (3) the post-processing of the morphologically annotated output.

3.1. Pre-processing

In Kyrgyz, letters *ю*, *я* and *ё* are clearly biphonemic: they represent respectively the sequences /j/ + /u/, /j/ + /a/ and /j/ + /o/. If such letters are inside a stem, e.g. *таяк* "stick", they have no influence on the morphological structure, but they can also stretch over a morphemic boundary. For example, the word form *коюп* combines the stem *кой* "to put" (also used as a frequent auxiliary) with the gerund suffix *-уп*, so that the letter *ю* is shared between the stem (the /j/ part) and the suffix (the /u/ part). In order to simplify the morphological analysis, *ю* and *я* are rewritten as a sequence of two letters, respectively *йу* and *йа*, e.g. *коюп* → **койуп* (to be analysed as *кой-уп*). The letter *е* can also be biphonemic, but only after a vowel, and is corrected accordingly in such a context, e.g. *тиер* "will touch"→ * *тийер*.

3.2. The core component

The morphological analysis was implemented using an incomplete finite-state machine. The automaton is represented as a simple Unicode text file, which is parsed and loaded by the Haskell program.

3.2.1 Description of the automaton

In the text file that stores the automaton, each non-empty line represents a transition. Each transition is made up of four fields separated by tabulations: current state, surface string, grammatical values, next state. Empty strings, used for null morphemes, are noted as /\ (for λ). For example, the polarity is represented by the following transitions (where the empty string stands for the affirmative):

```
25  /\      P:AFF  26
25  ба      P:NEG 26
25  па      P:NEG 26
25  бо      P:NEG 26
25  по      P:NEG 26
25  бе      P:NEG 26
25  пе      P:NEG 26
25  бө      P:NEG 26
25  пө      P:NEG 26
```

The automaton is made up of about one thousand transitions that represent the main Kyrgyz suffixes. No device was developed in order to derive the different surface forms out of the main suffix form, therefore each surface form had to be explicitly added as a transition. For example, as previously shown, eight transitions (with the same current state and next state), one for each concrete realisation, correspond to the negative suffix (*ба*, *па*, *бо*, …, *пө*).

As a rule, each transition corresponds to one morphemic surface form. The main exception concerns the sequence of the possession suffix + case suffix, since such sequences exhibit some irregularities, especially for the third person.

Table 1. **Irregularities in the morphemic sequences Possession + Case**

Nominative	Accusative	Genitive	Dative	Locative	Ablative	Adverbial
эл ‘people’	*эл-ди*	*эл-дин*	*эл-ге*	*эл-де*	*эл-ден*	*эл-дей* *эл-че*
эл-им ‘my people’	*эл-им-ди*	*эл-им-дин*	эл-им-е	*эл-им-де*	*эл-им-ден*	*эл-им-дей* *эл-им-че*
эл-и ‘his/her people’	эл-ин-	эл-ин-ин	эл-ин-е	*эл-ин-де*	эли-ин-ен	*эли-ин-дей* *эли-ин-че*

In the automaton, two types of suffixes, derivational and flexional, are distinguished. The first category includes suffixes that are expected in lexicon items and that are not systematic (that is, the suffix is combined with some lexical stems of a given part of speech only), e.g. *-чы*, *ыр* “song”, *ырчы* “singer”, *-мак*, *куй* “to pour”, *куймак* “pancake”, -л*а*, *жаза* “punishment”, *жазала* “to punish”, *-тыр*, *өл* “to die”, *өлтүр* “to kill”. The second category includes suffixes that are more utterance-based: they are compatible with any lexical stem (of a given part of speech) and cannot be retained in further derivation. It includes plural suffix, negative suffix, interrogative suffix, temporal-modal suffixes, possessive and person suffixes, case suffixes and some similar adverbial suffixes (*-ча*, *-дай*).

There is no clear boundary between the two categories: some suffixes are in the middle, they can be adapted to any lexical stem (of a given part of speech), but they can remain in derivation, e.g. *сыз*, *карындаштарым-сыз* “without my younger sisters” (utterance-based), vs. *жумуш-суз-дук* “unemployment” (embedded in a lexical derivative). Such suffixes appear in two distinct positions in the automaton, as case-like adverbial transitions (as a flexional suffix) and as derivational transitions.

In Kyrgyz, the difference between nouns and adjectives is not obvious from syntactic suffixes. Adjectives normally bear no syntactic suffixes, but they can be used as nouns and bear syntactic suffixes, especially with the plural suffix, e.g. *акылдуу* “clever”, *акылдуулардан* “from the clever (people)”. Nouns normally bear suffixes (although null morphemes are not obvious), but they can be used to express quality and have no flexional suffix in such a case. To a large extent,

the difference is syntactic, as illustrated in the following examples: *кыргыз сөз* “Kyrgyz word”, *жаш кыргыз* “a young Kyrgyz”, *кыргыз жаштар* “Kyrgyz youth (young people)”, *кымбат жыгач* “precious wood”, *жыгач табак* “wooden plate”. Some nominals are more quality-like, others - more substance-like, but it is not very obvious on the morphological level, except for some derivational suffixes. Therefore, the automaton differentiates between nouns and adjectives only for such suffixes, in other cases, word forms are tagged as nominal (marked as CL:N).

3.2.2 Analysis

The morphological analysis is performed from the end of the word form, since Kyrgyz morphological processes are based on suffixation. As mentioned before, the automaton described in the text file is incomplete: it contains only suffixes and no root or stem. When the parsing fails, the remaining string is tagged as *root* and returned together with the sequence of recognized suffixes. The analysis function provides all analyses formally possible for a given input string (internally structured as a binary tree of transition), without disambiguation. Thus, this analyser works as a guesser: it has no clue whether the remaining root exists or not, and provides the set of plausible morphological interpretations. Example (LC means lexical category):

кыш LC:N
кыш LC:V

Such analysis is formally possible, cf. *жаз* V “to write” / *жаз* N “spring”, but, in this case, it is factually wrong: there is no verb **кыш*, only a noun *кыш* “winter”.

The Haskell program can generate two different outputs, a “morphemic” view, which provides for each interpretation the root and the list of recognized suffixes (with the form and value of each matching transition), and the “lexical” view, which provides for each interpretation the root, the stem (the root with its derivational suffixes), and the flexional features (number, case, person, tense…).

3.3. Post-processing

Some operations are carried out after the FSM analysis. Inside stems, sequences *йа, йу, йо and йе* (generated during the pre-processing) are replaced by their genuine orthographic counterparts *я*, *ю, ё*

and e, e.g. **майак* is reversed back to its original written form *таяк*. Besides, Kyrgyz stems that end in *-к* and *-п* are voiced before a suffix that starts with a vowel, ex. *китеп* "book", *китеб-им* "my book", *чык* "to go out", *чыг-ат* "goes out". Therefore, both final consonants *-г* and *-б* are corrected as *-к* and *-п* before suffixes with initial vowels, although it could lead to mistakes with some nominal loanwords, such as *микроб* "microbe", where the *-б* is stable.

Further operations related to the stem are aimed at reducing the number of ambiguous morphological interpretations. They are very similar to the process described in Moral et al. (2014). Interpretations with one letter stem are removed, as well as two letter stems that end in a vowel, except *де* "to say" and *же* "to eat". Many final consonant clusters do not occur in Kyrgyz stems except in loanwords, so that the program discards such stems, e.g. **илимд* vs. *илим*, for the word form *илимди* "science" (accusative, singular), unless no other interpretation is available, e.g. for *паркты* "park" (accusative, singular), the (correct) base form *парк* will not be removed.

As a result of an incorrect segmentation, some interpretations suggest a too long stem (under-stemming), e.g. **арпага*, **арпак*, instead of *арпа*, for the word form *арпага* "barley" (dative, singular), or a too short one (over-stemming), e.g. **ил, *или*, instead *илим*, for the word form *илим* "science" (nominative, singular). Since the morphological analyser provides all plausible interpretations, the correct stemming should appear among the different options. It is possible to reduce the number of morphological interpretations with false stems automatically, but it should be cautiously weighted, because in some cases several alternative analyses are possible. For example, the word form *жаздык* means "pillow" as a noun (as base form without ending), but as a form of the verb *жаз* "to write", it means "we wrote". To avoid harmful suppressions, we decided to implement only the filtering of the longest stem, which may result from under-stemming, when several plausible stems are provided for a given word form: the longest stem is compared with the list of all extracted stems, and if this stem does not exist with other (flexional) suffixes, it is discarded. For example, the mentioned form *жаздык* appears as a possible stem in longer word forms, e.g. *жаздыктан*, and is preserved, while *жаздыктан*, which does not appear as a possible stem in longer forms, has to be removed. Nonetheless, this simple operation alone is not able to identify all cases of under-stemming.

In the whole pipeline, the list of morphologically annotated word forms is not the final result: it is used as a rough data to build the XML lexicon.

4. Lexicon building

The list of morphologically annotated word forms is manually corrected. This work is in progress, but it is time-consuming given the size of the data. If suitable interpretations are available, it is sufficient to add the plus sign before each correct interpretation. If no morphological interpretation is correct, it is always possible to enter directly the right one(s), but no specific device has been set in order to avoid mistyping and accidental data corruption. It is also possible to enter a minus sign before non-existent stems. Then, a program that can be called after each annotation session collects these incorrect stems and removes related interpretations from the list of annotated word forms.

The main lexicon is generated automatically out of the manually corrected annotated word forms. This lexicon can integrate corrected results block by block, so that it is not necessary to have the full word form list corrected before starting the generation process. In order to be easily shared and reusable, this lexicon is an XML file, which relies on two major standards, TEI P5 for the file structure (http://www.tei-c.org/index.xml) and Universal Dependency (http://universaldependencies.org/, Nivre:2015) for the grammatical features. The TEI structure of the lexicon directly follows examples presented by Budin, Majewski and Mörth (2012). Grammatical information is encoded with the Universal parts of speech and Universal features. Although some problematic aspects remain (Çöltekin, 2015), Turkish is largely integrated in the Universal dependency framework (Eryigit et alii, 2016) and provides a suitable background to describe Kyrgyz with UD categories.

Most Kyrgyz features have natural counterparts in Universal features, with the mechanism of layered features for possessive suffixes, e.g. *дос-убуз-суң* “(you) are our friend” would be annotated as : Number[psor]=Plur, Person[psor]=1 (*-убуз*, “our”, 1st person plural for the possessor), Number=Sing, Person=2, Polite=Infm (*-суң*, “are”, 2nd person, singular, informal)

We cannot present explicitly the whole set of categories in the present paper, but the main issues are related to the verb categories,

as in Turkish (Kaşıkara, 2015). The features used in the lexicon for the temporal-modal and the non-finite forms are illustrated with the verb *жаз* "to write" (other categories, e.g. person or number, are not mentioned):

Table 2. **Selected UD categories for verbs**

Finite verb forms (3rd person)	UD categories
жазды	Tense=Past, Aspect=Perf
жазган	Tense=Past, Aspect=Imp
жазучу	Tense=Past, Aspect=Iter
жазыптыр	Tense=Past, Evident=Nfh
жазат	Tense=Pres (although it often expresses future)
жазар	Mood=Pot
жазса	Mood=Cond
жазсын	Mood=Imp (although Mood=Opt may be better)
Participles and gerunds	UD categories
жазуу	VerbForm=Inf
жазган	VerbForm=Part (homonym of the finite imperfect form)
жаза	VerbForm=Conv, Tense=Pres (Aspect=Imp might be an option)
жазып	VerbForm=Conv, Tense=Past (Aspect=Perf might be an option)

Some decisions should be discussed among a wider audience, but at the current stage the most important is to have a clear description how verbal categories are annotated, which allows for future automatic re-tagging, if collective discussions lead to different choices.

5. Conclusive remarks

• The precise number of words in the lexicon is still unclear. A rough estimate based on a small sample of the word list allows us to evaluate the lexeme to word form ratio of about 1/5, so that the number

of lexemes should be around 20,000-25,000, which represents a good starting point for a general lexicon.

• The pipeline is slowed by the manual correction of the 130,000 word forms. A team of several annotators would speed up the process.

• The final decision about categories used for grammatical annotation should involve more researchers and should probably take into account a common Turkic perspective.

Future work includes the realization of a new morphological analyser based on the corrected annotated lexicon that would perform better than a guesser. Furthermore, some steps to improve the corpus are under way. More texts from news websites are already ready to be included in the corpus. Then, we intend to annotate this improved corpus with the new morphological analyser and to train a disambiguation tool.

REFERENCES

1. Baisa, V., & Suchomel, V. (2012). Large Corpora for Turkic Languages and Unsupervised Morphological Analysis, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey.

2. Çöltekin, C. (2015). A Grammar-Book Treebank of Turkish, Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14). Warsaw, Poland.

3. Eryigit, G., Gokirmak, M., Nivre, J., Sulubacak, U., Tyers, F.M., & Çöltekin, Ç. (2016). Universal Dependencies for Turkish. COLING.

4. Kaşıkara, H. (2015). Universal Dependency Representation of Turkish: The Challenge of the Verb, Master thesis. Uppsala University.

5. Moral, C., de Antonio, A., Imbert, R., & Ramírez, J. (2014). A survey of stemming algorithms in information retrieval, Information Research, 19, 1.

6. Nivre, J. (2015). Towards a Universal Grammar for Natural Language Processing. In A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing (pp 3–16). Springer/

УДК 811.512.145; 81'367.625

TATAR P-CONVERBS: SEMANTIC CLASSES AND DISTRIBUTION (ON CORPUS DATA)

A. Galieva, R. Gataullin
Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic Kazan, Russia
amgalieva@gmail.com

The paper deals with semantics and distribution of Tatar converbs functioning within analytical verbs and verb constructions. Available literature on Turkic converb constructions describes diverse grammatical, semantic and syntactic features of such items, nevertheless, there is a lack of quantitative data get on a large volume of linguistic material. The quantitative analysis based on peculiarities of structural components and the meaning of converb constructions as well as subsequent interpretation of linguostatistic data enables to detect core and peripheral classes of these items and to determine main trends in their formation and functioning. Using text collection of Tatar National corpus, we detected the most frequently used types of converb constructions. The main attention is paid to constructions built according to the model: p-converb + verb *alu* 'take' (1 335 unique constructions are selected). We distinguished the most frequent semantic classes of these constructions and present quantitative data on distribution of them in the Corpus focusing on the interaction of the lexical and grammatical semantics.

Verbs denoting taking are at the intersection of object moving and possession and they show a tendency to a greater extent of polysemy. The ambiguity of the verb *alu* 'to take' leads to the ambiguity of constructions it containing.

The most frequent constructions are coined by converbs denoting perceptions (especially visual perception verbs are frequent), possession, conquest and intellectual activity. The different semantic classes of converb constructions manifest different degree of grammaticalization of the second component of a construction.

Keywords: the Tatar language, converb, verb semantic classes, converb constructions, corpus.

ТАТАРСКИЕ КОНВЕРБЫ НА -П: СЕМАНТИЧЕСКИЕ КЛАССЫ И СТАТИСТИКА (НА КОРПУСНЫХ ДАННЫХ)

А.М. Галиева, Р.Р. Гатауллин
Институт прикладной семиотики Академии наук Республики Татарстан, Казань, Россия
amgalieva@gmail.com

В статье на корпусных данных рассматривается вопрос о семантике и статистическом распределении татарских конвербов на -п в составе аналитических глаголов и глагольных конструкций разных типов. Имеющиеся исследования по конструкциям с тюркскими конвербами описывают различные грамматические и семантико-синтаксические особенности таких единиц, но без приведения количественных данных на большом объеме лингвистического материала. Между тем количественный анализ, выполненный с учетом особенностей как структурных компонентов, так и семантики конвербных конструкций с последующей интерпретацией лингвостатистических данных позволяет выявить ядерные и периферийные классы единиц и основные тенденции в их образовании.

Основное внимание уделено аналитическим глаголам (конструкциям), образованным по модели: *конверб на -п + алу* ('брать'), выделены наиболее частотные семантические группировки таких глаголов, представлены количественные данные по их распределению в коллекции текстов Татарского национального корпуса.

Глаголы со значением 'брать, взять' находятся на пересечении глаголов перемещения объекта и обладания, и в языках мира, в том числе и в татарском языке, обладают чрезвычайно развитой полисемией. Многозначность глагола *алу* во многом обусловливает многозначность глагольных конструкций с его участием.

Самую частотную группировку представляют конструкции, образованные при помощи конвербов, обозначающих восприятие; среди них основную массу составляют глаголы, связанные с зрительным восприятием. В данной группе финитный глагол *алу* десемантизирован и выражает аспектуальные характеристики, связанные с результативностью действия по восприятию. Тем не менее, кроме грамматикализации глагола алу, здесь можно усмотреть освоение восприятий, вовлечение их в личное пространство человека.

Следующая значимая по частотности группировка образована конвербами приобретения, обозначающими взятие в руку, которые с глаголом *алу* образуют аналитические глаголы со значением приобретения, а также связанные с ними глаголы со значением завоевания и захвата. В данной группе прослеживается связь с прямым значением глагола *алу* ('брать').

Еще одна группа конструкций образована конвербами, обозначающую интеллектуальную деятельность). В данной группе глагол *алу* десемантизирован и выражает различные аспектуальные характеристики.

Ключевые слова: татарский язык, конверб, конструкции с конвербом, семантические классы глагола, корпус.

1. Введение

Тюркские конвербы на -п образуют различные глагольные конструкции (аналитические глаголы), образованные по модели: *конверб на -п + финитный глагол* (термин «финитный глагол» для обозначения второго компонента конструкции здесь использован с определенной долей условности, так как конструкции с конвербами могут быть использованы в нефинитной форме в составе более сложных глагольных цепочек; более того, в статье эти условно финитные глаголы представлены в форме имени действия). В данном материале термин «финитный глагол» мы считаем более предпочтительным, чем «вспомогательный глагол» или «сериальный глагол»; последние термины, хотя и используются в литературе, не всегда удовлетворяют описанию конвербных конструкций по причине того, что семантика и функции глаголов в анализируемых конструкциях достаточно разнообразны: это далеко не всегда глагол со служебным статусом или конструкция со значением сериализации. Термин «финитный глагол» удобен тем, что он противопоставлен конвербу и акцентирует лишь формальный аспект: в абсолютном большинстве корпусных контекстов второй глагольный компонент стоит в личной форме и может выражать время, лицо, число, залог, отрицательность.

Имеющиеся исследования по конструкциям с тюркскими конвербами описывают различные грамматические и семантико-синтаксические особенности таких единиц, но без приведения количественных сведений на большом объеме лингвистического материала (Юлдашев 1977, Гращенков 2011, Гращенков 2012, Гращенков 2015, Матушкина 2016 и др.). Между тем количественный анализ на большом массиве данных, выполненный с учетом особенностей как структурных компонентов, так и семантики конвербных конструкций и последующая интерпретация лингвостатистических данных позволяет выявить ядерные и периферийные классы анализируемых единиц и основные тенденции в их образовании и функционировании.

Лингвистический материал извлечен из Татарского национального корпуса «Туган тел» (http://tugantel.tatar/), объем текстовой коллекции, ставшей базой для данного исследования, составил 21 935 876 словоупотреблений. Нами были извлечены конструкции с конвербами на -п (всего 48996 уникальных конструкций), затем из них были сформированы классы с наиболее частотными финитными глаголами. Основное внимание в статье уделено конвербным конструкциям, образованным при помощи глагола *алу* ‘брать’ в качестве второго компонента.

Многие конвербные конструкции являются многозначными, и степень десемантизации или грамматикализации второго компонента точно можно определить лишь после детального анализа каждого контекста употребления. В статье отражены лишь наиболее типичные значения конвербных конструкций, без учета многозначности компонентов или конструкции в целом.

2. Татарские конвербные конструкции: предварительные сведения

Конструкции с конвербами, в зависимости от компонентов, могут выражать одновременные (*карап йөрү* 'ходить и смотреть', *сокланып карау* 'смотреть и любоваться') или последовательно разворачивающиеся события (*чыгып керү* 'зайти и выйти', *сугышып ару* 'устать в результате драки'). Кроме того, конструкции с конвербами могут быть лексикализованы (*уйлап табу* 'изобретать', *сатып алу* 'купить') или грамматикализованы их компоненты. Типичный случай грамматикализации – функционирование конверба *дип* 'сказать' в качестве цитатива при вводе прямой или косвенной речи. Грамматикализованными также можно считать многие случаи использования глаголов бытия, положения в пространстве и движения (например, *тору* 'быть, находиться', *яту* 'лежать', *бару* 'идти') в качестве вспомогательных глаголов в конструкциях с конвербами.

Образующие конвербную конструкцию глаголы с точки зрения переходности могут быть: оба переходные (*куып чыгару* 'гнать' + 'выпускать', *карап алу* 'смотреть' + 'брать'), оба непереходные (*кереп чыгу* 'заходить' + 'выходить', *йоклап яту* 'спать' + 'лежать'), переходный и непереходный (*укып чыгу* 'читать' + 'выходить', *карап тору* 'смотреть' + 'стоять'), непереходный и пере-

ходный (*сискәнеп кую* 'вздрогнуть' + 'положить', *көлеп җибәрү* 'смеяться' + 'отправлять').

Такие конструкции, в зависимости от структуры, могут выражать также различные модальные значения, так, конструкция типа: конверб на -п + *була* выражает идею объективной безличной возможности/невозможности:

- *барып була* 'можно пойти/дойти';
- *күтәреп булмый* ‘нельзя поднять’.

Представляет интерес количественное распределение типов конструкций с различными финитными глаголами для уточнения характера связи компонентов конвербной конструкции и взаимовлиянии грамматического значения конструкции и лексического значения компонентов в случаях грамматикализации второго компонента.

Поэтому на первом этапе нами были получены данные о глаголах, которые присоединяют к себе наибольшее число различных конвербов, образуя конструкции разного типа.

В таблице 1 представлены количественные данные о типах конструкций по частотности в зависимости от финитного глагола: в столбце 3 представлено количество уникальных конструкций по лемме (а не количество их употреблений в корпусе).

Таблица 1. **Распределение конвербов на -п, образующих аналитические глаголы (конструкции) с разными финитными глаголами**

Тип конструкций	Финитный глагол	Количество уникальных конструкций
-п тору	'быть, находиться, стоять'	1 762
-п йөрү	'ходить'	1 441
-п алу	'брать'	1 335
-п кую	'положить'	1 021
-п яту	'лежать'	1 007
-п утыру	'сидеть'	980
-п калу	'оставаться'	932
-п карау	'смотреть'	849
-п китү	'уходить'	766

-п бару	'идти'	727
-п булу	'быть'	678
-п чыгу	'выходить'	674
-п килү	'приходить'	666
-п яшәү	'жить'	647
-п бетерү	'закончить'	600
-п өлгерү	'успеть'	552
-п кайту	'вернуться'	508
-п җибәрү	'отправлять'	462
-п керү	'входить'	410
...		
Всего		**48 996**

Как показывает таблица 1, наибольшее разнообразие в образовании конвербных конструкций демонстрируют глаголы бытия (*тору, булу, яшәү*), положения в пространстве (*ятү, утыру*), движения (*йөрү, китү, бару, чыгу, килү, кайту, керү*), а также отдельные глаголы других семантических классов.

Далее для каждого класса, образованного при помощи соответствующего финитного глагола, были выделены наиболее частотные семантические группировки, обусловленные семантикой конверба и конвербной конструкции. Материал данной статьи далее ограничен данными о структурной и семантической организации конструкций с финитным глаголом *алу* ('брать').

3. Конструкции с финитным глаголом *алу*

Глаголы со значением 'брать, взять' находятся на пересечении глаголов перемещения объекта и обладания и в языках мира обладают чрезвычайно развитой полисемией. Представляется интересной задача по выделению и систематизации значений татарских частотных конвербных конструкций со вторым компонентом со значением 'брать, взять' и описанию связей между ними.

Методика исследования строится следующим образом: из общего списка аналитических глаголов (глагольных конструкций), образованных при помощи конверба на -п и финитного глагола *алу*, были выбраны 200 наиболее частотных аналитических глаго-

лов (конструкций) и далее исследована их структура и семантика с выделением основных семантических группировок.

Самую частотную группировку представляют глаголы, образованные при помощи конвербов, обозначающих восприятие (таблица 2).

Таблица 2. **Конструкции со значением восприятия**

Конструкция с конвербом	**Значение конверба**	**Значение конструкции**	**Употребление в корпусе**	**Количество разных форм финитного глагола в корпусе**
карап алу	'смотреть'	'посмотреть, просмотреть'	1 037	36
күреп алу	'видеть'	'увидеть'	1 037	42
сизеп алу	'чувствовать'	'почувствовать'	314	31
шәйләп алу	'замечать'	'заметить'	148	15
каранып алу	'посмотреть по сторонам'	'посматривать по сторонам'	133	6
абайлап алу	'замечать, почувствовать'	'заметить'	92	10
ишетеп алу	'слышать'	'услышать'	62	13
тоеп алу	'чувствовать'	'почувствовать'	54	12
сизенеп алу	'чувствовать'	'почувствовать'	39	10
карангалап алу	'посматривать по сторонам'	'посматривать по сторонам'	15	4
Всего			**2 931**	

Как показывает таблица 2, основную массу частотных аналитических глаголов со значением восприятия, образованных по модели конверб на -п + *алу*, составляют глаголы, связанные с зрительным восприятием. Это конвербы от глаголов со значением 'видеть' и 'смотреть', а также их дериваты, образованные при помощи аффикса рефлексива (*каранып*) и раритива (*карангалап*). Кроме того, компонент зрительного восприятия представлен в семантической структуре еще двух конвербов – *шәйләп* и *абайлап* ('заметить'). Еще три конверба обозначают ощущение, не дифференцируя канал восприятия информации: *тоеп, сизеп* и *сизенеп*.

Кроме того, к данной же группе семантически близок аналитический глагол *куренеп алу* ('показаться'), который в выборке встречается 24 раза. В данной группе финитный глагол *алу* десемантизирован и выражает аспектуальные характеристики, связанные с результативностью действия по восприятию. Тем не менее, кроме грамматикализации глагола *алу*, здесь можно усмотреть освоение восприятий, вовлечение их в личное пространство человека.

Следующая значимая по частотности группировка образована конвербами приобретения, обозначающими взятие в руку, которые с глаголом *алу* образуют аналитические глаголы со значением приобретения. В нашей выборке данный класс представлен всего двумя конструкциями, которые, тем не менее, характеризуются высокой частотностью. См. таблицу 3.

Таблица 3. **Конструкции со значением 'поймать, схватить'**

Конструкция с конвербом	**Значение конверба**	**Значение конструкции**	**Употребление в корпусе**	**Количество разных форм финитного глагола в корпусе**
тотып алу	'держать'	'поймать'	793	52
элəктереп алу	'подцепить'	'схватить'	454	40
Всего			**1 247**	

Следующая группа – глаголы со значением завоевания, захвата (таблица 4.). Завоевание предполагает использование силы при преодолении сопротивления противника и перемещение на территорию, изначально принадлежавшую противнику, овладев этой территорией, значение ‘завоевать, завладеть’ имеется среди вторичных значений глагола *алу*. В целом глаголы данной группы используются при описании боевых действий.

Таблица 4. **Конструкции со значением 'завоевать, завладеть'**

Конструкция с конвербом	**Значение конверба**	**Значение конструкции**	**Употребление в корпусе**	**Количество разных форм финитного глагола в корпусе**
яулап алу	'завоевывать’	'завоевать'	509	78

басып алу	'захватывать'	'захватить'	367	62
биләп алу	"овладеть'	"завладеть'	353	23
Всего			**1 229**	

Следующую группу образуют глаголы со значением окружения (таблица 5). Здесь представлено 2 основных типа конструкций: 1) поместить, расположить что-л. вокруг кого-либо или чего-либо; 2) окружить что-либо с целью захвата.

Среди конвербов есть явно мультисубъектные, выражающие значение большой или неопределенной совокупности субъектов (*камап, сырып, чолгап*) или потенциально мультисубъектные (*уратып*).

Таблица 5. **Конструкции со значением 'окружить'**

Конструкция с конвербом	**Значение конверба**	**Значение**	**Употребление в корпусе**	**Количество разных форм финитного глагола в корпусе**
уратып алу	'окружать, обступать'	'окружить, обступить'	161	34
сырып алу	'облеплять'	'облепить'	132	12
урап алу	' окружать'	'окружить'	111	15
чолгап алу	'окружать'	'окружить'	93	23
әйләндереп алу	'окружать'	'окружить'	101	24
чорнап алу	'мотать, обвивать'	на'мотать, обвить'	74	14
камап алу	'брать в окружение, осаждать'	'взять в окружение, осадить'	59	19
Всего			**731**	

Следующая группа конструкций образована конвербами, обозначающую интеллектуальную деятельность (таблица 6). В данной группе глагол *алу* десемантизирован и выражает аспектуальные характеристики, связанные со спектром значений конверба.

Таблица 6. **Конструкции, обозначающие интеллектуальную деятельность**

Конструкция с конвербом	**Значение конверба**	**Значение**	**Употребление в корпусе**	**Количество разных форм финитного глагола в корпусе**
сайлап алу	'выбирать'	'выбрать, отобрать'	471	65
аңлап алу	'понимать'	'понять'	452	21
уйлап алу	'думать'	'подумать'	432	18
танып алу	'узнавать'	'узнать'	282	23
белеп алу	'знать'	'узнать'	177	29
төшенеп алу	'понимать, постигать'	'понять, постигнуть	87	13
чамалап алу	'прикидывать в уме'	'прикинуть в уме'	87	15
отып алу	'выучить, запомнить'	'выучить, запомнить'	47	22
уйланып алу	'задумываться'	'задуматься'	40	8
ятлап алу	'учить наизусть'	'выучить наизусть'	15	9
Всего			**2 803**	

Выделение частотных конвербных конструкций, образованных при помощи глагола *алу,* позволяет увидеть цепочку модификаций, через который проходит глагол *алу,* в зависимости от тематического класса конверба и значения конструкции. Наиболее часто встречаются конструкции, где конверб является глаголом восприятия, приобретения (включая завоевание), окружения и интеллектуальной деятельности.

Заключение

Тюркские конструкции с конвербами на -п представляют собой обширный неоднородный пласт образований с разной семантикой, степенью грамматикализации и лексикализации. Корпусные данные показывают, что наибольшее разнообразие в образо-

вании конвербных конструкций демонстрируют глаголы бытия, положения в пространстве, движения, а также отдельные глаголы других семантических классов.

Основное внимание в статье было уделено конструкциям со вторым компонентом *алу* ‘брать’, который во многих случаях подвергнут десементизации. Исследование показывает, что самую частотную группировку в данном классе представляют конструкции, образованные при помощи конвербов, обозначающих восприятие; среди них основную массу составляют глаголы, связанные с зрительным восприятием. Следующие по частотности группировки образованы конвербами со значением приобретения, завоевания и захвата, окружения и интеллектуальной деятельности.

Привлечение большого объема количественных данных позволяет уточнить полученные ранее результаты о месте конструкций в конвербами в лексической и грамматической системе тюркских языков и особенностях грамматикализаций глаголов. В дальнейшем планируется аналогичное исследование частотных конвербных конструкций с другим составом компонентов.

Благодарность

Исследование выполнено при финансовой поддержке РФФИ (проект № 15-07-09214).

ЛИТЕРАТУРА

1. Гращенко, В.П. (2015). Тюркские конвербы и сериализация: синтаксис, семантика, сериализация – Москва: Языки славянской культуры.

2. Гращенков, П. В. (2011). Подлежащее в деепричастных конструкциях тюркских языков. Филология и культура, № 26. С. 182–185.

3. Юлдашев, А. А. (1977). Соотношение деепричастных и личных форм глагола в тюрских языках. Москва: Наука, С. 270.

4. Гращенков, П. В. (2012). Тюркские конструкции со вспомогательным глаголом и деепричастием на-р (на материале языков кыпчакской группы). Урало-алтайские исследования, (1), С. 55–77.

5. Матушкина, Н. А. (2016). К вопросу о функциональных особенностях тюркских деепричастий (на материале якутского языка). Актуальные вопросы тюркологических исследований. К 180-летию кафедры тюркской филологии Санкт-Петербургского государственного университета. С. 65–72.

6. Татарский национальный корпус «Туган тел». URL: http://tugan-tel.tatar/.

УДК 811.512.145; 811’111’42

AUTOMATIC CLASSIFICATION OF TATAR TEXTS BY STYLES: PRELIMINARY RESULTS

A. Galieva, R. Gataullin
Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic Kazan, Russia
amgalieva@gmail.com, ramil.gata@gmail.com

This paper studies the issue of correlation of formal quantitative (mainly morphological) and stylistic features of Tatar texts. Currently computational linguistics deals with two basic aspects of texts – their structure and their content. Developing methods of the automatic classification of text style and register is one of topical subjects in NLP.

Functional stylistics is a traditional topic in Russian linguistics with functional styles described in various aspects. For the Russian language, lexical and morphological parameters of texts of different functional styles are studied in detail and presented in special literature; this data is the starting point for a number of applied researches on the automatic detection of styles. The available studies in Tatar stylistics are mainly focused on the analysis of linguistic imagery and individual styles of Tatar writers. Morphological parameters of functional styles lack special literature on Tatar stylistics.

Our research work included the following main stages: first we analyzed an array of Tatar texts and manually distinguished a set of morphological and other formalizing characteristics of the texts for each style; then we developed methods of automatic classification of Tatar texts by styles.

Preliminary results of methodology development for automatic classification of Tatar texts according to their morphological features are presented. The results of our research mark an important step for methodology development in detecting text styles of Tatar corpus collection automatically. In future, the results may be used for a wide range of tasks in Tatar texts processing.

Keywords: style; the Tatar language; stylistic characteristics of the text; automatic classification of texts by styles.

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ТАТАРСКИХ ТЕКСТОВ ПО СТИЛЯМ: ПОСТАНОВКА ЗАДАЧИ И ПЕРВЫЕ РЕЗУЛЬТАТЫ

А.М. Галиева, Р.Р. Гатауллин
Институт прикладной семиотики Академии наук Республики Татарстан, Казань, Россия
amgalieva@gmail.com, ramil.gata@gmail.com

В настоящее время одной из важных задач прикладной лингвистики является разработка методов автоматической классификации текстов по стилевым и жанровым характеристикам. 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 по автоматической классификации текстов по стилям. Исследования по стилистике татарского языка ориентированы главным образом на анализ языковых средств выразительности и особенностей языка того или иного писателя. Морфологические параметры функциональных стилей в специальных работах по татарской стилистике не описаны.

Поэтому наше исследование включает следующие основные этапы:

1) на основе анализа массива текстов устанавливаются частеречные и иные формализуемые признаки текстов, харакиерные для того или иного функционального стиля;

2) разрабатываются методы автоматической классификации татарских текстов по стилям.

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, автоматическая классификация текстов по стилям.

1. Введение

В настоящее время одной из важных задач прикладной лингвистики является создание методов автоматической классификации массива текстов по стилевым и жанровым характеристикам, и эта задача решается для ряда языков, в частности, английского (Burrows 1992, Kessler, Numberg & Schütze 1997, Stamatatos, Fakotakis & Kokkinakis 2000 и др.) и русского (Браславский 2000, Емашова, Мальковский 2007, Поспелова, Ягунова 2014, Ермакова, и др. 2014).

Функциональный стиль в российской лингвистике определяется как разновидность литературного языка, в которой язык выступает в той или иной социально значимой сфере общественно-речевой практики людей и особенности которой обусловлены особенностями общения в данной сфере. (ЛЭС 1990: 567). Сферы применения языка в значительной мере влияют на тематику, содержание и форму высказывания. Важную роль при выборе стиля и регистра текста играют типичные для той или иной сферы деятельности автор речи (частное лицо, официальное лицо, учреждение и т. п.), адресат сообщения (собеседник, массовая аудитория), тематика и целевая установка общения и др. Особенности употребления языка в той или иной сфере определяют выбор языковых средств, и каждый стиль характеризуется своим набором лексических, морфологических и синтаксических признаков, которые могут быть положены в основу автоматической классификации текстов по стилям.

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 по автоматической классификации текстов по стилям. Исследования по стилистике татарского языка ориентированы главным образом на анализ языковых средств выразительности и особенностей языка того или иного писателя. Морфологические параметры функциональных стилей в специальных работах по татарской стилистике не описаны.

2. Постановка задачи и определение признаков

Поэтому наше исследование строится следующим образом:

1) на основе анализа массива текстов устанавливаются морфологические и иные формализуемые признаки, значимые для текстов для каждого стиля;

2) разрабатываются методы автоматической классификации татарских текстов по стилям.

Первоначально нами вручную была подготовлена коллекция текстов основных функциональных стилей – официально-делового, публицистического (на примере новостных текстов), научного и художественного, и получены количественные данные по

четырем типам критериев для каждого стиля. 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	32	220 992
2	Публицистический	32	277 958
3	Научный	32	109 937
4	Художественный	32	55 614

Анализ основывался на измеряемых параметрах текстов (преимущественно морфологических), без подключения специальных словарей. Сформулированы ряд критериев для классификации текстов, а именно:

– морфологический критерий;

– критерий «Средняя длина слов и предложений в тексте»;

– критерий «Средняя длина аффиксальной цепочки имен сущ.»

– критерий «Средняя длина аффиксальной цепочки глаголов»

– 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 дает общее представление о распределении ряда параметров в зависимости от стилевой принадлежности текста.

Таблица 2. **Распределение грамматических параметров текстов**

Количество по отношению к общему количеству слов	**Стили**			
	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

Ряд критериев, которые считаются надежными с точки зрения автоматического выделения стилей текстов русского языка (Поспелов, Ягунова 2014), 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.

Как показывает анализ, 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 художествен-

ные; 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.

3. Эксперименты и результаты

Несмотря на существование работ, опирающихся на прямое сопоставление текста с уже классифицированными документами или псевдо-документом, представляющим собой жанр (Rutherford 2005), классическим подходом классификации текстов стало использование методов машинного обучения. Машинное обучение предполагает наличие обучающей и контрольной тестовой выборки.

Для проведения экспериментов подготовлены обучающая (104 текстов) и тестовая (24 текстов) коллекции вручную отобранных разножанровых текстов на татарском языке четырех основных стилей (официально-деловой, художественный, научный, публи-

цистический на примере новостных текстов). При этом каждому документу соответствует вектор признаков и стиль, к которой документ относится.

В качестве начальных признаков были выбраны следующие 25 признаков:

1. средняя длина предложения по тексту
2. средняя длина (по всему тексту) аффиксальной цепочки у глаголов
3. средняя длина (по всему тексту) аффиксальной цепочки у сущ
4. среднее количество уникальных словоформ на 1000 словоформ
5. среднее количество уникальных лемм на 1000 словоформ
6. N – кол-во сущ.
7. V – кол-во глаголов
8. POST – кол-во послелогов
9. PART – кол-во частиц
10. INTRJ – кол-во междометий
11. Num – кол-во числительных
12. MOD – кол-во модальных слов
13. Adj – кол-во прилагательных
14. CNJ – кол-во союзов
15. IMIT – кол-во звукоподражательных слов
16. PN – кол-во местоимений
17. Adv – кол-во наречий
18. PROP – кол-во имен собственных
19. NR – кол-во нераспознанных морфоанализатором словоформа
20. Type2 – кол-во знаков препинания (“,”, “ ”, “–”, “:”, “–”)
21. Type3 – кол-во скобок (“]”, “{“, “)”, “}”, “(“, “[“)
22. Type4 – кол-во кавычек (“””, “““, “””, “»”, “«“, “””)
23. Rus – кол-во слов из русского языка
24. Letter – кол-во отдельно стоящих букв
25. Latin – кол-во словоформ, написанных латиницей

Значения всех признаков были нормализованы и сведены к интервалу [0, 1], используя формулу с минимальным и максимальным значением каждого признака по обучающей выборке.

В качестве методов машинного обучения было принято решение проверить следующие:

1. SVM (Support Vector Machine, eng. – Метод опорных векторов, рус.)

2. MLP (Multilayer Perceptron, eng. – Многослойный персептрон (нейронная сеть), рус.)

3. Decision Tree (Дерево принятия решений, рус.)

Таблица 3. **Результаты обучения и работы моделей**

Метод	Средняя ошибка на обучающей выборке	Средняя ошибка на тестовой выборке
SVM	3.54 %	4.91 %
MLP	0.25 %	4.58 %
Decision Tree	0.0 %	14.49 %

Кроме обучения моделей, была предпринята попытка выявить признаки больше всего влияющие на определение стиля. Используя метод SVM, был получен список 10 таких признаков (упорядочены по убыванию степени влияния):

1) PN – кол-во местоимений;

2) N – кол-во имен сущ.;

3) PROP – кол-во имен собственных;

4) Adv – кол-во наречий;

5) INT – кол-во аффиксов, выражающих общий вопрос;

6) PCP_FUT – кол-во аффиксов причастия будущего времени (омонимичных формам будущего времени на *-ыр*);

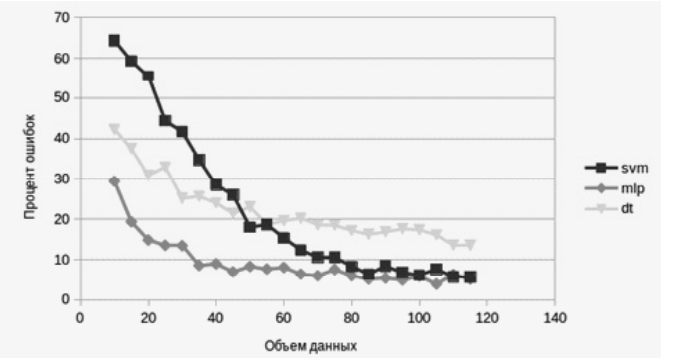


Рис. 1. Сравнительный график зависимостей точности моделей от объема данных

7) v_chain_len – средняя длина (по всему тексту) аффиксальной цепочки у сущ.;

8) V – кол-во глаголов;

9) PRES – кол-во морфем настоящего времени;

10) DESID – кол-во морфем “форма, выражающая значение намерения (дезидератив)”.

На рисунке 1 приведены зависимости точности моделей от объема данных обучения.

4. Заключение

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.

На тестовой выборке методы SVM и MLP показали примерно одинаковое количество ошибочного распознавания текстов по стилям (4,91% и 4,58% соответственно), количество ошибок при использовании дерева решений составил 14,49 %.

Нами была поставлена также задача по определению признаков, в наибольшей степени значимых для автоматического определение стилевой принадлежности текста. Используя метод SVM, был получен список 10 таких признаков.

Благодарность

Исследование выполнено при финансовой поддержке РФФИ (проект № 15-07-09214).

ЛИТЕРАТУРА

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

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

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

4. Ермакова, Л. М., Абашев, М. А., Никитин, Р. В., Ушаков, Р. И. (2014). Методы автоматической классификации текстов по функциональным стилям. ВЕСТНИК ПЕРМСКОГО УНИВЕРСИТЕТА. – № 4(27). С. 78–83.

5. Mason J.E., Shepherd M., Duffy J. (2009). An n-gram based approach to automatically identifying web page genre. System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on. IEEE, 2009. P. 1–10.

6. Лингвистический энциклопедический словарь / Главный редактор В. Н. Ярцева. – М.: «Советская энциклопедия», 1990. – 688 с.

7. Burrows, J. (1992). Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information. Literary and Linguistic Computing, 7(2), pp. 91–109.

8. Kessler, B., Numberg, G., Schütze, H. (1997). Automatic detection of text genre, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, p.32-38, July 07–12, 1997, Madrid, Spain

9. Lee, D. (2002). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. In Language and Computers, 42(1), 247–292.

10. Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against corpus-based methodologies. In English for Specific Purposes, 24(3), 321–332.

11. McEnery, T., Xiao, R., & Tono, Y. (2006). Corpus-based language studies: An advanced resource book. Taylor & Francis.

12. Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000, July). Text genre detection using common word frequencies. In Proceedings of the 18th conference on Computational Linguistics-Volume 2 (pp. 808-814). Association for Computational Linguistics.

13. Biber, D., & Conrad, S. (2009). Register, genre, and style. Cambridge University Press.

14. Rutherford, B. A. (2005). Genre Analysis of Corporate Annual Report Narratives A Corpus Linguistics–Based Approach. Journal of Business Communication, 42(4), 349–378.

УДК 81’33

MULTIFUNCTIONAL MULTILINGUAL INTERNET SERVICE BASED ON THE MODEL OF THE TURKIC MORPHEME

A. Gatiatullin
Institute of Applied Semiotics
of the Academy of Sciences of Tatarstan Republic
Kazan, Russia
ayrat.gatiatullin@gmail.com

The article describes a multifunctional multilingual Internet service, which helps to form a resource base for software products aimed to perform computer processing of Turkic languages, such as machine translation systems, information retrieval systems, electronic corpora markup and data extraction systems, etc. The service can be used as an information and reference system, which contains almost complete information on the Turkic language units, namely morphemes, and as a toolbox for turkology researchers. This toolkit can be used for comparative analysis of the Turkic language units and their proximity on the corresponding language levels of the Turkic languages.

The development and usage relevance of such a service for the Turkic languages is caused by the absence of an electronic database with full description of morphemes for any of the Turkic languages. Other reasons are the lack of effective research integration and coherence in the developments carried out by the Turkic language researchers in the field of computer linguistics. Thus, linguistic models and software modules, which by 70-80 percent are common to all Turkic languages, become duplicated in practice. This applies to structure and functionality development of electronic corpora and grammar analyzers, as well as search engines and machine translation systems.

The core of the multifunctional multilingual Internet service is a structural and functional model of the Turkic morpheme, which is a pragmatically oriented structural and functional description of morphological elements. It allows to perform complete enumeration of the Turkic morphemes with descriptions of their characteristics and the situations when they are used at all (namely, phonological, morphological, syntactic, and semantic) language levels. The architecture of the structural and functional model of the Turkic morpheme is also a hierarchical model consisting of many different submodels.

Keywords: Model of the Turkic morpheme; multifunctional multilingual Internet service, Turkic languages.

МНОГОФУНКЦИОНАЛЬНЫЙ МНОГОЯЗЫЧНЫЙ ИНТЕРНЕТ СЕРВИС НА БАЗЕ МОДЕЛИ ТЮРКСКОЙ МОРФЕМЫ

А.Р. Гатиатуллин
Институт прикладной семиотики
Академии наук Республики Татарстан
Казань, Россия
ayrat.gatiatullin@gmail.com

В статье описывается многофункциональный многоязычный Интернет сервис, основными задачами которого являются формирование ресурсной базы для программных продуктов, осуществляющих компьютерную обработку тюркских языков, таких как системы машинного перевода, информационно-поисковые системы, системы разметки электронных корпусов, извлечения данных и др. В набор функций сервиса также входят роли информационно-справочной системы, содержащей практически полную информацию о тюркских языковых единицах – морфемах, и инструментария для исследований ученых-тюркологов. Данный инструментарий может быть использован для проведения исследований по сравнительному анализу тюркских языковых единиц и степеней близости соответствующих языковых уровней тюркских языков.

Актуальность разработки и использования подобного многофункционального Интернет сервиса для тюркских языков определяется отсутствием электронной базы с полным описанием морфем ни для одного из тюркских языков, отсутствием реальной интеграции исследований и согласованности в разработках, осуществляемых исследователями тюркских языков в области компьютерной лингвистики. Таким образом, практически происходит дублирование лингвистических моделей и программных модулей их обработки, в основе своей на 70-80 и более процентов являющихся общими для всех тюркских языков, как при разработке структуры и функционала электронных корпусов языков, грамматических анализаторов, так и машин поиска и систем машинного перевода.

Ядром многофункционального многоязычного Интернет сервиса является структурно-функциональная модель тюркской морфемы, которая представляет собой прагматически-ориентированное структурно-функциональное описание элементов морфологии и позволяет осуществить полную «инвентаризацию» тюркских морфем с описанием характеристик и ситуаций их проявления на всех языковых уровнях (фонологическом, морфологическом, синтаксическом, семантическом). Архитектура самой структурно-функциональной модели тюркской морфемы также представляет собой иерархическую модель, состоящую из множества подмоделей.

Ключевые слова: Модель тюркской морфемы, многофункциональный многоязычный Интерннет сервис, Тюркские языки.

1. Введение

Перспективным и актуальным направлением в области компьютерной лингвистики является создание лингвистических баз данных, содержащих различные характеристики языковых единиц. Создание таких баз данных предполагает, обработку больших объемов фактического языкового материала, и глубокую разработку теоретико-методологических принципов, лежащих в основе построения таких моделей данных. Особую значимость имеют исследования, связанные с построением концептуальных и формальных моделей в группе тюркских языков, например, различные технологии сопоставительно-сравнительного исследования тюркских языков с использованием моделей тюркских морфем.

В статье описывается комплексный подход, в котором активные и пассивные лингвистические ресурсы для тюркских языков объединяются вместе в единый многофункциональный сервис. Пассивные – это лингвистические базы данных, а активные это программные модули для работ с этими базами данных. С одной стороны, эти программные модули используют информацию, представленную в базе данных для своей работы, а с другой сами служат дальнейшему наполнению лингвистической базы данных.

Актуальность разработки и использования подобного многофункционального Интернет сервиса для тюркских языков определяется следующими факторами:

– не имеется электронной базы с полным описанием морфем ни для одного из тюркских языков;

– отсутствует реальная интеграция исследований и согласованность в разработках, осуществляемых исследователями тюркских языков в области компьютерной лингвистики;

– практически происходит дублирование лингвистических моделей и программных модулей их обработки, в основе своей на 70–80 и более процентов являющихся общими для всех тюркских языков, как при разработке структуры и функционала электрон-

ных корпусов языков, грамматических анализаторов, так и машин поиска и систем машинного перевода. Очевидно, преодоление такого дублирования, объединение усилий на совместных разработках и даже обмен программными модулями позволят сэкономить финансы, направить усилия специалистов на нерешенные проблемы и достичь общего прорыва в области создания технологий для обработки тюркских языков, и даже создавать новые технологии обработки информации на основе лексико-грамматических особенностей тюркских языков;

– технологическим ядром данного многофункционального сервиса является модель тюркской морфемы, использование которой обусловлено исключительной значимостью морфологического языкового уровня при обработке естественно-языковых текстов. Особенно это актуально для языков агглютинативного типа с богатой морфологией, к которым относятся все языки тюркского семейства.

В настоящее время данный сервис находится на стадии реализации, в результате чего он должен стать ресурсной базой для программных продуктов, осуществляющих компьютерную обработку тюркских языков (систем машинного перевода, информационно-поисковых систем, системы разметки электронных корпусов, извлечения данных и др.); информационно-справочной системой, содержащей практически полную информацию о тюркских языковых единицах – морфемах; инструментарием для исследований ученых-тюркологов, в частности, для сравнительного анализа тюркских языковых единиц и степени близости различных языковых уровней тюркских языков.

При этом, этот сервис также должен включать модули, как для ручного, так и автоматического (полуавтоматического) наполнения базы данных. Аналогом таких систем можно считать программы машинного перевода накопительного типа, когда информация, получаемая в процессе работы, служит для дальнейшего улучшения качества работы программных модулей.

2. Архитектура многофункционального Интернет сервиса

Общая архитектура многофункционального многоязычного сервиса представлена на рис.1. На этом рисунке показано, что ядром этого многофункционального многоязычного Интернет

сервиса является структурно-функциональная модель тюркской морфемы (Рис.1.), которая представляет собой прагматически-ориентированное структурно-функциональное описание элементов морфологии и позволяет осуществить полную «инвентаризацию» тюркских морфем с описанием характеристик и ситуаций их проявления на всех языковых уровнях (фонологическом, морфологическом, синтаксическом, семантическом)(Сулейманов, Гатиатуллин, 2003).

В свою очередь архитектура самой структурно-функциональной модели тюркской морфемы представляет собой иерархическую модель, состоящую из множества подмоделей (Рис.2).

Многофункциональный многоязычный Интернет сервис содержит целый набор программных модулей для компьютерной обработки тюркских языков, реализованных с использованием модели морфем: морфологический анализатор, расширенный морфологический анализатор с аналитическими формами, семантико-синтаксический анализатор, подсистему сравнительного анализа близости тюркских языков, систему машинного перевода между тюркскими языками, спеллчеккер.

Рассмотрим основные программные модули сервиса, предствленные на рис.1.

Модуль администрирования. Модуль администрирования обеспечивает следующие функции:

– предоставление прав доступа пользователям сервиса и контроль за использованием базы данных (БД). Права доступа предоставляет администратор системы. Пользователям системы предоставляется один из режимов работы с базой данных: просмотра, редактирования или администрирования. Экспертам-лингвистам предоставляется доступ для редактирования информации только определенных языков или категорий.

– ведение журнала системы для контроля (мониторинг и фиксация) изменений, вносимых в базу данных и возможность разных режимов восстановления в случае необходимости.

– возможность обратной связи с разработчиками сервиса. Эта функция важна, если в сервисе не реализованы какие-либо параметры, необходимые с точки зрения пользователей - лингвистов-экспертов.

Модуль заполнения Базы данных. Этот модуль предоставляет возможность лингвистам-экспертам производить заполнение базы

данных в онлайн режиме, осуществляет проверку целостности и корректности заполнения базы данных. Каждый пользователь будет заполнять данные, войдя в систему с помощью своего аккаунта, поэтому будет возможно отслеживать изменения, вносимые каждым пользователем.

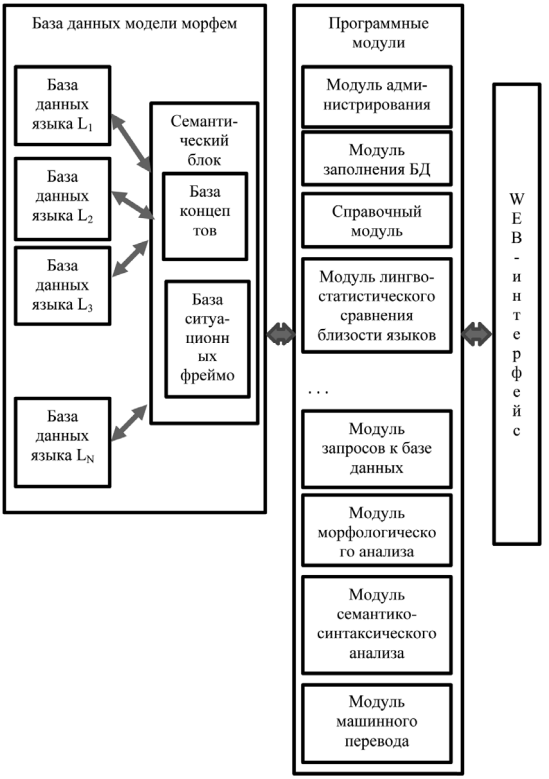


Рис. 1. Архитектура многофункционального сервиса

Модуль справочной системы (МСС). Модуль справочной системы предоставляет пользователю информацию с описанием возможностей сервиса и особенностей описания языковых единиц. Описание параметров языковых единиц, представленное в модели, является прагматически-ориентированным (в соответствии с решаемыми задачами) и не всегда совпадает с общепринятым описанием в академических грамматиках по тюркским языкам.

МСС включает таблицу соответствия обозначений (названий параметров, тэгов и др.)

Модуль запросов к базе данных. Модуль запросов к базе данных позволяет получать информацию о разных свойствах тюркских морфем. В частности, информация о свойствах синонимии, омонимии, антонимии морфем не представляется в базе данных в явном виде, а выдается как результат разных запросов к базе данных. Так омонимия между морфемами может быть разного вида: полная, когда омонимичны все алломорфы морфемы или частичная, когда омонимичны только некоторые алломорфы. Может быть омонимия между морфемами одного языка, а может быть омонимия между морфемами разных языков. Аналогично есть разные формы синонимии и антонимии. Вся эта информация может быть представлена с помощью разных параметров запроса к базе данных.

Модуль лингво-статистического сравнения близости языков. Модуль сравнительного анализа близости языков позволяет специалистам в области морфологической типологии заниматься вопросами сравнительного анализа, где в качестве предмета сравнения можно выбрать, например, сравнение грамматических категорий в различных языках, определение способов выражения грамматических категорий, установление синонимичных отношений аффиксальных морфем и служебных слов (послелогов, частиц и вспомогательных глаголов) и др.

Основными единицами измерения в морфологической типологии служат морфемы. Способ представления информации в модели морфем позволяет проводить лингво-статистические исследования для определения разных типов близости языков: морфологической, синтаксической, лексической.

Модуль морфологического анализа. Модуль морфологического анализа позволяет производить морфологический анализ, как в рамках словоформ, так и расширенный морфологический анализ с аннотированием аналитических форм. Этот модуль используется модулями семантико-синтаксического анализа и машинного перевода между тюркскими языками.

Модуль семантико-синтаксического анализа. Модуль семантико-синтаксического анализа производит анализ простых предложений тюркских языков, представляя их в виде ситуационного фрейма. Полученные таким образом ситуационные фреймы можно

использовать в системах машинного перевода, в информационно-поисковых системах и др.

Модуль машинного перевода. Модуль машинного перевода представляет собой программу для осуществления машинного перевода между тюркскими языками, информация о которых заполнена в модели морфем. Программа работает на основе Rule Based подхода и основными компонентами этого модуля являются модули морфологического и семантико-синтаксического анализа, представленных в этом сервисе (Gatiatullin, Bashirov, 2017).

Все модули в программе взаимосвязаны между собой и используют информацию, представленную в единой базе данных. Рассмотрим структуру базы данных многофункционального многоязычного Интернет сервиса.

3. Структура базы данных модели

База данных модели морфем состоит из следующих компонентов:

1. Семантический блок.
2. База идентификаторов (тэгов)
3. Базы языковых единиц для каждого тюркского языка Li.

База тэгов и семантический блок образуют языконезависимые блоки, которые являются общими для всех тюркских языков и служат для связи между собой баз данных каждого из тюркских языков.

База идентификаторов (тэгов) представляет собой базу тэгов разного уровня: морфологических, синтаксических, семантических. Создание единой системы индентификаторов (тэгов) является задачей семинара UniTurk, таким образом этот инструмент позволит автоматизировать работу участников семинара. Полученная база тэгов может стать единой базой для тюркских электронных корпусов.

Семантический блок состоит из базы концептов и базы ситуационных фреймов. База концептов представляет собой тезаурус, в котором концепты связаны между собой семантическими отношениями. Основное предназначение базы концептов в данной программе – это выстраивание системы взаимосвязей между корневыми морфемами разных языков. База ситуационных фреймов предназначена в первую очередь для представления значе-

ний аффиксальных морфем и служебных слов (Сулейманов и др, 2013).

База языковых единиц для каждого тюркского языка состоит из базы самих языковых единиц и базы правил сочетания этих языковых единиц. Базовой единицей здесь является морфема. Кроме морфемы в базе данных представлены языковые единицы, образуемые путем сочетания морфем с использованием комплекса правил. База морфем представляет собой базу морфем двух видов: корневых и аффиксальных.

Общая структура базы данных модели представлена на рис. 2.

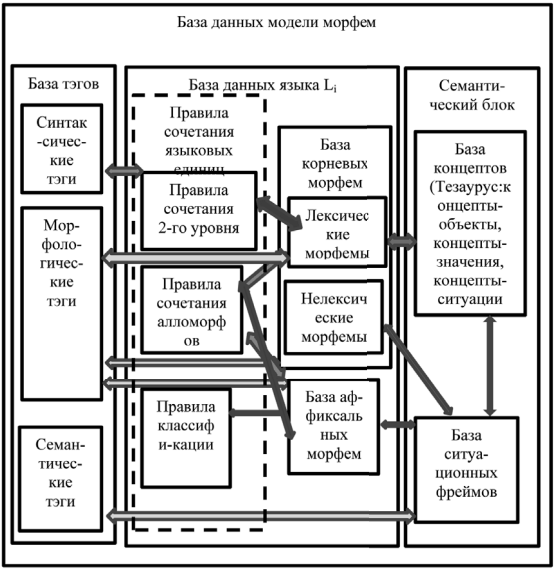


Рис.2. Структура базы данных модели

База правил состоит из правил сочетания разного уровня и правил классификации.

Примером правил классификации являются правила, по которым алломорфы объединяются в морфемы, так как, морфема это абстрактный класс, состоящий из группы алломорфов с общими признаками. Этот класс может состоять и из одного алломорфа.

Правила сочетания в базе данных делятся на правила сочетания разных типов:

– правила сочетания 1-го типа – сочетания алломорфов в одной словоформе, которые делятся на правила сочетания алломорфов двух видов: сочетания двух аффиксальных алломорфов и сочетания корневого и аффиксального алломорфа;

– правила сочетания уровня 2 типа – сочетания алломорфов в аналитических формах;

– правила сочетания 3-го уровня – сочетания словоформ в словосочетаниях, где каждая словоформа представляет собой комбинацию из корневых и аффиксальных морфем.

4. Заключение

В статье дано концептуальное описание многофункционального многоязычного Интернет сервиса, который в настоящее время находится в состоянии разработки. Идет процесс уточнения структуры, создания информационно-программной оболочки и заполнения базы данных для разных тюркских языков. Тот факт, что данная программа представляет собой Интернет сервис, позволяет уже на стадии разработки разным коллективам, работающим в сфере компьютерной обработки тюркских языков подключаться к процессу заполнения многоязычной базы данных и участвовать в уточнении, как самой структурно-функциональной модели, так и базы данных, с учетом языковых особенностей 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. Данный проект должен послужить интеграции усилий ученых-тюркологов для расширения базы данных описаниями различных тюркских языков, что обеспечит эффективное использование многофункционального многоязычного Интернет сервиса в качестве технологического инструментария 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.

3. Ayrat Gatiatullin, Artur Bashirov On the development of the translation system for Closely related languages on the basis of the Multifunctional model of the turkic morpheme // INTERACTIVE SYSTEMS: Problems of Human – Computer Interaction. – Collection of scientific papers. − Ulyanovsk: USTU, 2017. – pp. 132–135.

УДК 81’33

THE USE OF STATISTICAL METHODS IN DESIGN OF CYRILLIC-LATIN CONVERTER FOR TATAR LANGUAGE

A. Danilov[1], T. Ilyasov[2]
[1] Kazan Federal University. Kazan, Russia
[2] BARS Group, Kazan, Russia
tukai@yandex.ru

The relevance of the study is conditioned by the need of computer program development which allow to transliterate texts from Cyrillic graphics during the operation in non-Russian systems. During transliteration a text written with the use of a particular alphabet, is presented by the alphabet of another system. The correspondence of only two alphabet letters is usually taken into account. However, during the transliteration of living languages they try to take into account the sound aspect in order not to detach a word from its livesounding form. Thus, they transliterate not an alphabet, but the graphics adopted in this language system. Modern Tatar language has Cyrillic graphics. In this regard, the article shows the process of software development, which allows you to convert a text written in Tatar language using Cyrillic symbols into the Latin symbols. The principle of conversion is proposed, based on etymology. Original Tatar words are proposed to convert according to phonetic principle, and the borrowed words are proposed to convert according to transliteration rules. In order to determine the origin of a Tatar word it is proposed to use the following algorithms: digram analysis, combined analysis and search. An algorithmic model of conversion was developed. The software designed by the authors allows to transliterate native Tatar words nowadays.

Keywords: transliteration, Cyrillic, Latin, converter, Tatar, Russian.

СТАТИСТИЧЕСКИЕ МЕТОДЫ В РАЗРАБОТКЕ КОНВЕРТЕРА КИРИЛЛИЧЕСКОЙ ГРАФИКИ НА ЛАТИНСКУЮ ДЛЯ ТАТАРСКОГО ЯЗЫКА

А.В. Данилов[1], Т.А. Ильясов[2]
[1] Казанский Федеральный Университет, Казань, Россия
[2] Барс Груп, Казань, Россия
tukai@yandex.ru

В статье рассматриваются аспекты конвертации кириллической графики на латинскую графику для татарского языка. Авторы изучают применение различных статистических методов, необходимых для работы конвертера и анализируют скорость и точность работы алгоритмов конвертации. Был построен алгоритм и разработаны программные модули, позволяющие преобразовывать сообщения, написанные в татарской кириллице на татарскую

латиницу. В работе рассмотрены и разработаны программные инструменты, основанные на статистической обработке лингвистических данных: комбинированный биграммный анализ, наивная байесовская классификация и поиск на основе прямого перебора. Каждый из этих алгоритмов используется для определения этимологии слова, от которого зависит применение определенных правил конвертации с кириллицы на латиницу. Результатом исследования является разработанный программный продукт, который способен проводить процесс конвертации кириллической графики на латинскую для татарского языка. В дальнейшем авторами планируется усовершенствовать программный продукт и использовать в образовательной деятельности.

Ключевые слова: конвертер, кириллица, латиница, Наивный Байесовский Классификатор.

1. Введение

Письменность – это одно из средств коммуникации человека. Потребность в общении на расстоянии привела к возникновению письменности, она расширила круг общения и объединила людей не только в пространстве, но и во времени. Изобретение письменности привело к информационной революции, благодаря которой появились новые возможности для обмена и передачи информации Умение писать, как и умение читать, является одним из необходимых условий обучения. Несмотря на появление современных технологий передачи данных, письменная коммуникация не утратила своего значения и в наши дни.

За свою историю развития татарский язык несколько раз менял свою письменность. До 1927 года использовалось арабское письмо, с 1927 по 1939 года применялась латинская графика, с 1939 по настоящее время употребляется кириллица. В тридцатые годы 20-ого века после принятия алфавита «Яналиф» многие книги были перепечатаны с использованием латиницы. По мнению В.Г. Хакова (Хаков В.Г., 1993), практика тех лет показала, что использование латинской графики для татарской письменности позволило облегчить усвоение европейских языков, возможность чтения книг, написанных на тюркских языках.

Современная Республика Татарстан является многонациональным регионом Российской Федерации, где проживают представители различных национальностей и культур, по данным Всероссийской переписи населения 2010 года, большинство населения Республики Татарстан составляют татары (53,15%) и русские (39,1%) [2]. В соответствии с Законом «Об использовании татарского языка как государственного языка республики Татарстан»

государственными языками Республики Татарстан являются русский и татарский языки [3]. Таким образом, увеличивается количество текстов различных стилей (официально-деловых, научных, художественных, публицистических, разговорных), написанных на татарском языке. Повышается функциональность татарского языка, как языка хранения, обработки и передачи информации в области компьютерных технологий [Nevzorova O., Suleymanov D., Khakimov B, 2013; Danilov A, Salekhova L., 2015]. Существует большое количество транслитераторов слов с русского языка, однако практически нет разработанных транслитераторов для татарского языка. Этим обусловлена актуальность нашей разработки.

2. Методы

Цель исследования состояла в проектировании алгоритмической модели конвертации и разработки на её основе программного продукта, позволяющего «переводить» сообщения, записанные на татарском языке с помощью кириллицы, в латинскую графику.

В основу разработки модели транслитератора были положены правила использования латинской графики в татарском языке и перехода с кириллицы на латиницу, изложенные в Законе РТ «Об использовании татарского языка как государственного языка республики Татарстан» от 24.12.2012. В нем регулируется использование татарского языка в трех вариантах – на кириллице, латинице и арабице [3]. Также были использованы правила перевода и использования латинской графики, предложенные В. Хаковым в работе «Теленбелгән ил ачар: Латин графикасында уку һәм язу кунекмәләре» (1993).

При проектировании алгоритмической модели пришлось столкнуться с рядом трудностей. Одной из них является несоответствие количества гласных и согласных букв в двух алфавитах. В расширенной латинской графике 9 гласных букв: **a, ä, ü, u, o, i, e, ı, ö,** тогда как в татарском языке с использованием кириллической графики 13 букв: **а, ә, у, ү, о, ө, ы, е, э, и, я, ю, ё**. Важно отметить, что буквы «я», «ю», «е» представляют собой соединение двух звуков: я - [йа], ю - [йу], е - [йы], [йе]. То есть, в данных случаях нарушается закон сингармонизма.

В расширенной латинице для обозначения согласных используется 26 букв. В таблице 1 представлено соответствие между согласными буквами двух алфавитов.

Таблица 1. **Соответствие согласных букв, употребляемых в татарском языке (латиница-кириллица)**

Латиница	Кириллица	Латиница	Кириллица
Bb	Бб	Nn	Нн
Cc	Җҗ	Ņņ	Ңң
Çç	Чч	Pp	Пп
Dd	Дд	Rr	Рр
Ff	Фф	Ss	Сс
Gg	Гг	Şş	Шш
Ğğ	Гъ	Tt	Тт
Hh	Һһ	Vv	Вв
Jj	Ж ж	Ww	Уау
Kk	Кк	Xx	Хх
Qq	Къ	Yy	Йй
Ll	Лл	Zz	Зз
Mm	Мм	Şç	Щщ

Как любой живой язык, татарский язык развивается. Заимствования и неологизмы являются неотъемлемой составляющей процесса функционирования и исторического изменения языка, одним из основных путей пополнения его словарного запаса. При выработке основного принципа конвертации решено остановиться на принципе, основанном на этимологии. Для определения происхождения татарских слов выбраны следующие алгоритмы: биграммный анализ, комбинированный анализ и перебор. Предлагается применять различные правила транслитерации в зависимости от происхождения слова. Исконно-татарские слова конвертируются по фонетическому принципу, заимствованные слова – по правилам транслитерации.

К исконно-татарским словам применяется собственный набор правил, основанный преимущественно на фонетическом принципе (как слышу, так и пишу) – *тавык – tawıq*.

К заимствованным словам (арабско-персидского, русского и английского происхождения) применяется упрощенный набор правил, близкий к механической транслитерации (строгое соответствие символу на кириллице одному символу на латинице).

3. Результаты

Отобраны и модифицированы с учетом особенностей системы татарского языка такие алгоритмы определения происхождения слова, как биграммный анализ, комбинированный анализ, включающий в себя биграммный и морфемный анализ, а также метод перебора (brute-force). Обсудим преимущества и недостатки выбранных алгоритмов.

1) Биграммный анализ. Биграммой является идущая подряд пара лингвистических единиц (в нашем случае – пара букв). Некоторые биграммы в языке встречаются чаще, чем другие, следовательно, существует возможность определить происхождение слова с помощью статистических методов, проанализировав входящие в него биграммы. Подобные алгоритмы широко используются в веб-индустрии [Гречников Е.А., Гусев Г.Г., Кустарев А.А.,. Райгородский А.М., 2001; Attenberg, J., Suel, T., 2008; Benczur, A., Biro, I., Csalogany, K., and Sarlos, T., 2007]. В частности, в Интернет-браузерах для определения кодировки веб-страницы. Интернет-компания Яндекс применяет аналогичные методы для определения автоматически сгенерированного текста на веб-страницах. Была разработана специальная программа на выявление биграмм, которая анализировала тексты, состоящие исключительно из исконно-татарских слов.

Безусловно, алгоритм имеет свои недостатки. Это связано с тем, что татарский язык является агглютинативным, т.е. доминирующим принципом словообразования выступает агглютинация – «приклеивание» новых морфем к концу слова. Такой принцип словообразования создает трудности при определении происхождения слова.

Пример: Рассмотрим слово «Андрейныкыларгадыр». Слово заимствованное, и большинство биграмм в корне («Андрей») встречаются редко. Однако большую часть слова составляет суффиксальная часть, идущая после корня («-ныкыларгадыр»). Биграммы, представленные в суффиксальных морфемах, очень часто используются в татарском языке. При использовании вышеописанного алгоритма биграммного анализа, программа определит это слово как родное. Поэтому при анализе необходимо выделять корневую часть слова, после чего анализировать лишь корень.

2) Комбинированный анализ (биграммный + морфемный). Принцип работы этого алгоритма заключается в том, что слово

перед проверкой делится на коренную часть и суффиксальную, суффиксальная часть латинизируется по правилам родного языка, а коренная часть проверяется по биграммам. Для выделения корня используется специальное программное обеспечение –морфоанализатор.

3) Метод перебора или Brute-force. Brute-force – прием в программировании, при котором исходное слово проверяется на соответствие из заранее подготовленного списка. При использовании данного алгоритма необходимо создать множество из исконно татарских слов, каждое слово проверяется на вхождение в составленный список.

Недостаток алгоритма в том, что он будет работать только в том случае, если слово входит в множество. Если его нет в списке, программа определит исходное слово как неродное. Использование данного алгоритма приводит к увеличению времени конвертации слов.

В дальнейшем, при совершенствовании обобщенного алгоритма для определения происхождения слова, необходимо ориентироваться также на такие факторы, как скорость работы алгоритма и точность. По нашему мнению, среди предложенных алгоритмов самым оптимальным на данный момент является комбинированный метод, так как он позволяет максимально точно конвертировать слова без потери скорости работы программы.

Разработана алгоритмическая модель перевода татарского слова с кириллической графики на латиницу (Рис.1). Данная алгоритмическая модель легла в основу разработки компьютерной программы.

Алгоритм работает следующим образом:

Input – пользователь программы вводит слово на татарском языке, записанное на кириллице.

Check – проверка этимологии слова (родное или заимствованное).

Check (word)=own – блок условного выбора, в зависимости от результата проверки, слово конвертируется по определенному набору правил:

1 случай **(Latinise_Own)** – если слово исконно-татарское, то оно конвертируется посимвольно. Данный блок содержит в себе комплекс процедур и функций, предназначенных для конвертации. Все процедуры построены с учетом правил перевода с кириллицы на латиницу.

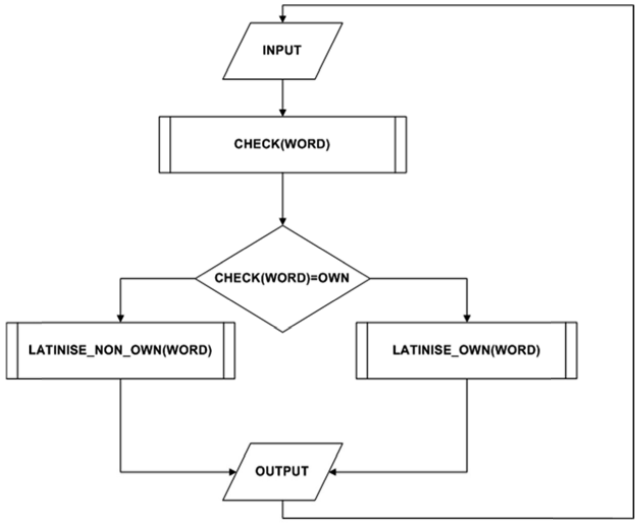


Рис.1. Алгоритм перевода сообщения, записанного на татарском языке с помощью кириллической графики, на латиницу

2 случай **(Latinise_non_Own)** – если слово заимствованное, то оно посимвольно конвертируется согласно правилам конвертации для заимствованных слов. Данный блок содержит комплекс процедур и функций, реализующий процесс механической транслитерации.

Примечание! Печатные знаки, не являющиеся буквами (знаки препинания, цифры) не обрабатываются.

Блок Output – выводится сообщение, записанное на латинице. Далее считывается следующее слово (возврат к блоку **Input**).

Разработанная алгоритмическая модель и принципы легли в основу создания компьютерной программы. Для её разработки в 2015 году была создана группа, состоящая из сотрудников кафедры образовательных технологий и информационных систем в филологии и студентов и магистров Института филологии и межкультурной коммуникации КФУ. В команду входят программисты и специалисты в области татарского языкознания. Приложение реализуется с помощью интегрированной среды разработки (IDE) Embarcadero Delphi 2009. На данный момент разработан программный продукт, который позволяет конвертировать

в латиницу исконно-татарские слова. Пример работы программы представлен на рис.2.

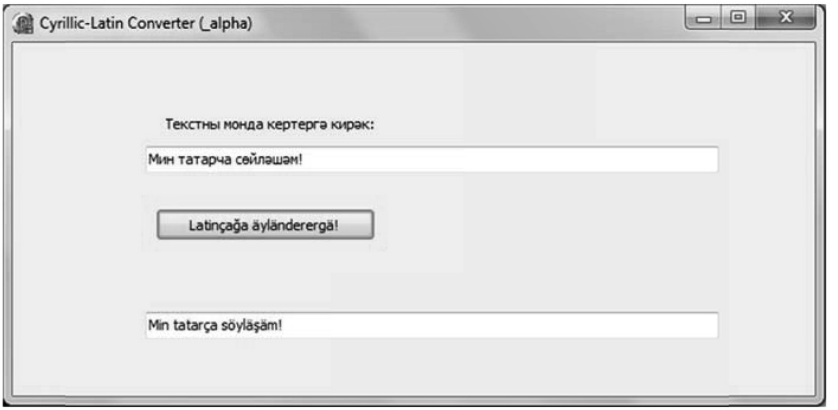


Рис.2. Демонстрация работы конвертера.

4. Обсуждение

На данный момент задача определения этимологии слова татарского языка окончательно не решена, будет разрабатываться более совершенный алгоритм проверки. Создан программный продукт, позволяющий проводить процесс конвертации. Помимо программного обеспечения для операционной системы Windows, планируется разработка мобильного приложения для мобильных операционных систем (iOS, Android или Windows Phone/Mobile), Интернет-ресурса, предоставляющего возможность конвертации в режиме он-лайн и надстройки для распространенного текстового процессора Microsoft Office Word.

Разработка данного программного продукта откроет новые возможности по использованию татарского языка в сфере письменной коммуникации, ИТ-индустрии и образовании. Использование разработанного программного обеспечения, наряду с другими решениями [Danilov, A., Salekhova, L., 2015; Zaripova, R., Salekhova, L., Tuktamyshov, N., Salakhov, R. 2014; Nevzorova, O., Suleymanov, D., Gilmullin, R., Gatiatullin, A., Khakimov, B., 2013; Fatkhullova K., Zamaletdinov R., Yusupova A., 2013], будет способствовать повышению уровня информационной культуры татаро-

язычных пользователей Интернета и информационных технологий, транслитерация татаро-язычных текстов даст возможность отечественным зарубежным ученым совместно проводить совместные исследования.

ЛИТЕРАТУРА

1. Хаков В.Х. Телен белгән ил ачар: Латин графикасында уку һәм язу күнекмәләре [Текст]/ Хаков В.И. – Казань: Издательство Мәгариф,1993 – 140 с.

2. 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)

3. Об использовании татарского языка как государственного языка республики Татарстан [Электронный ресурс]: закон Республики Татарстан от 12 января 2013 года №1-ЗРТ – Режим доступа: http://mon.tatarstan.ru/rus/file/pub/pub_227812.pdf

4. Nevzorova, O., Suleymanov, D., Gilmullin, R., Gatiatullin, A., Khakimov, B. Tatar National Corpus "Tugantel": structure and features of grammatical mark-up // Procedia - Social and Behavioral Sciences. Vol. 95. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013) // Ed. Chelo Vargas-Sierra, pp. 68–74.

5. Danilov, A., Salekhova, L. Design of Virtual Keyboard for Tatar-Speaking Users on the Basis of the Mobile Operating System Android// International Journal of Soft Computing, №10 (5): p. 348–352, 2015.

6. Гречников Е.А., Гусев Г.Г., Кустарев А.А.,. Райгородский А.М. Поиск неестественных текстов // Proceedings of VLDB-2001, 2001, 306–308.

7. Attenberg, J., Suel, T. Cleaning search results using term distance features // Proceedings of AIRWeb-2008, p. 21–24.

8. Benczur, A., Biro, I., Csalogany, K., and Sarlos, T.Web spam detection via commercial intent analysis. //Proceedings of AIRWeb-2007, NewYork, USA, 2007, p. 89–92.

9. Zaripova, R., Salekhova, L., Tuktamyshov, N., Salakhov, R. Definition of development level of communicative features of mathematical speech of bilingual students //Life Science Journal. – 2014. – . №11 (8). – URL: http://www.lifesciencesite.com/lsj/life1108/ (дата обращения : 23.03.2016)

10. Fatkhullova K.S., Zamaletdinov R.R., Yusupova A.S. Information-Communicative Devices for Tatar Language Teaching // World Applied Sciences Journal, 2013, Volume 26, Issue 1, pp. 103–107.

УДК 004.912

ON THE DEVELOPMENT OF THE COMPONENTS OF THE NATIONAL CORPORA OF THE CHUVASH LANGUAGE

V. Zheltov, P. Zheltov,
Federal state budget educational institution of higher education «Chuvash State University named after I.N. Ulyanov», Cheboksary
e-mail: chnk@mail.ru

The subject of the research is the peculiarities of the development of a national corpora of the Chuvash language. The aim of this work is the software implementation of national corpus of the Chuvash language, as well as the analysis and the identification of the most similar existing models and algorithms. Were revealed specific features arising from the design and development of the national corpora of Chuvash language related to the creation of the dictionary of morphemes, to the processing of the input information, to the character encoding, to the representation of the output information and to the search parameters. Are being considered the problems and prospects of the development of the system of national corpora of the Chuvash language. The practical result of the program is presented.

Keywords: national corpora, morphological analysis, syntax analysis, Chuvash language, linguistic corpora.

О РАЗРАБОТКЕ КОМПОНЕНТОВ НАЦИОНАЛЬНОГО КОРПУСА ЧУВАШСКОГО ЯЗЫКА

В.П. Желтов, П.В. Желтов,
Федеральное государственное бюджетное образовательное учреждение высшего образования «Чувашский государственный университет имени И. Н. Ульянова», Чебоксары
e-mail: chnk@mail.ru

Предметом исследования являются особенности разработки национального корпуса чувашского языка. Целью работы является программная реализация национального корпуса чувашского языка, анализ и выявление наиболее близких существующих аналогов и алгоритмов их работы. Выявлены специфические особенности, возникающие при проектировании и разработке национального корпуса чувашского языка, связанные с созданием словаря морфем, обработкой входной информации, кодировкой

символов, представлением выходной информации, параметрами поиска. Рассмотрены проблемы и перспективы разработки системы национального корпуса чувашского языка. Приведен практический результат работы программы.

Ключевые слова: национальный корпус, морфологический анализ, синтаксиыеский анализ, семантический анализ, чувашский язык, лингвистический корпус.

В настоящее время большую актуальность приобретает создание национального корпуса чувашского языка. Такие корпуса созданы уже для многих языков Российской Федерации (русский, татарский, башкирский, калмыцкий, марийский, мордовский, удмуртский, коми, хакасский) и являются огромными структурированными хранилищами текстов с возможностью быстрого поиска на нескольких уровнях языка: морфемном, морфологическом, синтаксическом, текстовом и семантическом [1-3].

Эти тексты доступны любому пользователю через Интернет. Для этого разрабатываются специализированные сайты (или наборы связанных сайтов – порталы), с необходимым функционалом, позволяющие пользователям проводить поиск необходимых языковых единиц (слов и их парадигм) и их частотный анализ, автоматическую разметку (по результатам морфологического и синтаксического анализа) пользовательских текстов, а также ряд других действий (например, построение картины мира текста или концептуального облака для какого-либо слова – набора слов связанных синтаксическими отношениями с искомым словом). В данной работе представлена структура разработываемгого портала для национального корпуса чувашского языка, доступного по адресу http://yuman21.ru, и его поисковика, доступного через вкладку «Поиск» на указанном портале.

Структура портала национального корпуса чувашского языка включает в себя следующие компоненты, реализованные ввиде вкладок: главная страница, с кратким описанием проекта; «Поисковик» – поисковик национального корпуса чувашского языка; «Морфологический анализатор» – морфологический анализатор национального корпуса чувашского языка; «Синтаксический анализатор» – синтаксический анализатор национального корпуса чувашского языка; «Семантический анализатор» – семан-

тический анализатор национального корпуса чувашского языка; «Тезаурус» – тезаурус анализатор национального корпуса чувашского языка, на основе известного 17-ти томного словаря (тезауруса) Н.И. Ашмарина и на основе [4-6].

Вкладка «Поисковик», реализующая поиск в текстовой базе портала, содержит следующие элементы пользовательского интерфейса:

1) текстовое поле для ввода искомого слова на чувашском языке с кнопками для ввода специальных чувашских символов (ă,ĕ,ӳ,ç);

2) кнопку «Поиск», расположенную рядом;

3) раскрывающийся каталог текстов (ввиде древовидного контейнера), расположенный ниже на странице; в случае свернутого каталога или отсутствия выбора какого-либо из его элементов поиск производится по всему корпусу текстов; первое разделение каталога производится по жанрам (словари, художественный произведения, публицистика, пресса, энциклопедии, справочные издания, научная литература, материалы конференций, учебная литература, детская литература, фольклор), второе – по конкретным произведениям в рамках выбранного жанра; реализована возможность группировки произведений в рамках выбранного жанра по фио (псевдониму и имени автора), а также по названию произведения;

4) текстовое поле, содержащее абзац, в котором найдено искомое слово, с указанием на выходные данные произведения (автора, название произведения, год издания, место издания, редакцию) с кнопками навигации (“переход к следующему абзацу” и «возврат к предыдущему абзацу») и кнопкой и чекбоксом “сохранить в файл” (вызывает окно выбора текстового файла, в который будет добавлен найденный абзац, если имя такового не было задано, а чекбокс был отмечен, если же имя файла было уже задано, то при переходе к следующему найденному в корпусе абзацу, содержащему искомое слово, предыдущий абзац добавляется к указанному файлу автоматически).

Морфологический анализатор национального корпуса чувашского языка реализован на основе теории, изложенной в [7-8] и включает в себя следующие компоненты:

1. dll-библиотеку, содержащую морфологический анализатор и реализованную на языке C# с использованием платформы .NET Microsoft.

2. Вкладку «Морфологический анализатор» на сайте национального корпуса чувашского языка (http://yuman21.ru/Morf).

Вкладка ««Морфологический анализатор» содержит следующие элементы пользовательского интерфейса:

1) текстовое поле для ввода анализируемого слова на чувашском языке с кнопками для ввода специальных чувашских символов (ӑ,ӗ,ӳ,ҫ);

2) кнопку «Анализ», расположенную рядом;

Морфологический анализатор, реализованный офлайн, содержит окно отображения результатов морфологического анализа как для отдельно введеного слова или фразы, так и возможность вставлять текст для анализа в специальном поле текствого ввода.

База данных морфологического анализатора состоит из словаря основ чувашского языка и базы аффиксов.

Словарь представляет собой текстовый файл, в котором слова представлены следующим образом: слово, часть речи, информация об источнике [9]. В нем собрано более тридцати одной тысячи слов чувашского языка.

Аффиксы играют решающую роль и несут основную словообразовательную нагрузку. Их в чувашском языке около 170 [1]. Исходная база аффиксов морфологического анализатора имеет схожую структуру со словарем.

Процесс морфологического анализа разделен на два этапа. На первом этапе слово в исходной форме ищется в словаре основ. Грамматические характеристики в данном случае определяются по умолчанию в зависимости от части речи. На втором этапе производится непосредственный анализ слова, разбиение его на пары «корень-аффиксы» и извлечение характеристик. Оба этапа возвращают произвольное количество омонимов в зависимости от найденных совпадений. При отсутствии совпадений слово возвращается с «неопределенными» характеристиками.

Прямой поиск входного слова в словаре основ выполняет функция *SearchInDictionaries*.

Второй этап реализован в функции *DetermineOnYourOwn*. Исходное слово подвергается пошаговому разбиению на аффиксы и происходит исследование на основе выделения его компонент (основы и аффиксов) на предмет принадлежности к какой-либо части речи.

Основной задачей синтаксического блока лингвистического процессора (ЛП) является преобразование морфологической структуры (МорфС) предложения, поступающей с выхода морфологического блока, в синтаксическую структуру (СинтС).

Так как МорфС предложения состоит из МорфС отдельных словоформ, то переход от МорфС предложения к его СинтС осуществляется путем установления синтаксических связей между МорфС слов и между самими связями. При этом МорфС отдельных словоформ служат для установления этих связей или, как принято их называть в компьютерной лингвистике, отношений. Поэтому от того, какую модель данных для представления СинтС мы примем за основу, будет зависеть эффективность работы синтаксического блока ЛП [1].

Для представления СинтС предложения в данном ЛП был выбран подход, основанный на методе, предложенном Л. Теньером и А.М. Пешковским, используемом в большинстве современных ЛП [2].

Согласно данному подходу, СинтС предложения называется размеченное дерево зависимостей, при котором: множество его узлов образует имена всех лексем, входящих в предложение; каждая дуга помечена именем какого-либо синтаксического отношения, описывающего синтаксическую связь между компонентами предложения; все дуги ориентированы от родительского узла к дочернему [3].

Программная часть синтаксического анализатора включает в себя следующие компоненты: вкладку «Синтаксический анализатор» на сайте ресурса компьютерной лингвистики чувашского языка, размещенного на специально выделенном для этого домене (http://yuman21.ru/Synt) и реализующую синтаксический анализатор библлиотеку.

Вкладка «Синтаксический анализатор» будет содержать следующие элементы пользовательского интерфейса:

1) текстовое поле для ввода анализируемого предложения на чувашском языке с кнопками для ввода специальных чувашских символов (ӑ,ӗ,ӳ,ҫ);

2) кнопку «Анализ», расположенную рядом;

3) окно отображения синтаксических отношений и синтаксиче кой структуры предложения (ввиде дерева).

В настоящий момент также не существует никаких аналогов электронного тезауруса чувашского языка, поэтому разработка мультимедийного тезауруса чувашского языка приобретет большую актуальность [4].

Программная часть мультимедийного тезауруса разрабатываемого для национального корпуса включает в себя следующие компоненты:

1. Список наиболее употребительных слов чувашского языка с их переводами на русский язык и со статьями из 17-ти томного словаря Н.И. Ашмарина (введено пока около 500 статей из томов 1-3).

2. Вкладку «Тезаурус» на сайте национального корпуса чувашского языка с переадресацией на сайт тезауруса (https://thesaurus21.herokuapp.com), который реализуется на основании готовой платформы создания облачных приложений (https://www.heroku.com).

3. Панель администрирования, позволяющая подгружать в каталог/список слов тезауруса новые слова со связанными с ними статьями из словарей и индексировать их.

4. Модуль индексации словарных статей.

5. Специальное веб-приложение, которое позволяет зарегистрированным пользователям просматривать словарь и тезаурус чувашского языка, добавлять и удалять лексико-семантические и словообразовательные пометы, добавлять и редактировать детализацию слов: перевод на русский язык, синонимы, аллоформы, статьи из словаря чувашского языка Н. И. Ашмарина.

Разработанное для управления содержимым веб-приложение работает в операционной системах семейства *Windows*, *Linux, MacOS* и мобильных операционных системах *Windows Phone*, *Andriod* и *iOS* и было выполнено в среде разработки *Visual Studio Code* с использованием фреймворка *Django*. Результатами работы программы является выводы заполненной базы данных в формате СУБД *PostgreSQL* и текстового файла в формате *JSON*.

В настоящее время также разрабатывается семантический анализатор чувашского языка (доступен для тестирования по адресу http://yuman21.ru/Semant).

Семантический анализатор в настоящее время способен находить слова, которые связаны с искомым синтаксическими отношениями. Таким образом, можно находить для исследуемых слов

определения и эпитеты, а также глаголы, которые употребляются вместе с ними, т.е. в полуавтоматическом режиме строить некую упрощенную картину мира или облако концептов для исследуемого текста.

Благодарность

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 15-04-00532.

ЛИТЕРАТУРА

1. Желтов П.В. Лингвистические процессоры. Формальные модели и методы: теория и практика: Монография. – Чебоксары: Изд-во Чуваш. ун-та, 2006. – 208 с.

2. Желтов П.В. Формальные методы и модели в сравнительно-сопоставительном языкознании: Монография. – Чебоксары: Изд-во Чуваш. ун-та, 2006. – 252 с

3. Желтов П.В. Сравнительные исследования морфем чувашского языка. – Чебоксары: Изд-во Чуваш. ун-та, 2013. – 166 с.

4. Желтов П.В. Создание национального корпуса чувашского языка: проблемы и перспективы // Современные проблемы науки и образования. – 2015. -№ 1–1. С. 338.

5. Желтов П.В. Лингвистические процессоры в системах искусственного интеллекта: Монография. – Чебоксары: Изд-во Чуваш. ун-та, 2007. – 100 с.

6. Желтов П.В., Губанов А.Р., Желтов В.П. Морфологический стандарт национального корпуса чувашского языка // Современные проблемы науки и образования. – 2015. № 2. С.180.

7. Желтов П.В. Исследования исторического развития чувашского языка. – Чебоксары : Изд-во Чуваш. ун-та, 2013. – 166 с.

8. Желтов П.В. Компьютерное моделирование многоагентных систем. – Чебоксары: Изд-во Чуваш. ун-та, 2008. – 112 с.

9. Желтов П.В. Модели и методы обработки символьной информации на сетях Петри. – Чебоксары: Издательство Чуваш. ун-та, 2012. – 108 с.

10. Желтов П.В. Моделирование многоагентных систем сетями Петри: Монография. – Чебоксары: Изд-во Чуваш. ун-та, 2008. – 108 с..

11. Zheltov P, Fomin E., Luutonen J. Reverse Dictionary of Chuvash. Societe Finno-ougrienne, Helsinki. 2009. – 344 p.

УДК 004.912

THE PARSER OF THE NATIONAL CORPORA OF THE CHUVASH LANGUAGE

V. Zheltov, P. Zheltov
Federal state budget educational institution of higher education «Chuvash State University named after I.N. Ulyanov», Cheboksary
e-mail: chnk@mail.ru

The article presents a description of the syntax analyzer (SA) of the national corpora of Chuvash language. The syntax analysis typically consists of two steps: the segmentation of sentences and the linking of words. Segmentation allocates simple sentences in complex ones. The next step is to establish links between words in the selected segments. The parser was realized as a library consisting of ten classes. The results of the work of the library can be widely applied in the systems of automatic processing of texts in Chuvash, the library can be part of a linguistic processor of Chuvash language. Have been revealed specific features arising from the design and development of the syntactic parser of the national corpora of the Chuvash language.

Keywords: national corpora, morphological analysis, syntax analysis, Chuvash language, linguistic corpora.

СИНТАКСИЧЕСКИЙ АНАЛИЗАТОР НАЦИОНАЛЬНОГО КОРПУСА ЧУВАШСКОГО ЯЗЫКА

В.П. Желтов, П.В. Желтов
Федеральное государственное бюджетное образовательное учреждение высшего образования «Чувашский государственный университет имени И. Н. Ульянова», Чебоксары
e-mail: chnk@mail.ru

В статье представлено описание синтаксического анализатора (СА) национального корпуса чувашского языка. Синтаксический анализ разделен на два этапа: сегментацию предложения и установление связей между словами. Сегментация выделяет простые предложения в составе сложного. На следующем этапе происходит установление связей между словами в выделенных сегментах. Программа реализована как библиотека из десяти классов. Результаты работы библиотеки могут широко применяться в системах автоматической обработки чувашских текстов, быть составной частью лингвистического процессора чувашского языка и систем машинного перевода.

Выявлены специфические особенности, возникающие при проектировании и разработке синтаксического анализатора национального корпуса чувашского языка.

Ключевые слова: национальный корпус, морфологический анализ, синтаксический анализ, чувашский язык, лингвистический корпус.

В статье представлено описание синтаксического анализатора (СА) национального корпуса чувашского языка. Синтаксический анализатор реализован ввиде *dll* и базы знаний [1].

Основной задачей синтаксического анализа является, как известно, преобразование морфологической структуры (МорфС) предложения, поступающей с выхода морфологического блока, в синтаксическую структуру (СинтС) [2]. Так как МорфС предложения состоит из МорфС отдельных словоформ, то переход от МорфС предложения к его СинтС осуществляется путем установления синтаксических связей между МорфС слов и между самими связями. При этом морфологические характеристики отдельных словоформ служат для установления этих связей или, как принято их называть в компьютерной лингвистике, отношений. Поэтому от того, какую модель данных для представления СинтС мы примем за основу, будет зависеть эффективность работы синтаксического блока. Для представления СинтС предложения выбран подход, основанный на методе, предложенном Л. Теньером и А.М. Пешковским, используемом во многих современных лингвистических процессорах (ЛП) [3]. Согласно данному подходу, СинтС предложения называется размеченное дерево зависимостей, при котором: 1) множество его узлов образует имена всех лексем, входящих в предложение; 2) каждая дуга помечена именем какого-либо синтаксического отношения, описывающего синтаксическую связь между компонентами предложения; 3) все дуги ориентированы от родительского узла к дочернему.

Программа синтаксического анализа состоит из двух компонентов: сегментации предложения и установления связей между словами. Компоненты работают параллельно или последовательно, в зависимости от архитектуры синтаксического модуля.

К началу синтаксического анализа весь текст представляется в виде последовательности характеристик к словоформам, так что алгоритм синтаксического анализа имеет дело не со словоформами, а лишь с соответствующими характеристиками.

Сегментация выделяет простые предложения в составе сложного. В любом простом предложении могут быть причастные или деепричастные обороты, придаточные предложения, которые, в свою очередь, тоже могут быть «разбиты» другими оборотами. Существуют примеры, когда части цельного высказывания находятся на значительном расстоянии друг от друга, а глубина вложения небольших предложений теоретически не ограничена.

На следующем этапе происходит установление связей между словами в выделенных сегментах. На этом этапе появляется проблема морфологической омонимии, то есть неоднозначности. Морфологическая омонимия возникает, когда одна и та же форма может выражать разные морфологические значения. Пример: форма «*çурĕ*» может быть глаголом прошедшего времени ‘*помыл*’ и глаголом будущего времени ‘*порвет*’. Явление морфологической омонимии весьма негативно отражается на скорости работы программы синтаксического анализа. На «длинных» предложениях количество комбинаторных вариантов иногда достигает нескольких сотен, поэтому используются разного рода математические и лингвистические ухищрения, позволяющие избежать анализа всех комбинаторно возможных вариантов [4].

База знаний синтаксического анализа содержит перечислимые типы; правила анализа; метаправила управления алгоритмом анализа, а также структуру рабочей модели (содержащую элементы: текст, предложение, вариант разбора предложения, простое предложение, слово и словосочетание, отношения).

Приведем пример СинтС следующего предложения: «*Хастар каччăсем урамра чупаççĕ*» (рис. 1).

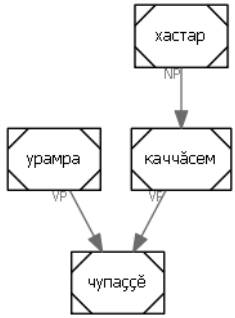


Рис.1.

Фактически такое представление нужно лишь для пользователей при показе СинтС. Для внутреннего представления и анализа удобно и выгодно использовать не само дерево, а лишь информацию, которая однозначно его представляет и которую можно получить непосредственно при анализе предложения. Это набор входящих в него синтаксических отношений (СинтО) и набор узлов с пометкой о них. В разработанном СА и принята такая структура данных [5-7].

Структура классов синтаксического анализатора

Проект библиотеки синтаксического анализатора состоит из 10 классов (рис. 2).

1. SyntParser.cs – основной класс. Содержит поля и методы для работы с библиотекой.

2. MainWork.cs – класс, который отвечает за определение характеристик предложения.

3. AttributesDeterminer.cs – класс, который предоставляет основные методы по анализу предложений. Объединяет все полученные данные под одним объектом.

4. AttrDeterHelper.cs – вспомогательный класс класса *AttributesDeterminer*.

5. Combinator.cs – класс, который на основе входного предложения формирует всевозможные вариации предложения.

6. Helper.cs – классе, в котором собраны общие вспомогательные функции.

7. BorderingWork.cs – класс, который отвечает за определение границ простых предложений в сложном.

8. Relations.cs – класс, отвечающий за определение отношений (связей) между словами.

9. GraphMaker.cs – класс, который формирует граф на основе выявленных отношений

10. SyntConstants.cs – класс синтактических констант, используемых в проекте.

Разработанная библиотека выполняет следующие действия.

1. Определение подлежащих и сказуемых предложения.
2. Определение границ простых предложений.
3. Определение связей между словосочетаниями.
4. Визуализация выявленных связей.

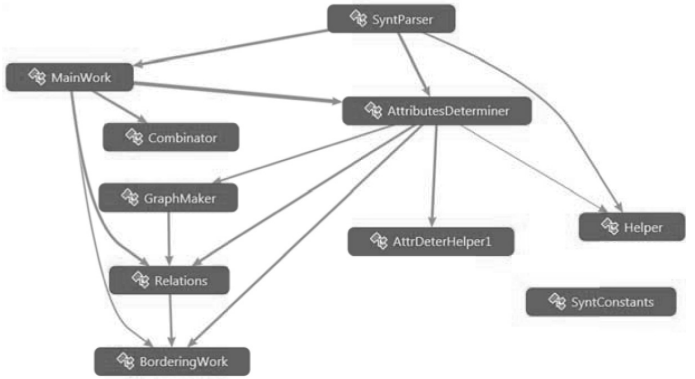


Рис. 2.

5. Возможность вывода результатов в файл.
6. Режим логгирования для удобной отладки.

Определение подлежащего и сказуемого в предложении

Первоочередной задачей СА является корректное распознавание главных членов предложения, от которого зависит весь синтаксический анализ и правильное построение структуры.

Алгоритм определения подлежащего и сказуемого в предложении выглядит так.

Шаг 1. Из входного предложения после морфологического анализа (МА) получить набор МорфС.

Шаг 2. Из МорфС отобрать по первичному признаку кандидатов в подлежащие и сказуемые.

Шаг 3. По вторичному признаку перекрестно сравнить отобранные элементы. Отсечь кандидаты несовместимые друг с другом.

Шаг 4. Объединить разрозненные члены, которые остались без пары.

Под первичным признаком подразумевается набор условий отбора кандидатов первого уровня. Отбор происходит по общим признакам, поэтому в сложных предложениях список кандидатов может быть довольно обширным. В число подлежащих включаются: существительные, местоимения, числительные, прилагательные, причастия в основном падеже. В список сказуемых заносятся все глаголы и причастия.

На втором этапе кандидаты проходят проверку на совместимость друг с другом. Подлежащие проверяются на связь со сказуемыми по следующим правилам.

1. Если подлежащее выражено местоимением, то сказуемое должно быть в том же лице, что и подлежащее.

2. Подлежащее и сказуемое должны быть в одном числе, кроме исключительных случаев.

Определение границ простых предложений

Если предложение сложное, то возникает вопрос обозначения границ простых предложений.

Алгоритм определния границ простых предложений в сложном следующий.

Шаг 1. Преобразовать входное предложение в МорфС. Запомнить позиции знаков препинания.

Шаг 2. Если количество пар подлежащее-сказуемое равно одному, то предложение простое. Границами будут начало и конец предложения. Иначе, переход к шагу 3.

Шаг 3. Если количество пар главных членов предложения равно количеству простых предложений, то запускается цикл. Между i-м и (i+1)-м элементами подсчитывается количество знаков препинания.

Шаг 4. Учитывая знаки препинания определяется принадлежность «нейтральных» слов (слов, которые могут быть отнесены к тому или иному простому предложению, т.к. либо не имеют явных связей согласования в лице или числе) к i-му или (i+1)-му простому предложению.

Определение связей

СинтО являются базовыми элементами СинтС, поэтому рассмотрим их подробнее. Для анализа простых предложений естественного языка достаточно около 16 СинтО. Все СинтО являются бинарными и ориентированными. Ориентированность подразумевает то, что все отношения представлены в формате

$$Y(\text{левая часть}) - X(\text{правая часть}),$$

где главным словом является X, а зависимым – Y.

Все отношения группируются по признаку их типологической близости и делятся на актантные, атрибутивные и сочинительные.

Актантные СинтО – это отношения, компоненты которых, по сути, выражают законченное по смыслу высказывание, и таким образом, сами являются предложениями.

Атрибутивные СинтО – это отношения, при которых один из компонентов является определяемым (*Y*), а другой определяющим (*X*).

Сочинительные СинтО – это отношения, компоненты которых связаны между собой союзами.

Структура данных, в которой содержится набор СинтО, представляет собой список

$$SyntRelations = array\ of\ SyntR.$$

Атрибутами каждого элемента СинтО *X* (главное слово) или *Y* (зависимое слово) является часть речи, к которой относится *X* или *Y* в определенном формате, а также морфологические характеристики данного элемента, такие как падеж, род, число и другие, тоже в принятом в данной работе формате [8,9].

Алгоритмы синтаксического анализа представлены двумя основными алгоритмами: 1) алгоритмом предсинтаксического анализа; 2) алгоритмом формирования СинтС. Второй алгоритм тоже состоит из двух алгоритмов: а) алгоритма установления СинтО; б) алгоритма установления связей между СинтО. На вход данного алгоритма поступает предложение в виде строки. Затем выделяются компоненты предложения, т.е словоформы, разделенные разделителями. В число разделителей входят: точка –´.´; точка с запятой –´;´, запятая –´,´, двоеточие –´:´, пробел и т. д. Все буквы преобразуются к строчным. Выделенные из предложения словоформы заносятся в список *Components* структуры данных *TSentence*. После этого каждый из выделенных компонентов подается на блок морфологического анализатора. В результате МорфС словоформы из *Components*[*i*] возвращается в *MorphS*[*i*] (см. рис. 3).

Как видно из данной схемы, алгоритм представляется в виде *n*-1 итераций для предложения, состоящего из *n* МорфС. На каждом шаге (*i* – номер шага) делается попытка установить СинтО между МорфС *i* и оставшимися *n-i* МорфС. Проверка на возможность существования СинтО между двумя МорфС обозначается как <СинтО?>. Как только СинтО между МорфС устанавлива-

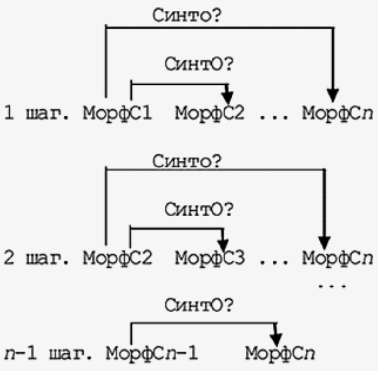


Рис. 3.

ются, номера МорфС *i* и МорфС *j*, образовавших данное СинтО, заносятся в список СинтО данного предложения (*SyntRelations*), и происходит переход к следующему шагу.

За выявление синтаксических отношений в проекте отвечает класс *Relations*. В общем все отношения поделены на две части. К первому типу относятся связи актантные, которые выражают законченную мысль. К ним относятся связи подлежащего и сказуемого. Во вторую группу входят остальные атрибутивные связи [10,11].

Алгоритм установления синтаксических связей выглядит следующим образом.

Шаг 1. Установить связи между подлежащими и сказуемыми.

Шаг 2. Все оставшиеся слова «отлавливаются» и к ним подбираются пары. Для каждого простого предложения выявление отношений производится независимо.

Шаг 3. Если пара подобрана, тип связи устанавливается на основе частей речи.

Шаг 4. Вся информация о связях (зависимое, главное слово, тип связи) записывается в динамические списки.

Визуализация графа

Для визуализации структуры предложения использован инструментарий *QuickGraph*. *QuickGraph* позволяет создавать ориентированные и неориентированные графы для *.NET*. *QuickGraph*

имеет набор готовых для использования алгоритмов, таких как поиск по глубине, *A** поиск, кратчайший путь, *k*-кратчайший путь, максимальный поток, минимальное дерево и т. д. *QuickGraph* поддерживает *MSAGL*, *GLEE* и *Graphviz* для формирования графов, сериализация в *GraphML*.

GraphViz – предоставляет способ представления структурной информации в виде диаграмм абстрактных графов и сетей. Он имеет важные приложения в области сетевого взаимодействия, биоинформатики, разработки программного обеспечения, базы данных и веб-дизайна, машинного обучения и визуальных интерфейсов для других технических областей.

Особенности. Программы компоновки Graphviz содержат описания графиков на простом текстовом языке и делают диаграммы в полезных форматах, таких как изображения и *SVG* для вебстраниц; *PDF* или *Postscript* для включения в другие документы; отображение в интерактивном графическом браузере. *Graphviz* имеет много полезных функций для конкретных диаграмм, таких как варианты цветов, шрифтов, макетов табличных узлов, стилей линий, гиперссылок и пользовательских фигур.

GraphViz предоставляет набор продуктов. Для формирования графов использовался формат *dot*. *Dot* применяется для представления иерархических или слоистых визуализаций ориентированных графов. Это инструмент по умолчанию, который следует использовать, если ребра имеют направленность.

Алгоритм внедрения визуализация графов с помощью указанного инструментария в *.NET* приложения следующий:

1. Используя пространства имен *QuickGraph* и *QuickGraph*. *GraphViz*, сформировать граф и записывать его в файл *.dot*. Также можно отсортировать граф топологически, это полезно для выявления последовательности выполнения связанных процессов.

2. С помощью командной утилиты *dot.exe* из пакета *GraphViz* формируется изображение в желаемом формате и отображается в приложении.

Для визуализации связей был создан класс *GraphMaker*. Задача класса – используя утилиты *GraphViz* и *QuickGraph*, сформировать *dot*-файл и изображение графа.

Dot-файл создается с помощью функции *Generate* класса *GraphVizAlgorithm*. До вызова функции необходимо в объект класса добавить информацию о связях следующим образом: зави-

симое и главное слово отношения добавляются в вершины, описание связи добавляется в название ребра.

Пример *dot*-файла для предложения «*Урамра чупаççĕ хастар каччăсем*». *Dot*-файл содержит информацию о шрифтах, размере текста, форме представления, стилях, вершинах и ребрах.

```
digraph G {
0 [fontname="Times New Roman",  fontsize=10, shape=box, style=diagonals, label="урамра"];
1 [fontname="Times New Roman",  fontsize=10, shape=box, style=diagonals, label="чупаççĕ"];
2 [fontname="Times New Roman",  fontsize=10, shape=box, style=diagonals, label="хастар"];
3 [fontname="Times New Roman",  fontsize=10, shape=box, style=diagonals, label="каччăсем"];
0 -> 1 [ color="#808080FF", fontcolor="#FF0000FF", taillabel="VP", fontname="Times New Roman",
2 -> 3 [ color="#808080FF", fontcolor="#FF0000FF", taillabel="NP", fontname="Times New Roman",
3 -> 1 [ color="#808080FF", fontcolor="#FF0000FF", taillabel="VP", fontname="Times New Roman",
}
```

Для получения изображения сформированный файл поступает на вход инструмента *Graphviz*. Команда вызова библиотеки выглядит следующим образом: «*dot.exe -T png* %~1.*dot*>%~1.*png*». В ней указывается входной *dot*-файл, формат и имя выходного изображения.

Изображение сохраняется в исходной директории.

Выводы

Таким образом разработана библиотека, реализующая синтаксический анализатор чувашского языка. Библиотека создана на платформе *.NET.Framework* в среде *Visual Studio* 2014 на языке *C#*. Разработанная библиотека выполняет следующие задачи:

– определение главных членов предложения;
– определение границ простых предложений в сложном;
– выявление синтаксических связей;
– визуализацию графа синтаксических отношений;
– логгирование работы для удобной отладки.

Результаты работы библиотеки могут широко применяться в системах автоматической обработки чувашских текстов, быть составной частью лингвистического процессора.

В настоящее время синтаксический анализатор чувашского языка доступен для тестирования на сайте национального корпуса чувашского языка по адресу http://yuman21.ru/Syntax.

Благодарность

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 15-04-00532.

ЛИТЕРАТУРА

1. Желтов П.В. Лингвистические процессоры. Формальные модели и методы: теория и практика: Монография. – Чебоксары: Изд-во Чуваш. ун-та, 2006. – 208 с.

2. Желтов П.В. Формальные методы и модели в сравнительно-сопоставительном языкознании: Монография. – Чебоксары: Изд-во Чуваш. ун-та, 2006. – 252 с

3. Желтов П.В. Сравнительные исследования морфем чувашского языка. – Чебоксары: Изд-во Чуваш. ун-та, 2013. – 166 с.

4. Желтов П.В. Создание национального корпуса чувашского языка: проблемы и перспективы // Современные проблемы науки и образования. – 2015. -№ 1-1. С. 338.

5. Желтов П.В. Лингвистические процессоры в системах искусственного интеллекта: Монография. – Чебоксары: Изд-во Чуваш. ун-та, 2007. – 100 с.

6. Желтов П.В., Губанов А.Р., Желтов В.П. Морфологический стандарт национального корпуса чувашского языка // Современные проблемы науки и образования. – 2015. № 2. С.180.

7. Желтов П.В. Исследования исторического развития чувашского языка. – Чебоксары : Изд-во Чуваш. ун-та, 2013. – 166 с.

8. Желтов П.В. Компьютерное моделирование многоагентных систем. – Чебоксары: Изд-во Чуваш. ун-та, 2008. – 112 с.

9. Желтов П.В. Модели и методы обработки символьной информации на сетях Петри. – Чебоксары: Издательство Чуваш. ун-та, 2012. – 108 с.

10. Желтов П.В. Моделирование многоагентных систем сетями Петри: Монография. – Чебоксары: Изд-во Чуваш. ун-та, 2008. – 108 с.

11. Zheltov P, Fomin E., Luutonen J. Reverse Dictionary of Chuvash. Societe Finno-ougrienne, Helsinki. 2009. – 344 p.

УДК 519.25

THE USE OF DIGRAMS AND TRIGRAMS TO IDENTIFY THE AUTHOR OF THE TEXT ON THE MATERIAL OF THE YAKUT LANGUAGE

N. Leontiev
M. K. Ammosov North-Eastern Federal University
leonza@mail.ru

In this paper we consider the problem of identifying the author in newspaper texts. Using the machine body of the Yakut language is used in the correlation analysis digrams and trigrams. Discusses texts by different authors, as well as the texts of one author. Application of the method digrams and trigrams improves the accuracy of the method. Identifies the causes of low probability, and it provides the ways to improve the accuracy of identification of the author of the text.

Keywords: author identification, Sakha corpora, Yakut language.

ПРИМЕНЕНИЕ ДИГРАММ И ТРИГРАММ ДЛЯ ИДЕНТИФИКАЦИИ АВТОРА ТЕКСТА НА МАТЕРИАЛАХ ЯКУТСКОГО ЯЗЫКА

Н.А. Леонтьев
Северо-Восточный федеральный университет им. М.К. Аммосова
leonza@mail.ru

В данной работе рассматривается проблема идентификации автора в газетных текстах. Используя машинный корпус якутского языка проводится корреляционный анализ диграмм и триграмм. Рассматриваются тексты разных авторов, а также тексты одного автора. Применение метода диграмм и триграмм позволяет повысить точность метода. Выявляются причины низкой вероятности и приводятся способы повышения точности идентификации автора текста.

Ключевые слова: идентификация автора, метод диграмм, метод триграмм якутский язык, машинный корпус.

Развитие письменности дало возможность для распространения знаний, но вместе с увеличением количества авторов, наступило и возможность скрытия авторства письма. Возникновение анонимных источников влечет снижение истинности изложенных сведений письменных источников. Установление авторства также необходимо в случае обнаружение текстов с неизвестным автором, а также в установлении истинного автора письменного текста.

Для идентификации автора также используется такой критерий как энтропия [1]. Применение N-грамм и нейронных сетей позволяет увеличить точность определения автора [2]. Применение нейронных сетей для метода опорных векторов также требует большого времени для обучения сети и существенных объемов текстов [3]. Предлагают также строить модель автора и на его основе производить идентификацию [4].

Также исследуются проблема идентификации автора для других языков с помощью биграмм, например азербайджанского [5].

Некоторые авторы создают полный информационный портрет на основе

– буквы, идущие подряд в слове;

– буквы, встречающиеся в слове через один– символ; двубуквенные сочетания (биграммы),

– идущие подряд в слове (из всех возможных биграмм для исследования были выбраны 30 наиболее часто встречаемых в русских текстах: ва, ка, ла, на, ра, та, ов, не, ре, ли, ни, ал, ел, ол, ен, он, во, го, ко, ло, но, по, ро, то, ер, ор, пр, ат, от, ст);

– гласные буквы (не обязательно стоящие– рядом);

– служебные слова в тексте [6].

Применяют также статистические характеристики определенных длин слов [7], в том числе и для якутского языка [8].

Методика

Для исследования используется машинный газетный корпус, разработанный автором для задач автоматизированной обработки якутского языка (Саха тыла) [9]. Объем машинного корпуса составляет более 12 млн. словоупотреблений, содержится более 21 тысяч единиц текста.

Применение биграмм и триграмм возможно также в иденти-

фикации языка текстового сообщения [10,11], где он показывается хорошие результаты.

Для анализа для каждого текста вычисляется частотная таблица встречаемых диграмм и триграмм.

Основная часть

Для идентификации автора используется корреляционная функция Пирсона:

$$P(x) = \frac{\sum_{i=0}^{N-1} (a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=0}^{N-1} (a_i - \overline{a})^2 \sum_{i=0}^{N-1} (b_i - \overline{b})^2}},$$

где N – число элементов массива.

Элементами массива являются значения частоты диграмм и триграмм, а индексами массива сами диграммы и триграммы.

Для проверки метода идентификации автора с помощью диграмм и триграмм были выбраны 200 статей из машинного корпуса. Вычислена взаимная корреляция Пирсона по всем статям.

Для анализа были взяты 9 статей и произведены выборка их коэффициентов корреляции.

Список авторов:

1. Филипп Охлопков
2. Мария Христофорова
3. А.Аммосов
4. А.Желобцов
5. А. Кривошапкин-Айыһа
6-8. А. Максимов

В результате вычисления корреляционной функции, была получена таблица 1, где в первой строке и столбце, расположен номер автора.

Таблица 1. **Вероятность взаимной корреляции диграмм**

№	1	2	3	4	5	6	7	8
1	1,000	0,747	0,878	0,741	0,829	0,810	0,843	**0,886**
2	0,765	1,000	**0,802**	0,665	0,798	0,718	0,776	**0,802**
3	0,879	0,794	1,000	0,742	0,864	0,844	0,901	**0,909**
4	0,757	0,676	0,751	1,000	0,756	0,712	0,731	**0,800**

5	0,823	0,789	0,865	0,738	1,000	0,767	0,824	**0,879**
6	0,806	0,721	0,840	0,694	0,748	1,000	**0,857**	0,854
7	0,853	0,752	0,902	0,723	0,826	0,859	1,000	**0,917**
8	0,887	0,788	0,907	0,790	0,880	0,861	**0,915**	1,000

В таблице 1 главная диагональ это вероятность автокорреляции, которая составляет в данном случае 100%. Максимальная вероятность по строкам сосредоточена на столбец №9, что не позволяет точно определить автора. Также у столба №3 также имеется выброс вероятность, что является не корректным результатом. Если ввести определение автора по величине вероятности корреляции от 0,9, то результаты точности определение автора в столбцах №6-9 падает. Вероятность определения автора зависит от точности получения вектора диграмм, при этом большое значение имеет объем текста и наличием в нем собственного письменного стиля.

При анализе всех 200 текстов, получилась аналогичная картина. Имеются статьи, которые не имеют коэффициента корреляции с другими больше чем 0,9, но и почти половина авторов статей не может быть корректно идентифицирована с помощью метода диграмм.

Таблица 2. **Вероятность взаимной корреляции триграмм**

№	Светлана ТИМОФЕЕВА						Туйаара СИККИЭР					
	1	2	3	4	5	6	7	8	9	10	11	12
1	1	0,7	0,74	0,75	**0,79**	0,73	0,75	0,72	0,68	0,68	0,75	0,71
2	0,72	1	0,68	0,73	0,71	0,67	0,73	0,73	0,76	0,68	**0,75**	0,7
3	**0,79**	0,7	1	0,72	0,72	0,72	0,68	0,68	0,67	0,68	0,75	0,7
4	0,74	0,72	0,68	1	0,75	0,7	0,74	0,74	0,7	0,7	**0,76**	0,73
5	**0,75**	0,66	0,63	0,7	1	0,63	0,7	0,68	0,64	0,65	0,72	0,69
6	**0,74**	0,69	0,71	0,72	0,69	1	0,69	0,68	0,67	0,72	0,73	0,64
7	0,76	0,73	0,67	0,76	0,74	0,68	1	0,75	0,74	0,74	**0,79**	0,74
8	0,73	0,73	0,65	0,75	0,74	0,67	0,75	1	0,74	0,71	**0,8**	0,71
9	0,68	0,75	0,65	0,72	0,69	0,66	0,74	0,74	1	0,69	**0,8**	0,67
10	0,69	0,68	0,66	0,73	0,71	0,71	0,75	0,72	0,69	1	**0,79**	0,69
11	0,76	0,75	0,71	0,77	0,76	0,72	0,79	0,79	**0,8**	0,77	1	0,76
12	0,69	0,69	0,62	0,7	0,7	0,62	0,72	0,68	0,67	0,66	**0,74**	1

В таблице приведены коэффициент взаимной корреляции в случае наличия текстов двух авторов, вычисленные с помощью триграмм. Точно идентифицировать авторов не получается, хотя в некоторых случаях средний коэффициент корреляции среди текстов одного автора чуть-чуть выше.

Выводы

Для получения более точной вероятности определения автора необходимо:

– Собрать более точный портрет автора на основе текстов с более объемным содержанием;

– Точно установить авторство текста для эталонного шаблона;

– Применить более сложные технологии определения авторства.

Влияние агглюнативности якутского языка смазывает точную картину при определении автора текста, также на вероятность определения влияют часто используемые обороты в газетных публикациях.

ЛИТЕРАТУРА

1. Зверева Ю.В. Идентификация автора на основе энтропии его текстов // В сборнике: Интеграционные процессы в науке в современных условиях Материалы Международной (заочной) научно-практической конференции. Научное (непериодическое) электронное издание. Под общей редакцией А.И. Вострецова. 2016. С. 16–21.

2. Романов А.С., Шелупанов А.А., Бондарчук С.С. Обобщенная методика идентификации автора неизвестного текста // Доклады Томского государственного университета систем управления и радиоэлектроники. 2010. Т. 1. № 1. С. 108–112.

3. Романов А.С. Методика идентификации автора текста на основе аппарата опорных векторов // Доклады Томского государственного университета систем управления и радиоэлектроники. 2009. Т. 1. № 2. С. 36–42.

4. Воробьева А.А. Модель автора и текста в решении задачи идентификации автора анонимных сообщений в сети Интернет // В сборнике: Региональная информатика и информационная безопасность Сборник трудов. Санкт-Петербургское Общество информатики, вычислительной техники, систем связи и управления. 2015. С. 95–98.

5. Айдазаде К.Р., Талыбов С.Г. Идентификация авторства текстов на азербайджанском языке // Прикладная математика и фундаментальная информатика. 2016. № 3. С. 142–146.

6. Суркова А.С. Идентификация авторства текстов на основе информационных портретов // Вестник Нижегородского университета им. Н.И. Лобачевского. 2014. № 3-1. С. 145–149.

7. Пигарева А.В., Черкасова Т.Х. Определение автора текста статистическими методами // В сборнике: Фундаментальные и прикладные науки сегодня. Материалы X международной научно-практической конференции: в 3-х томах. 2016. С. 157–160.

8. Леонтьев Н.А., Протопопова В.Ф. Программное определение автора текста на якутском языке статистическим методом // В сборнике: Высокие технологии и инновации: векторы, проблемы и приоритеты Сборник научных трудов по материалам I Международной научно-практической конференции. 2017. С. 40–44.

9. Leontiev N.A. The newspaper corpus of the yakut language// Proceeding of the International Conference «Turkic Languages Prosessing: TurkLang -2015». – 2015. – P. 233–235.

10. Леонтьев Н.А. Распознавание языка текстовых сообщений с помощью биграмм на материалах якутского языка // Современное состояние естественных и технических наук. 2014. № XIV. С. 88–91.

11. Леонтьев Н.А., Слепцов И.А. Идентификация текстового документа с помощью триграмм на материалах якутского языка // Вестник Северо-Восточного федерального университета им. М.К. Аммосова. 2015. № 4 (48). С. 45–50.

УДК 004

CREATING A NOUN DATABASE FOR THE ELECTRONIC CORPUS OF TEXTS OF THE TUVAN LANGUAGE

B. Oorzhak, A. Khertek,
V. Ondar, A. Salchak,
M. Kuzhuget
Tuvan State University, Kyzyl
kuzhuget.m55@mail.ru

The proposed article of the authors' team outlines the principles of creating a noun database for the ECTTI.

Creating a database of noun names for Tuvan is a necessary stage of work. All nouns of the Tuva language are divided into the main semantic classes: man, animal, object, natural objects and phenomena, abstract concepts. Semantic classes, subclasses and descriptors are assigned tags in Tuvan, Russian and English languages, with the help of which an automated search will be performed. The database of full-valued tokens of Tuvan language will serve to identify lexical compatibility of lexemes. Four basic semantic classes are distinguished: human, 2) animal, 3) thing, 4) natural objects and phenomena, 5) abstract concepts. In semantic classes, subclasses are identified: 1) substantive nouns; 2) proper names; 3) non-prudential nouns (abstract concepts). All subject names are divided into subgroups: "person" - names of kinship, profession, ethnonym; "Animal" - wild animals, domestic animals, birds - wild birds, domestic birds, fish, insects; "Subject" - substances and materials, household items, buildings and structures, tools; "Natural objects and phenomena" - plants, names of weather phenomena, celestial bodies, landscape objects; "Abstract concepts" - emotions, sensory perceptions, universal submission. Further, the selected subclasses are subdivided into smaller semantic groups.

Keywords: Tuvan language; lexical-semantic classes and subclasses; handles; tags; noun.

СОЗДАНИЕ БАЗЫ ДАННЫХ ИМЕНИ СУЩЕСТВИТЕЛЬНОГО ДЛЯ ЭЛЕКТРОННОГО КОРПУСА ТЕКСТОВ ТУВИНСКОГО ЯЗЫКА[1]

Б.Ч. Ооржак, А.Б. Хертек, В.С. Ондар, А.Я. Салчак, М.А. Кужугет
Тувинский государственный университет, Кызыл
kuzhuget.m55@mail.ru

В предлагаемой статье коллектива авторов излагаются принципы создания базы данных имени существительного для ЭКТТЯ.

Создание базы данных имен существительных тувинского языка является необходимым этапом работы. Все имена существительные тувинского языка распределяются на основные семантические классы: человек, животное, предмет, природные объекты и явления, абстрактные понятия. Семантическим классам, подклассам и дескрипторам присваиваются тэги на тувинском, русском и английском языках, при помощи которых будет производиться автоматизированный поиск. Создаваемая база данных полнозначных лексем тувинского языка будет служить для выявления также лексической сочетаемости лексем. Выделено четыре базовых семантических класса: 1) человек, 2) животное, 3) предмет, 4) природные объекты и явления, 5) абстрактные понятия. В семантических классах выявлены подклассы: 1) предметные имена существительные; 2) имена собственные; 3) непредметные имена существительные (абстрактные понятия). Все предметные имена подразделяются на подгруппы: «человек» – имена родства, профессия, этноним; «животное» – дикие животные, домашние животные, птицы – дикие птицы, домашние птицы, рыбы, насекомые; «предмет» – вещества и материалы, бытовые принадлежности, здания и сооружения, инструменты; «природные объекты и явления» – растения, названия погодных явлений, небесные тела, объекты ландшафта; «абстрактные понятия» – эмоции, чувственные восприятия, универсальные представления. Далее выделенные подклассы подразделяются на более мелкие семантические группы.

Ключевые слова: тувинский язык; лексико-семантические классы и подклассы; дескрипторы; тэги; имя существительное.

[1] Работа выполнена при поддержке РГНФ (проект «Создание базы данных лексического фонда тувинского языка», грант №16-04-12020).

В статье представлен ход работ над созданием базы данных для Электронного корпуса текстов тувинского языка (ЭКТТЯ). Этот этап является продолжением работы коллектива авторов Тувинского государственного университета (Научно-образовательного центра «Тюркология» в сотрудничестве с кафедрой информатики) по включению текстов на тувинском языке в электронную базу и разработке разметки корпуса.

Электронная база данных лексического фонда будет функционировать в ЭКТТЯ как справочно-поисковая система, при помощи которой будет автоматизирован поиск необходимых фрагментов текстов с искомой семантической информацией с последующим их использованием при составлении учебников (в том числе и электронных) и других учебных материалов и контента, а также при составлении словарей.

База данных имени существительного ЭКТТЯ основывается на распределении всех полнозначных лексем тувинского языка на семантические разряды (классы) слов. Выделяются четыре базовые семантические классы: 1) человек, 2) животное, 3) предмет, 4) природные объекты и явления, 5) абстрактные понятия. При этом, имена существительные подразделяются на подклассы: 1) предметные имена существительные; 2) имена собственные; 3) непредметные имена существительные (абстрактные понятия). Все предметные имена, в свою очередь, подразделяются на: «человек» – имена родства, профессия, этноним; «животное» – дикие животные, домашние животные, птицы – дикие птицы, домашние птицы, рыбы, насекомые; «предмет» – вещества и материалы, бытовые принадлежности, здания и сооружения, инструменты; «природные объекты и явления» – растения, названия погодных явлений, небесные тела, объекты ландшафта; «абстрактные понятия» – эмоции, чувственные восприятия, универсальные представления. Далее выделенные подклассы подразделяются на более мелкие семантические группы.

Пример распределения предметных имен существительных тувинского языка и их помет приведен в таблице 1. Названия лексико-семантических классов, подклассов и дескрипторов обозначаются тэгами на тувинском, русском и английском языках.

Таблица 1. **Имена существительные тувинского языка. Предметные имена. Nouns in the Tuvan language. Object names**

Кижи / Человек / Human	***Дөргүл-төрел аттары***/Имена родства / Names of kinship	*ада, ача* ‘отец, папа’, *ава* ‘мама’, *кырган-ава* ‘бабушка’, *кырган-ача* ‘дедушка’, *угба* ‘сестра’, *акы* ‘брат’, *дуңма* ‘младший брат / младшая сестра’, *даай* ‘дядя’, *күүй* ‘жена дяди’	
	Профессия/ Profession	*эмчи* ‘врач’, *башкы* ‘учитель’, *ыраажы* ‘певец’, *чолаачы* ‘водитель, шофер’	
	Этноним/ Ethnonym	*кыдат* ‘китаец, китаянка’, *бурят* ‘бурят, бурятка’, *моол* ‘монгол, монголка’, *орус* ‘русский, русская’	
Дириг амытан / Животное / Animal	***Дириг амытан***/ Животные / Animal	***черлик***/дикие/ wild	*адыг* ‘медведь’, *дииң* ‘белка’
		азырал/домашние/ domestic	*инек* ‘корова’, *ыт* ‘собака’
	Куштар/ Птицы / Birds	***черлик***/дикие/ wild	*хартыга* ‘коршун’, *ускушкаш* ‘ремез’
		азырал/домашние/ domestic	*дагаа* ‘курица’, *кас* ‘гусь’
	Балыктар/ Рыбы / Fishes	*ак-балык* ‘елец’, *шортан* ‘щука’, *кадыргы* ‘хариус’	
	Курт аймаа/ Насекомые/ Insects	*шартылаа* ‘кузнечик’, *ары* ‘пчела’, *ымыраа* ‘комар’, *сээк* ‘муха’, *шыйлашкын* ‘дождевой червь’	

Чүүл / Предмет / Thing	***Бүдүмелдер***/ Вещества и материалы / Substances and materials	*суг* ‘вода’, *чугай* ‘известь’, *торгу* ‘шелк’, *алдын* ‘золото’, *каң* ‘сталь’, *хөмүр* ‘уголь’, *кидис* ‘войлок’, *маны* ‘мрамор’, *хүлер* ‘бронза’, *хола* ‘медь, жёлтая медь’, *мөңгүн* ‘серебро’	
	Эт-херексел / Бытовые принадлежности, утварь/ Household accessories	***Аяк-сава /*** Посуда / Dishes	*аяк* ‘пиала’, *паш* ‘чугунная чаша для приготовления пищи’, *дииэе* ‘блюдце’, *бижек* ‘нож’, *хууң*‘ведро’
		Өг, бажың дерии / Мебель / Furniture	*аптара* ‘сундук’, *орун* ‘кровать’, *сандай* ‘табуретка, стул’
	Идик-Хеп / Одежда, обувь/ Clothing, footwear	*тон* ‘национальная верхняя одежда, пальто’, *хөйлең* ‘рубашка’, *чүвүр* ‘штаны, брюки’, *идик* ‘национальная обувь, сапоги’, чеңижок ‘безрукавка, жилетка’.	
	Бажың-балгат, тудуг объектилери/ Здания и сооружения / Buildings and constructions	***Чурттаар оран-сава /*** Жилище / Housing	*өг* ‘юрта’, *чадыр* ‘чум, шалаш’, *алажы өг* ‘чум’, *кидис өг* ‘войлочная юрта’, *доорбаш өг* ‘бревенчатая юрта*’, бажың* ‘дом, здание’
		Тудуг объектилери / Нежилые постройки / Constructions	*көвүруг* ‘мост’, *кажаа* ‘загон для скота’, *дес кажаа* ‘закрытый загон для скота’

Чүүл / Предмет / Thing	***Эдилел /*** Принадлежности / Accessories	***Эр кижиниң эдилели /*** Мужские принадлежности/ Men's accessories	*балды* ‘топор’, *сыырткыыш* ‘удочка’, *кестик* ‘ножик’, *ча* ‘лук’, *чаныы*‘брусок, оселок, мягкий точильный камень’, *хол хирээзи* ‘ножовка’
		Кыс кижиниң эдилели / Женские принадлежности / Women's accessories	*баш шүүрү* ‘гребень’, *билзек*‘кольцо’, *билектээш* ‘браслет’, *боошкун* ‘девичье накосное украшение из трех нитей бус’, *ине хавы* ‘футляр для иглы’, *сырга*‘серьги’
Бойдус объектилери болгаш бойдустуң болуушкуннары/ Природные объекты и явления / Natural objects and phenomena	***Үнүш***/растения / Plants	*оът-сиген* ‘трава’, *ыяш* ‘дерево’, *шиви* ‘ель’, *чечек* ‘цветок’	
	Агаар байдалы / Погодные явления / Weather conditions	*чаъс* ‘дождь’, *кызаңнаашкын* ‘гроза’, *челээш* ‘радуга’, *хат* ‘ветер’, *диңмирээшкин* ‘гром’, *шуурган* ‘буря’	
	Дээр объектилери / Небесные тела / Heavenly bodies	*ай* ‘месяц, луна’, *хүн* ‘солнце’, *сылдыс* ‘звезда’, *Шолбан* ‘Венера’, *Үгер* ‘Плеяды’, *Чеди-хаан* ‘Большая Медведица’	
	Ландшафт / Landscape	*аяң* ‘горный луг’, *даг* ‘гора’, *хем* ‘река’, *хову* ‘степь’, *чоога* ‘ложбина, впадина, овраг’, *кырлаң* ‘небольшой горный хребет, отрог’, *хая* ‘скала’, *баалык* ‘седловина горы’	

Подкласс «Имена собственные» подразделяются на: «человек» – имя, отчество, фамилии, название рода; «животное» – ло-

шадь, корова, собака; «природные объекты» (названия местностей) – топонимы, гидронимы. См. таб. 2.

Таблица 2. **Имена существительные в тувинском языке. Имена собственные. Nouns in the Tuvan language. Own names**

Кижи / Человек / Human	***Ат***/Имя / Name	*Чечек-оол*, *Артыш*, *Менди, Кара-кыс*
	Адазының ады / Отчество / Middle name	*Дүрген-оолович*, *Бай-Караевна*
	Фамилиялар / Фамилии / Surname	*Сарыг-оол, Шыырап, Сагаачы*
	Аймак-сөөк ады / Названия родов / Names of genera	*Кыргыс, Монгуш*, *Куулар*, *Түлүш*
Дириг амытан / Животное / Animal	***Аът***/Лошадь / Horse	*Калчан-Шилги, Сарала, Доругдай.*
	Инек/Корова / Cow	*Дөңгүр, Дагыр-Мыйыс, Шокар*
	Ыт/Собака / Dog	*Ак-Төш, Көстүк, Калдарак*
Черлер / Местность / Terrain	***Черлер аттары*** / Топонимы / Placenames	*Кызыл*, *Чаа-Хөл, Кунгуртуг, Бай-Тайга*
	Суглар аттары / Гидронимы / Hydronyms	*Улуг-Хем, Дус-Хөл, Шивилиг*

Непредметные имена (абстрактные понятия) подразделяются на: эмоции, чувственное восприятие, универсальные представления. Семантические пометы и распределение непредметных имен существительных приводится в таблице 3.

Таблица 3. **Имена существительные тувинского языка. Непредметные имена. Nouns in the Tuvan language. Non-object names**

Туугай билиш-киннер/ Абстрактные понятия / Abstract concepts	***Сагыш-сеткил илерээшкини***/ Эмоция / Emotion	*өөрүшкү* ‘радость’, *муңгарал* ‘горе’, *дадагалзал* ‘сомнение’
	Миннишкин/ Чувственное восприятие / Sensory perception	*дааш* ‘шум’, *амдан* ‘вкус’, *өң* ‘цвет’, *ыыт* ‘звук’
	Ниити билиишкиннер/ Универсальные представления / Universal submission	*болуушкун* ‘событие’, *кылдыныг* ‘действие’, *байдал* ‘обстоятельство’, *үе* ‘время’

Составленная база данных имени существительного в тувинском языке является частью лексического фонда и будет функционировать в ЭКТТЯ как справочно-поисковая система, при помощи которой будет автоматизирован поиск необходимых фрагментов текстов с искомой семантической информацией с последующим их использованием при составлении учебников (в том числе и электронных) и других учебных материалов и контента, а также при составлении словарей.

ЛИТЕРАТУРА

1. Ооржак, Б. Ч., Хертек, А. Б. Разработка семантической разметки электронного корпуса тувинского языка // Материалы 3-ей Международной конференции по компьютерной обработке тюркских языков «TurkLang 2015». Казань, 17–19 сентября 2015. Казань: Изд-во АН Республики Татарстан, 2015. – С. 351–362.

2. Ооржак, Б. Ч, Хертек, А. Б., Кужугет, М. А., Салчак, А. Я., Ондар, В. С., Чамзырын, Е. Т. Создание базы данных лексического фонда тувинского языка // Труды Международной конференции по компьютерной и когнитивной лингвистике. TEL-2016. Казань, 21–24 апреля 2016. Казань: Изд-во Казанского госуниверситета, 2016. – Вып. 17. – С. 278–281.

УДК 512.141:004.655

ON THE SUBCORPUS OF APHORISTIC GENRES OF BASHKIR FOLKLORE

Z. Sirazitdinov, L. Buskunbaeva, A. Ishmukhametova, G. Shamsutdinova
Institute for history, language and literature of Ufa scientific center of RAS
sazin11@mail.ru

The article discusses the principles of the design of corpora of texts aphoristic genres in the Bashkir folklore. Small genres of folklore included in the database of will allow the reader to explore the language of folklore. The authors dwell in detail on the problem of annotating data texts. Also the article discusses the possibility of using the developed case in the scientific works and educational process.

Keywords: the Bashkir language, Turkic languages, corpus linguistics, database, folklore, aphoristic genres.

О ПОДКОРПУСЕ АФОРИСТИЧЕСКИХ ЖАНРОВ БАШКИРСКОГО ФОЛЬКЛОРА

З.А. Сиразитдинов, Л.А. Бускунбаева, А.Ш. Ишмухаметова, Г.Г. Шамсутдинова
Института истории, языка и литературы Уфимского научного центра РАН, Уфа, Россия
sazin11@mail.ru

В статье рассматриваются принципы разработки корпуса текстов афористических жанров башкирского фольклора. Данный подкорпус разрабатывается как часть корпуса фольклорных текстов в русле общей концептуальной модели корпусов башкирского языка, включающих в себя корпусы прозаических, публицистических (газетных и журнальных) и фольклорных текстов. Концептуальная модель корпусов разработана на основе системы управления базами данных ORACLE. Включенные в базу данных тексты малых жанров фольклора явятся источником для исследования языка фольклора. Авторы подробно останавливаются на проблеме аннотирования

данных текстов. Рассматриваются возможности использования корпуса в научных трудах, учебном процессе.

Ключевые слова: башкирский язык, тюркские языки, корпусная лингвистика, база данных, фольклор, афористические жанры, система разметок.

1. Введение

Фольклорные материалы, отражающие быт и мировоззрение народа, содержащие архаические элементы, являются ценным источником как для теоретических исследований истории развития языка, определения языковой картина мира, так и для практической лексикографии. Поэтому в отечественной корпусной лингвистике интерес к текстам народного творчества с каждым днем возрастает: активно ведутся работы по созданию корпусов фольклорных текстов русского [Николаев, 2015], калмыцкого [Куканова, 1912], нганасанского [Корпус нганасанских фольклорных текстов], тувинского [Салчак, 2012] языков. Институтом этнологии и антропологии разрабатываются корпусы фольклора ряда языков Сибири (эвенкийского, шорского, ненецкого, телеутского) [Корпусы ИЭА РАН]. Объектами этих корпусов выступают эпические, сказочные, библейские и мифологические тексты. Афористический жанр до сегодняшнего дня ни в одном из языков народов России не является объектом построения корпуса, корпусного исследования. Данная проблема впервые поднимается башкирскими языковедами.

Афористические жанры башкирского народного творчества – небольшие по объему фольклорные произведения, которые включают в свой состав пословицы (мәҡәлдәр), поговорки (әйтемдәр), загадки (йомаҡтар), приметы (һынамыштар), запреты (тыйыуҙар), предсказания (юрауҙар) и т.д. В лаконичных и емких суждениях находят отражение жизненные наблюдения и правила житейской мудрости башкирского народа.

Сотрудниками отдела фольклористики и искусства Института истории, языка и литературы Уфимского научного центра Российской академии наук (далее – ИИЯЛ УНЦ РАН) во время многочисленных экспедиций по районам Республики Башкортостан, соседним областям и республикам, где компактно про-

живает башкирское население, был собран богатый материал по афористическим жанрам башкирского фольклора. Данный материал представлен в виде отдельных томов “Башкирского народного творчества” [Башҡорт халыҡ ижады, 1995; Башҡорт халыҡ ижады, 2007; Башҡорт халыҡ ижады, 2006], словарей [Башҡортса-русса фразеологик һүҙлек, 1973; Духовное наследие.., 2008; Башкирско-англо-русский словарь адекватных пословиц и поговорок, 2002] и монографий [Надршина Ф.А., 2008; Нәҙершина, 1983].

Объемы собираемого материала по фольклору растут, и в основном они представлены на бумажных носителях. Создание электронной базы данных текстов афористических жанров, интегрированной в корпус фольклора башкирского языка, позволяет зафиксировать их в единой базе и пополнять по мере фиксации. Данный проект предоставляет возможность решать задачи сохранения культурного наследия башкирского народа с помощью новых технологий и способствовать широкому распространению материалов башкирского фольклора в ознакомительных и научных целях.

2. Архитектура подкорпуса

Данный подкорпус разрабатывается как часть корпуса фольклорных текстов в русле общей концептуальной модели корпусов башкирского языка, включающих в себя корпусы прозаических, публицистических (газетных и журнальных) и фольклорных текстов [Бускунбаева 2011, 45–51; Бускунбаева 2012, 139–141; Бускунбаева 2012, 54–58; Бускунбаева 2013, 135–140; Сиразитдинов 2014, 86–89; Сиразитдинов, 2015, 658–664; Сиразитдинов 2013; Сиразитдинов 2011, 269–274].

Для функционирования корпусов в сети Интернет лабораторией лингвистики и информационных технологий ИИЯЛ УНЦ РАН разработана интегрированная система, позволяющая создавать корпусы, осуществлять широкий круг поисковых задач и администрирования баз данных [Сиразитдинов, Полянин 2014; Sirazitdiniv, 2014]. Она разработана на основе системы управления базами данных ORACLE. Корпусы представлены в Машинном фонде башкирского языка (mfbl2.ru).

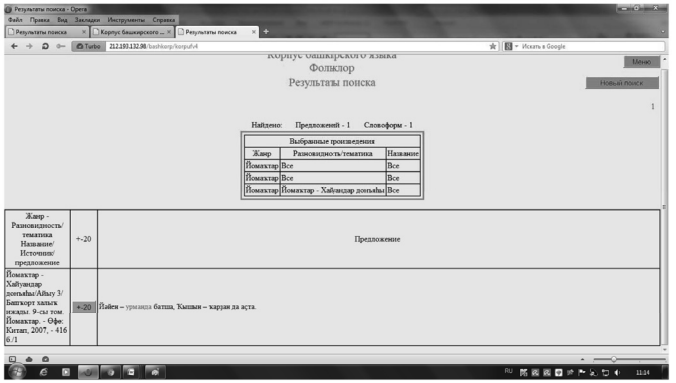


Рис. 1. Интерфейс подкорпуса текстов афористических жанров башкирского фольклора

3. Система разметок текстов афористических жанров башкирского фольклора

На сегодняшний день разразработаны метатекстовая и лингвистическая (морфологическая) разметки текстов афористических жанров башкирского фольклора.

Метатекстовая разметка текстов афористических жанров (информация о тексте):

– **жанр** (пословица ‘мәҡәл’, поговорка ‘әйтем’, загадка ‘йомаҡ’, примета ‘һынамыш’, запрет ‘тыйыу’, предсказание ‘юрау’);

– **тематика** (например, для загадок представлены следующие виды: земля и небо, явления природы ‘ер һәм күк, тәбиғәт күренештәре’, мир растений ‘үҫемлектәр донъяһы’, мир животных ‘хайуандар донъяһы’, человек и его жизнедеятельность ‘кеше һәм уның тормошо’);

– **название** (если имеется, например, в загадках в качестве названия дается его ответ);

– **источник** (название источника, год издания);

– **объем текста** (количество предложений, словоформ).

Система морфологической разметки в данном корпусе ориентирована на представление всех регулярных словоизменительных грамматических форм.

Морфологическая информация башкирской словоформы в корпусе включает:

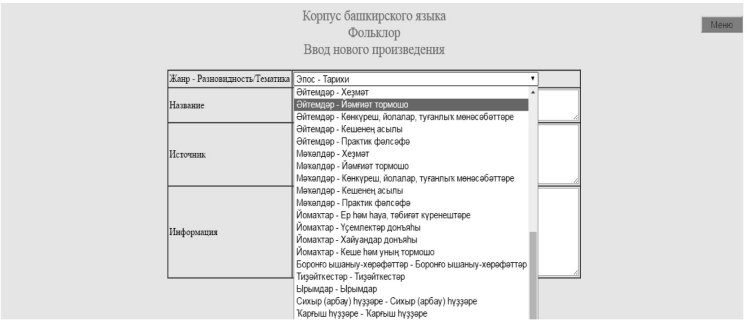


Рис.2. Блок администратора, включающий программные средства для ввода текстов

– исходную форму слова (лемму);
– частеречную характеристику;
– совокупность морфологических признаков по типу агглютинативных аффиксов словоизменения, которые подразделяются на именные и глагольные формы.

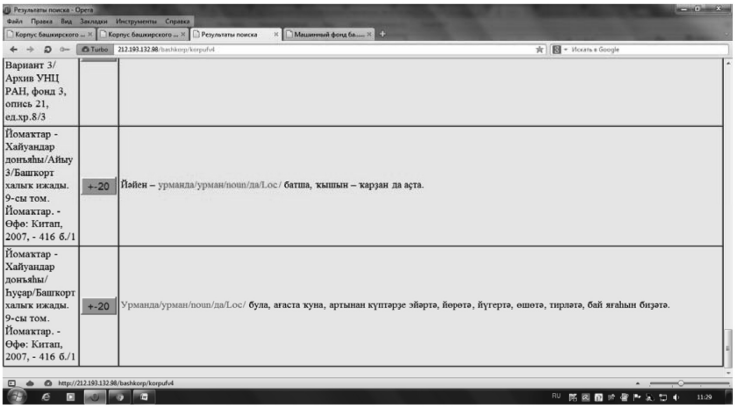


Рис. 3. Морфологическая разметка текстов афористических жанров башкирского фольклора

4. Заключение

Таким образом, создание электронной базы текстов афористических жанров башкирского фольклора, интегрированной в

корпус фольклора дает возможность пользователю исследовать язык текстов афористических жанров башкирского фольклора, получить лингвистические данные относительно морфологической информации представленных текстов.

Данный подкорпус ориентирован для исследователей с разными интересами и задачами – лингвистам, фольклористам, журналистам, преподавателям, учащимся.

Благодарность

Исследование выполнено при финансовой поддержке РФФИ и Правительства Республики Башкортостан в рамках научного проекта № 17-14-02010 а/р.

ЛИТЕРАТУРА

1. Башҡортса-русса фразеологик һүҙлек. – Өфө: Башҡ. Китап нәшр., 1973. 168 б. (Нәҙершина Ф.А., Ураҡсин З.Г., Йосопов Х.Г.)

2. Башҡорт халыҡ ижады. I т. Йола фольклоры. Өфө, 1995.

3. Башҡорт халыҡ ижады. IX т. Йомаҡтар. Өфө, 2007.

4. Башҡорт халыҡ ижады. X т. Мәҡәлдәр һәм әйтемдәр. Өфө, 2006.

5. Башкирско-англо-русский словарь адекватных пословиц и поговорок. Авт.-сост. – Ф.А. Надршина, Э.М. Зубаирова [Созинова]. – Уфа: Китап, 2002. – 160 с.

6. Бускунбаева Л.А., Сиразитдинов З.А., Ишмухаметова А.Ш., Ибрагимова А.Д., Мигранова Л.Г. Корпус текстов периодической печати на башкирском языке // Актуальные проблемы языков народов России: Материалы XII региональной конференции. Уфа, 2012. С. 139–141.

7. Бускунбаева Л.А., Сиразитдинов З.А., Ибрагимова А.Д., Мигранова Л.Г., Полянин А.И. Башкирские языковые корпусы в Лаборатории лингвистики и информационных технологий // Урал и просторы Евразии сквозь века и тысячелетия. Уфа, 2013. С. 135–140.

8. Бускунбаева Л.А., Сиразитдинов З. А. О проблемах создания национального корпуса башкирского языка // Материалы Международной научно-теоретической конференции «Современное казахское языкознание: актуальные вопросы прикладной лингвистики», посвященная 75-летнему юбилею известного ученого профессора Жубанова Аскара Кудайбергенулы. Алматы, 2012. С. 54–58.

9. Бускунбаева Л.А., Сиразитдинов З.А. Система разметок в национальном корпусе башкирского языка / Языки меньшинств в компьютерных технологиях: опыт, задачи и перспективы. Йошкар-Ола, 2011, С. 45–51

10. Духовное наследие: фольклор свердловских башкир. – Уфа: ООО «Деловая династия», 2008. – С. 207–247. (на башк. яз.).

11. Корпусы Института этнологии и антропологии РАН http://corpora.iea.ras.ru/corpora/ (дата обращения: 07.06.2017).

12. Корпус нганасанских фольклорных текстов http://www.iling-ran.ru/gusev/Nganasan/ texts/index.php (дата обращения: 07.06.2017).

13. Куканова В.В. Фольклорный подкорпус: проблемы, структура и перспективы использования//Участие калмыков в укреплении российской государственности: материалы региональной научно-практической конференции Элиста: КИГИ РАН, 2012. С. 193–198

14. Надршина Ф.А. Русско-башкирский словарь пословиц эквивалентов. – Уфа: Китап, 2008. – 196 с.

15. Николаев Д.С. Создание электронного корпуса фольклорных текстов на русском языке// V социологическая Грушинская конференция, 13 марта 2015 https://wciom.ru/fileadmin/file/nauka/grusha2015/s2_6/Nikolaev.pdf. (дата обращения: 07.06.2017).

16. Нәҙершина Ф.А. Халыҡ һүҙе. Өфө, 1983.– 160 б.

17. Салчак А.Я. Электронный корпус текстов тувинского языка// Новые исследования Тувы (Электронный журнал). №3, 2012 URL: https://www.tuva.asia/journal/issue_15/ (дата обращения: 07.06.2017).

18. Сиразитдинов З.А., Бускунбаева Л.А., Барлыбаева А.Д., Ишмухаметова А.Ш. К разработке корпуса прозаических текстов башкирского языка с 1917 по 1940-е годы // Этногенез. История. Культура. I Юсуповские чтения: Материалы Международной научной конференции, посвященной памяти Рината Мухаметовича Юсупова. Уфа, 2011. С. 269–274.

19. Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Ибрагимова А.Д. Информационные системы и базы данных башкирского языка. Уфа: Книжная палата РБ, 2013. – 116 с.

20. Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Ибрагимова А.Д. О создании корпуса башкирского фольклора // Урал-Алтай: через века в будущее: Материалы VI Всероссийской тюркологической конференции (с международным участием). Уфа, 2014. С. 86–89.

21. Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш. Лингвистические корпусы Лаборатории лингвистики и информационных технологий ИИЯЛ УНЦ РАН // Проблемы изучения национальных литератур: Материалы международной научной конференции. Махачкала, 2015. С. 658–664.

22. Сиразитдинов З.А., Полянин А.И. Об опыте разработки интегрированной корпусной системы на базе СУДБ Оракл // Труды казанской школы по компьютерной и когнитивной лингвистике TEL–2014. Казань, 2014. С.85–88.

23. Sirazitdinov Z.A. The corpora of the bashkir language//Turklang 14 Proceedings of the International Scientific Conference. Istanbul, 2014. P. 125–129.

УДК 81'322.2

LINGUISTIC ANNOTATION OF GRAMMATICAL MOOD IN THE SAKHA LANGUAGE

G. Torotoev, A. Nogovitsyna
M.K. Ammosov North-Eastearn Federal University
torgav@mail.ru, Erkin2007@mail.ru

This article which is based on the works of turkologists and Yakut language scholars deals with the problem of linguistic annotation of Sakha language grammatical mood system, represented by 10 modal forms.

Keywords: annotation, tag system, Sakha language, morphemes, allomorphs, grammatical mood.

ПРОБЛЕМА АННОТИРОВАНИЯ ГРАММАТИЧЕСКИХ КАТЕГОРИЙ В ЯЗЫКЕ САХА (НА ПРИМЕРЕ НАКЛОНЕНИЙ ЯКУТСКОГО ГЛАГОЛА)

Г. Г. Торотоев, А. Н. Ноговицына
Северо-Восточный федеральный университет имени М.К. Аммосова
torgav@mail.ru, Erkin2007@mail.ru

В данной статье, основываясь на работах якутоведов и тюркологов, рассматривается проблема лингвистического аннотирования системы наклонений глагола якутского языка, представленной 10 модальными формами.

Ключевые слова: аннотирование, система тэгов, язык саха, морфемы, алломорфы, наклонения глагола.

Введение

Аннотирование словоформ – это очень трудоемкая работа, требующая от исследователя глубоких знаний в области теоретической и прикладной лингвистики. В компаративистике пристальное внимание уделяется плану выражения и плану содержания, иными словами, компаративистами учитывается и структурная (формальная) близость, и функционально-семантическое соответствие тех или иных грамматических категорий.

Важнейшим компонентом электронного корпуса любого языка является грамматическая разметка (система тэгов), позволяющая морфологическому анализатору автоматически обрабатывать лингвистические объекты в формализованном виде. С точки зрения типологии языков, тюркские языки относятся к агглютинативным языкам. В длинной цепочке каждый аффикс имеет свое определенное место, отличается определенной закономерностью "приклеивания" и функционально-семантической нагруженностью. Такая особенность тюркских языков дает большое преимущество в описании их морфологии в автоматическом режиме.

В ближайшей перспективе в сравнительно-сопоставительных исследованиях тюркских языков будет применен метод автоматического лингвистического анализа, что требует унификации систем грамматической разметки в корпусах тюркских языков. В 2014 г. во время проведения международной конференции по компьютерной и когнитивной лингвистике TEL-2014 из числа исследователей тюркских языков была сформирована Казанская рабочая группа по унификации системы грамматической разметки в корпусах тюркских языков. С тех пор ученые представляют результаты своих исследований – и теоретического, и прикладного характера – в научно-практическом семинаре UniTurk. Исходя из вышесказанного, при аннотировании грамматических категорий языка саха мы оперируем условными символами, используемыми в корпусах других тюркских языков.

Система наклонений якутского глагола

Наклонение как модальная форма глагола с давних пор притягивает к себе пристальное внимание многих исследователей-языковедов. Наклонения глагола в якутском языке впервые описаны академиком О.Н. Бетлингком в научном труде "О языке якутов" (1851), им установлено всего пять наклонений глагола: изъявительное (Indicativus), повелительное (Imperativus), возможное (Potentialis), условное (Conditionalis), совершенное (Perfectiv)[1].

Термин *киэп,* что в переводе с якутского на русский язык означает *наклонение,* в якутский язык введено языковедом и талантливым писателем А.А. Ивановым-Күндэ. В работе "Биһиги саҥабыт" (1925) он различает 6 наклонений якутского глагола: повелительное (соруйар киэп), предупредительное (сэрэтэр киэп),

желательное (баҕарар киэп), наклонение уверенности (эрэнэр киэп), просительное (көрдөһөр киэп), постепенное (сыыра-баара гынар киэп) [2]. В своем учебнике “Саха тыла. Кырамаатыка уонна бырааапсайдык суруйуу” (1934) он вносит следующую поправку: из вышеназванных 6 наклонений он оставляет только первых четырех, двух последних по каким-то причинам исключает из парадигмы наклонений [3]. Наклонения глагола нашли свое отражение в работах Н.Н. Поппе [4], Н.Н. Павлова [5], Н.С. Григорьева [6], Л.Н. Харитонова [7] и др.

Наклонения в якутском языке основательно исследованы доктором филологических наук Е.И. Коркиной в фундаментальном труде “Наклонения глагола в якутском языке”. Ею установлено, что в якутском языке существует десять форм глагольного наклонения, которые отличаются друг от друга морфологическими показателями и модальной семантикой: 1) изъявительное наклонение; 2) повелительное наклонение; 3) условное наклонение; 4) возможное наклонение; 5) утвердительное наклонение; 6) долженствовательное наклонение; 7) наклонение обычно совершаемого действия; 8) сослагательное наклонение; 9) предположительное наклонение; 10) наклонение несовершившегося (неосуществленного) действия [8]. Данная классификация одобрена всеми ведущими учеными-якутоведами и введена в активный научный оборот. [9].

Изъявительное наклонение

Изъявительное наклонение (*кэпсиир киэп*) в якутском языке, как и в других тюркских языках, не имеет специального грамматического маркера – аффиксов наклонения. Через формы индикатива выражается объективная модальность в аспектах настоящебудущего, прошедшего и будущего времен. Здесь уместно процитировать определение, данное доктором филологических наук В.И. Рассадиным: «Эта группа специальных форм глагола, выражающих действия/состояния во времени (прошедшем, настоящем, будущем) как раз и составляет особое наклонение, темпоральное (временное) наклонение, которое в грамматике принято называть индикативом или изъявительным наклонением. В связи с тем, что в нем больше выражена категория времени, а не модальности, то некоторые исследователи исключают его из состава наклонений» [10, С. 101].

Повелительное наклонение

Повелительное наклонение глагола (*соруйар киэп*) выражает различные степени модальности побуждения говорящего к действию: приказ, повеление, просьбу,призыв, пожелание и т.д.

В повелительном наклонении различают близкое будущее время и отдаленное будущее время. «Близкое будущее время повелительного наклонения имеет три лица, три числа: единственное, двойственное и множественное» [9, С. 320]. Основным отличием отдаленного будущего времени повелительного наклонения от близкого будущего времени является употребление его исключительно во 2 лице ед. и мн. ч. И в близком будущем, и в отдаленном будущем выделяют два варианта – основной и усилительно-просительный.

В.А. Плунгян отмечает, что «для тюркских или дагестанских языков типично морфологическое противопоставление императива, юссива и гортатива» [11, С. 319]. Для обозначения императивных конструкций мы будем оперировать следующими терминами: *гортатив* для 1 лица, собственно *императив* для 2 лица и *юссив* для 3 лица. Кроме того, для обозначения двойственного числа якутского языка нами предлагается помета Dual, она использована А.В. Дыбо при аннотировании морфологической системы хакасского языка.

Для различения императива близкого будущего и отдаленного будущего времени нами использованы тэги IMP – для близкого будущего времени, а IMP2 – для отдаленного будущего времени, таким образом, условный символ IMP2 будет подчеркивать использование отдаленного будущего времени исключительно во 2 лице ед.ч. и мн. ч. Усилительно-просительная форма императива отдаленного будущего времени обозначена символом PREC. Форма прекатива на –ыый обозначена PREC_SG, т.к. она обозначает единственное число: *бар=аар=ыый>бараарыый* ‘пойди же (потом)’, форма прекатива на –ҥытыый помечена тэгом PREC_PL, так как здесь наличествует значение множественности: *бар=аар=ыҥ=ҥыт(=)ыый > бараарыҥҥытыый* ‘пойдите же (потом)’.В «Грамматике современного якутского литературного языка. Фонетика и морфология» прекатив во множественном числе отражен морфемами -ытыый/-ытыныый, вероятнее всего, в данном случае структура глагола представлена неадекватно, по-

скольку сложный аффикс не выдерживает критики при объективном компонентном анализе. Якутский литературный язык требует обязательное использование в составе глаголов с усилительно-просительным значением удвоенного согласного [ҥҥ], например: *кэлиҥҥитиий, кэлээриҥҥитиий*. В данном случае мы можем наблюдать за эволюцией морфемы -ҥытыый, не вызывает сомнения тот факт, что сложный аффикс образован из 2 самостоятельных морфем -ҕыт и -ыый.

Таблица 1. **Морфологическое аннотирование повелительного наклонения якутского глагола**

Сокращения	Расшифровка	Название категории	Алломорфы	Морфемы
HOR.SG	Hortative (1st person singular)	Императив 1 лица ед.числа(гортатив)	-ыым/-иим/ -үүм/-уум	-ЫЫм
HOR.PL	Hortative (1st person plural)	Императив 1 лица мн.числа(гортатив)	-ыаҕыҥ/ -иэҕиҥ/ -үөҕүҥ/ -уоҕуҥ -ыаххайыҥ/ -иэххэйиҥ/ -үөххэйиҥ/ -уоххайыҥ	-ЫАҕЫҥ -ЫАххАйЫҥ
IMP.SG	Imperative (2nd person singular)	Императив близкого будущего времени 2 лица ед.числа	–	–
IMP.PL	Imperative (2nd person plural)	Императив близкого будущего времени 2 лица мн.числа	-ыҥ/-иҥ/ -уҥ/-үҥ	-Ыҥ
IMP.DUAL	Imperative (dual)	Императив близкого будущего временидвойственного числа	-ыах/-иэх/ -уох/-үөх	-ЫАх
IMP2.SG	Imperative2 (2ndperson singular)	Императив отдаленного будущего времени 2 лица ед.числа	-аар/-ээр/ -өөр/-оор	-ААр

IMP2.PL	Imperative2 (2nd person plural)	Императив отдаленного будущего времени 2 лица мн.числа	-аарын/ -ээриҥ/ -өөрүҥ/ -ооруҥ	-ААрЫҥ
JUS.SG	Jussive (3rd person singular)	Императив 3 лица ед.числа(юссив)	-дын/-дин/ -дун/-дүн -тын/-тин/ -тун/-түн -лын/-лин/ -лун/-лүн -нын/-нин/ -нун/-нүн	-ДЫн
JUS.PL	Jussive (3rd person plural)	Императив 3 лица мн.числа(юссив)	-дыннар/ -диннэр/ -дуннар/ -дүннэр -тыннар/ -тиннэр/ -туннар/ -түннэр -лыннар/ -линнэр/ -луннар/ -лүннэр -ныннар/ -ниннэр/ -нуннар/ -нүннэр	-ДЫннАр
PREC.SG	Precative (2ndperson singular)	Просительный императив (прекатив) 2 лица ед.числа	-ыый/-ий/ -ууй/-үүй	-ЫЫй
PREC.PL	Precative (2nd person plural)	Просительный императив (прекатив) 2 лица мн.числа	-ҥытыый/ -ҥитиий/ -ҥүтүүй/ -ҥутууй	-ҥЫтЫЫй

Рассмотрим примеры аннотирования:
Барыым ‘Я пойду-ка’
‘go’=HOR.SG
Бардыннар‘Пусть они уходят’
‘go’=JUS.PL

В данной статье все наклонения якутского глагола представлены исключительно в положительной форме. При глоссировании наклонений в отрицательной форме впереди формулы ставится помета NEG, например:

Барымаарыый ‘Ты не уходи, пожалуйста’
‘go’=NEG_IMP2.SG=PREC.SG

Условное наклонение

Условное наклонение в якутском языке (*болдьуур киэп*) выражает модальное значение условия, предпосылки для совершения другого действия. Кондиционалис в якутском языке представлен аффиксом -тар и моделью -тах+Poss+на, которые непосредственным образом присоединяются к основе глагола. Параллелью аффикса -тар в тюркских языках выступает условное наклонение на -са/-за. «Вторая форма условного наклонения – чисто якутская форма, не имеющая в плане наклонений параллелей в других тюркских языках” [9, С.326]. Данная модель состоит из 3 компонентов: аффикс -тах=аффикс принадлежности=реликтовый аффикс -на.

Таблица 2. **Морфологическое аннотирование условного наклонения якутского глагола**

Сокращения	Расшифровка	Название категории	Алломорфы	Морфемы
COND	Conditional	Кондиционалис	-тар/-тэр/-тор/-төр -дар/-дэр/-дор/-дөр -лар/-лэр/-лор/-лөр -нар/-нэр/-нор/-нөр	-TAp
			-тах=Poss=на/-тэх=Poss=нэ/ -тох=Poss=на/-төх=Poss=нэ -дах=Poss=на/-дэх=Poss=нэ/ -дох=Poss=на/-дөх=Poss=нэ -лах=Poss=на/-лэх=Poss=нэ/ -лох=Poss=на/-лөх=Poss=нэ -нах=Poss=на/-нэх=Poss=нэ/ -нох=Poss=на/-нөх=Poss=нэ -	-TAx= Poss= нA

Утвердительное наклонение

Утвердительное наклонение (*бигэргэтэр киэп*) впервые было описано академиком О.Н. Бетлингком в работе “О языке якутов” (1851) под названием Perfectiv. Он образуется путем присоединения к глагольной основе морфемы -ыıһы и аффиксов сказуемости. «Основное модальное значение глаголов данного наклонения состоит в том, что ими на основе определенных примет, признаков или оценки и обобщения конкретной ситуации говорящим лицом выражается несомненная уверенность в возможности действия, которое должно иметь место после момента речи» [8, С. 204].

Для аннотирования утвердительного наклонения, в поисках оптимального решения данной проблемы, нами рассмотрены два варианта:

– ASS (Assertive) от англ. *настойчивый*, *утвердительный*;

– AFFR (Affirmative) от англ. *позитивный, утвердительный*.

Affirmative используется в ряде языков для обозначения утвердительной формы наклонений, например, в испанском языке используется Imperativo afirmativo (утвердительная форма повелительного наклонения).

Категория ассертива выражает эпистемическую уверенность, иными словами, говорящий с большой уверенностью утверждает о том, что действие будет реализовано в недалеком будущем. Как отмечает В.А. Плунгян, «Эпистемические наклонения (возникающие на базе показателей возможности и необходимости) выражают различные виды эпистемической оценки: эпистемическую невозможность, или сомнение (дубитатив), эпистемическую возможность, или вероятность (пробабилитив), эпистемическую необходимость, или уверенность (ассертив); адмиратив, как уже отмечалось, чаще имеет тенденцию выражаться совместно с показателями эвиденциальности» [11, С. 317]. С этой точки зрения, нам кажется, что тэг ASS наиболее точно отражает модальное значение глаголов утвердительного наклонения. В разрабатываемой сравнительной таблице UniTurk данное наклонение отсутствует, согласно Е.И. Коркиной, «утвердительное по своей модальной семантике наклонение глагола из живых тюркских языков имеется в одном только якутском языке» [8, С. 207].

Таблица 3. **Морфологическое аннотирование утвердительного наклонения якутского глагола**

Сокращения	Расшифровка	Название категории	Алломорфы	Морфемы
ASS	Assertive	Ассертив	-ыыһы/-иинһи/-үүһү/ -уухуу -ыыһык/-иинһик/-үүһүк/-уухук	-ЫЫІhЫ -ЫЫІhЫк

Долженствовательное наклонение

Долженствовательное наклонение (*буолуохтаах киэп*) представлено тремя плоскостями времени: настояще-будущим, будущим и прошедшим.

Настояще-будущее время долженствовательного наклонения образуется от вторичного причастия на –ардаах и аффикса сказуемости. Будущее время долженствовательного наклонения образуется от вторичного причастия на -ыахтаах и аффикса сказуемости. Прошедшее время представляет собой аналитическую форму, образуемую от вторичного причастия на –ыахтаах и спрягаемой формы недостаточного глагола э-.

Для обозначения долженствовательного наклонения якутского языка используется помета OBL (Obligative). Данный тэг использован татарскими учеными при аннотировании модальных форм глагола, выражающих значение необходимости. В татарском языке «формы на -асы и -асы иде, выступая в предикативной функции и в личном значении, выражают долженствование: а) Буарады митинг буласы бит» (Г.Колəхметов) «На днях ведь должен быть митинг»; б) Аларөйгə кереп китүгə, без бакча янына сызасы идек (Г. Бəширов) “Как только они войдут в дом, мы должны были удрать к саду» [12, С. 147].

Для обозначения вспомогательных слов предлагаем использование условного сокращения AUX от английского Auxiliary Verbs (вспомогательные глаголы) и при этом учитываем спряжение, например:

Барыахтаах этим ‘я должен (обязан) был идти (пойти)’
‘go’=OBL+AUX.1SG

Таблица 4. **Морфологическое аннотирование долженствовательного наклонения якутского глагола**

Сокращения	Расшифровка	Название категории	Алломорфы	Морфемы
OBL	Obligative	Облигатив	-ардаах/- эрдээх/ -ордоох/-өрдөөх -ыахтаах/-иэхтээх/ -уохтаах/-үөхтээх -ыахтаах э-/-иэхтээх э-/-уохтаах э-/ - үөхтээх э- -ыах/-иэх/-уох/-үөх тустаах- -ыах/-иэх/-уох/-үөх кэриҥнээх-	-АрдААх -ЫАхтААх -ЫАхтААх э- -ЫАх тустаах- -ЫАх кэриҥнээх-

Наклонение обычно совершаемого действия

Наклонение обычно совершаемого действия (*үгэс киэбэ*) образуется путем присоединения к глагольной основе морфемы –ааччы и аффикса сказуемости. “Модальное значение данного наклонения состоит в том, что оно выражает действия, совершающиеся с точки зрения говорящего лица обычно, регулярно, более или менее постоянно в силу проявления имманентных самому субъекту свойств, постоянной способности, привычки, склонности, пристрастия его к совершению действий такого рода” [8, С. 333].

Повторяемость действия или итеративность изучается в рамках современной аспектологии. Как отмечает Е.М. Самсонова, “помимо специальных форм с аффиксами многократности, составляющих ядро функционально-семантического поля итеративности, имеются и другие синтетические репрезентанты, которые в той или иной степени используются для выражения данной семантики. Например, форма наклонения обычно совершаемого действия («үгэс киэп»)” [13].

В современной аспектологии регулярно повторяющиеся ситуации, «привычные» действия, становящиеся характеристиками свойств субъекта принято называть хабитуалисом.

Для глоссирования наклонения обычно совершаемого действия в якутском языке нами предлагается помета HAB (Habitualis).

А.В. Дыбо для отображения хабитуального причастия хакасского языка использует тэг PrtHab. При этом необходимо отметить, что между якутским глаголом на -ааччы и хакасским причастием на -ааччы/-еечі можно провести параллель, однако в якутском языке аффикс -ааччы «полностью сохранило свое глагольное действие и легло в основу наклонения обычности», а в хакасском языке он «рассматривается как форма, утратившая глагольные признаки, полностью перешедшая в разряд имен прилагательных и существительных, обозначающих действие в качестве постоянного признака предмета, а также наименование деятеля (действующего лица)» [8, С. 224].

Таблица 5. **Морфологическое аннотирование наклонения обычно совершаемого действия якутского глагола**

Сокращения	Расшифровка	Название категории	Алломорфы	Морфемы
HAB	Habitualis	Хабитуалис	-ааччы/-ээччи/ -өөччү/-ооччу -ааччык/-ээччик/ -өөччүк/-ооччук	-АА**ччЫ** -АА**ччЫк**

Возможное наклонение

Возможное наклонение (*сэрэтэр киэп)* образуется путем присоединения к глагольной основе морфемы –аайа и аффикса сказуемости.

Л.Н. Харитонов в якутском языке возможное наклонение передает термином саарбаҕалыыр киэп. Об этом в своей работе упоминает Е.И. Коркина: «…безусловный шаг вперед представляет учебник Л.Н. Харитонова. Им зарегистрировано и выявлено шесть наклонений: изъявительное (кэпсиир киэп), повелительное (соруйар киэп), возможное (саарбаҕалыыр киэп), условное (усулуобунай киэп), утвердительное (бигэргэтэр киэп) и предположительное (сэрэйэр киэп)» [8, С. 14-15]. Далее о вариантах данного термина исследователь пишет следующее: “Название «возможное» или potentialis этого наклонения впервые в якутоведение было введено Бетлингком и вслед за ним повторено Радловым, С.В. Ястремским, Л.Н. Харитоновым. Н.Н. Поппе, Г.У. Гермогенов и И.Д. Моруо называли это наклонение формой опасения, а А.А.

Иванов-Кюндэ и Н.С. Григорьев – предупредительным наклонением» [8, с. 230].

При этом тюркологи отмечают родство возможного наклонения якутского языка с желательным наклонением, существующим во всех тюркских языках, однако «желательная семантика данной формы в значительной степени переосмыслена и приобрела другие оттенки значений» [8, С. 233]. Желательное наклонение в тюркских языках образуется посредством форманта на -yaj/-gej~ - qaj/-kej [14, С. 330].

Для аннотирования возможного наклонения нами рассматривались два варианта: во-первых, POT от Potentialis, использованного О.Н. Бетлингком; во-вторых, PREM (Premonitive), используемый в татарском корпусе тэг, выражающий значение предостережения.

Со значением предостережения, боязни какого-либо действия, нежелательности его совершения в татарском языке используется форма на -гай и при этом всегда следует за отрицательной формой. «Икенче яктан, бу вакытта тырмалау үзе бик хәтәр нәрсә. Бодайның тамыры *өзелмәгәе*, йолкынып *чыкмагае*. (Г. Бәширов) «Бороновать-то можно… С другой стороны, бороновать в это время очень опасно. Как бы не повредить корень пшеницы, как бы его не выдернуть» [12, С. 114].

Учитывая модальные значения возможного наклонения в якутском языке, в своих будущих исследованиях мы будем обращаться к условному символу PREM (Premonitive).

Таблица 6. **Морфологическое аннотирование возможного наклонения якутского глагола**

Сокращения	Расшифровка	Название категории	Алломорфы	Морфемы
PREM	Premonitive	Премонитив	-аайа/-ээйэ/-өөйө/-оойо -аайык/-ээйик/-өөйүк/-оойук	-ААйА -ААйЫк

При создании лингвистической базы данных и совершенствовании системы аннотирования всплывают интересные языковые факты и исключения из правил. Так в якутском языке 3 лицо ед. ч. возможного наклонения в виде исключения представлена формой *-аарай*:

1 л., ед.ч. - бар-аайа-бын ‘авось пойду’
2 л., ед.ч. - бар-аайа-ҕын ‘авось пойдешь’
3 л., ед.ч. - бар-**ааpай**‘авось пойдет’
1 л., мн.ч. - бар-аайа-быт ‘авось пойдем’
2 л., мн.ч. - бар-аайа-ҕыт ‘авось пойдете’
3 л., мн.ч. - бар-аайа**л**-лар‘авось пойдут’

Мы предполагаем, что в структуре лексемы, который представлен в форме 3 лица единственного числа, в процессе эволюции произошла позиционная фонетическая трансформация, в результате которого *бараайар* превратился в *бараарай*. В первоначальной форме четко прослеживаются морфемы -аайа и -ар: бар=аа**й**а=а**р**, что с точки зрения морфонологии, вполне естественно и закономерно. Этим можно и объяснить трансформацию фонемы **р** в **л** в форме 3 лица множественного числа (бар=аайа=ар=лар) под влиянием регрессивной фонетической ассимиляции.

Наклонение несовершившегося (неосуществленного) действия

Наклонение несовершившегося (неосуществленного) действия (*буола илик хайааһын киэбэ*) в якутском языке относится к числу аналитических форм. Она образуется путем сочетания деепричастия на -а/-ыы и форманта *илик*. С точки зрения темпоральности различают два времени: настоящее и прошедшее время.

Для обозначения наклонения несовершившегося (неосуществленного) действия мы будем использовать тэг CUNC (Cuncative). Согласно А.В. Дыбо, «Cunc – кункатив, еще не совершившееся действие» [15: С.131]. Форма еще не совершившегося действия в хакасском языке соответствует наклонению несовершившегося (неосуществленного) действия в якутском языке. Как отмечает Е.И. Коркина «формант *илик* и его параллели *элек, гелек* (*галак, келек, калак*) зарегистрированы не во всех тюркских языках, а только в ряде тюркских языков Сибири (хакасском, алтайском, шорском, чулымском, тувинском)» [8, С. 238]. Семантика вышеуказанных параллелей одинакова, обозначает еще не совершившееся действие на момент речи, при этом в тувинском и хакасском языках отмечается оттенок ожидаемого действия. Так, исследователь хакасского языка Н.П. Дыренкова отмечает, что «Форма причастия несовершенного выражает действие, еще не совершив-

шееся, но ожидаемое; состояние, еще не наступившее, но ожидаемое… Форма причастия несовершенного образуется путем присоединения аффикса халах-галах, келек-гелек» [16, С. 78].

Таблица 7. **Морфологическое аннотирование наклонения несовершившегося (неосуществленного) действия якутского глагола**

Сокращения	Расшифровка	Название категории	Алломорфы	Морфемы
CUNC	Cuncative	Кункатив	-а/-о/-э/-ө илик -ыы/-ии/-уу/-үү илик	-А илик -ЫЫ илик

Сослагательное наклонение

Сослагательное наклонение (*буолуон сөптөөх хайааһын киэбэ*) представлено несколькими аналитическими формами:

1) Основная форма сослагательного наклонения образуется путем присоединения к глагольной основе первичного причастия на -ыах и формы прошедшего категорического времени от недостаточного глагола э-.

2) Вторая форма сослагательного наклонения образуется сочетанием причастия на -ыах/-ыа и спрягаемой формы *эбит*.

3) Третья форма сослагательного наклонения образуется путем присоединения к основе глагола первичного причастия на –ар и вспомогательного глагола э- [9, С.337-340].

Для аннотирования сослагательного наклонения нами были рассмотрены две версии:

– SUBJ (Subjunctive) – сослагательное наклонение в английском языке;

– CONJ (Conjuctivus) – конъюнктив или сослагательное наклонение в немецком языке.

И конъюнктив, и субъюнктив в полной мере, емко отражают значение сослагательного наклонения. При этом отмечаем, что конъюнктив является исторически сложившимся наименованием, так в сравнительно-исторической грамматике тюркских языков отмечено, что «для тюркских языков характерно наличие вполне сформировавшихся следующих трех основных категорий косвенных наклонений, противопоставленных изъявительному: 1) желательного (Optativus), 2) повелительно-побудительного (Impertivus-Jussivus), 3) условного (Condicionalis) и двух вторич-

ных косвенных наклонений: 4) долженствовательного (Devitivus) и 5) сослагательного (Conjuctivus)» [14, С.329].

Таблица 8. **Морфологическое аннотирование сослагательного наклонения якутского глагола**

Сокращения	Расшифровка	Название категории	Алломорфы	Морфемы
CONJ	Conjuctivus	Конъюнктив	-ыах/-иэх/-уох/-үөх э- -ыа/-иэ/-уо/-үө э- -ыах/-иэх/-уох/-үөх *эбит* -ыа/-иэ/-уо/-үө *эбит* -ар/-эр/-ор/-өр э- -ар/-эр/-ор/-өр *эбит*	-ЫАх э- -ЫА э- -ЫАх *эбит* -ЫА *эбит* -Ар э- -Ар *эбит*

Предположительное наклонение

В основе предположительного наклонения (*сэрэйэр киэп*) лежит причастие на -тах. При спряжении в положительной форме к нему присоединяются аффиксы принадлежности, например, *бар-даҕ-ым*, а при отрицательном спряжении приклеиваются аффиксы отрицания и принадлежности плюс модальная частица *буолуо*, например, *бар-батаҕ-ым буолуо*. [9].

Для обозначения предположительного наклонения мы будем использовать тэг ASSUM (Assumptive), использованный А.В. Дыбо в корпусе хакасского языка: «Assum – ассумптив, предположительное наклонение, вводится оборотом «похоже, что …» [15]. Предположительное наклонение в хакасском языке образуется путем присоединения к основе глагола аффиксов -ғадағ/-гедек, -хадағ/-кедег, -адағ/-едег. «Форма предположительного наклонения на -ғадағ имеется, главным образом, в языках уйгурской группы (тувинском, шорском, уйгурском, якутском, хакасском) почти с одинаковым значением, обозначая предполагаемые действия: говорящий на основе своих наблюдений и опыта делает умозаключение о возможности, вероятности или невозможности, недопустимости совершения действия в будущем самим или другими лицами, предметами» [17: С. 198-199].

Аффикс -ғадағ является сложным аффиксом, состоящим из двух компонентов. «Первый из них, по мнению многих исследователей тюркских языков, восходит к древнетюркскому при-

частию будущего времени с аффиксом -гу/-гы… Вторая часть его восходит к общеизвестному тюркскому аффиксу уподобления -дағ/-деғ…» [17, С. 198].

Таблица 9. **Морфологическое аннотирование предположительного наклонения якутского глагола**

Сокращения	Расшифровка	Название категории	Алломорфы	Морфемы
ASSUM	Assumptive	Ассумптив	-тах/-тэх/-тох/-төх/ -дах/-дэх/-дох/-дөх -лах/-лэх/-лох/-лөх/ -нах/-нэх/-нох/-нөх/	-ТАх

Заключение

Лингвистическое аннотирование текстов является одним из наиболее актуальных направлений современной компьютерной лингвистики. Разметка корпусов нужна как для получения собственно лингвистических результатов, так и для обработки естественного языка (Natural Language Prosessing, NLP).

Таблица 10. **Лингвистическое аннотирование наклонений якутского глагола**

Сокращения	Расшифровка	Название категории	Алломорфы	Морфемы
IND	Indicative	Изъявительное наклонение	*не имеет специального грамматического маркера - аффиксов наклонения*	*темпоральные аффиксы*
IMP	Imperative	Повелительное наклонение	-ыым/-иим/-үүм/-уум -аар/ -ээр/ -өөр/ -оор -дын/-дин/-дун/-дүн -тын/-тин/-тун/-түн -лын/-лин/-лун/-лүн -нын/-нин/-нун/-нүн -ыах/-иэх/-уох/-үөх -ыаҕын/-иэҕин/-үөҕүн/ -уоҕуң -ыаххайың/-иэххэйин/ -үөххэйин/-уоххайың	-ЫЫм -ААр -ДЫн -ЫАх -ЫАҕЫң -ЫАххАйЫң

			-ын/-ин/-ун/-үн -аарын/-ээрин/-өөрүн/ -оорун -дыннар/-диннэр/-дуннар/ -дүннэр -тыннар/-тиннэр/ -туннар/-түннэр -лыннар/ -линнэр/-луннар/-лүннэр -ныннар/-ниннэр/-нуннар/ -нүннэр	-Ын -ААрЫн -ДЫннАр
CONJ	Conjuc-tivus	Сослага-тельное накло-нение	-ыах/-иэх/-уох/-үөх э- -ыа/-иэ/-уо/-үө э- -ыах/-иэх/-уох/-үөх *эбит* -ыа/-иэ/-уо/-үө *эбит* -ар/-эр/-ор/-өр э- -ар/-эр/-ор/-өр *эбит*	-ЫАх э- -ЫА э- -ЫАх *эбит* -ЫА *эбит* -Ар э- -Ар *эбит*
ASS	Assertive	Утверди-тельное накло-нение	-ыыһы/-ииһи/-үүһү/-ууһу -ыыһык/-ииһик/-үүһүк/- ууһук	-ЫЫһЫ -ЫЫһЫк*
HAB	Habitualis	Наклоне-ние обыч-но совер-шаемого действия	-ааччы/-ээччи/-өөччү/ -ооччу -ааччык/-ээччик/-өөччүк/ -ооччук	-ААччЫ -ААччЫк*
PREM	Premo-nitive	Возмож-ное накло-нение	-аайа/-ээйэ/-өөйө/-оойо -аайык/-ээйик/-өөйүк/ -оойук	-ААйА -ААйЫк*
ASSUM	Assum-ptive	Предпо-ложитель-ное накло-нение	–тах/-тэх/-тох/-төх -дах/-дэх/-дох/-дөх	–ТАх
COND	Condi-tional	Услов-ное накло-нение	-тар/-тэр/-тор/-төр -дар/-дэр/-дор/-дөр -лар/-лэр/-лор/-лөр -нар/-нэр/-нор/нөр -тахпына/-тэхпинэ/ -лахпына/-лэхпинэ -дахпына/-дэхпинэ/ -дохпуна/-дөхпүнэ	-ТАр -ТАх= Poss=нА

			-таххына/-тэххинэ -лаххына/-лэххинэ -даххына/-дэххинэ/ -доххуна/-дөххүнэ-таҕына/ -тэҕинэ -лаҕына/-лэҕинэ -даҕына/-дэҕинэ/-доҕуна/ -дөҕүнэ -тахпытына/-тэхпитинэ -лахпытына/-лэхпитинэ -дахпытына/-дэхпитинэ/ -дохпутуна/ -дөхпүтүнэ -таххытына/-тэххитинэ -лаххытына/-лэххитинэ -даххытына/-дэххитинэ/ -доххутуна/ -дөххүтүнэ -тахтарына/-тэхтэринэ -лахтарына/-лэхтэринэ -дахтарына/-дэхтэринэ/ -дохторуна/ -дөхтөрүнэ	
OBL	Obliga-tive	Должен-ствова-тельное накло-нение	-ардаах/- эрдээх/-ордоох/ -өрдөөх -ыахтаах/-иэхтээх/-уохтаах/ -үөхтээх -ыахтаах э-/-иэхтээх э-/ -уохтаах э-/- үөхтээх э- -ыах/-иэх/-уох/-үөх тустаах- -ыах/-иэх/-уох/-үөх кэриҥнээх-	-АрдААх -ЫАхтААх -ЫАхтААх э- -ЫАх тустаах- -ЫАх кэриҥнээх-
CUNC	Cunca-tive	Накло-нение несовер-шивше-гося (неосу-ществ-ленного) действия	-а/-э/-о/-ө илик -ыы/-ии/-үү/-уу илик	-А илик -ЫЫ илик

* Используется в разговорной речи.

Как видно из таблицы, в двух случаях, а именно в повелительном и условном наклонениях, грамматические показатели лица и глагольной модальности сливаясь, образуют сложную морфему, а в других случаях аффиксы лица следуют за морфологическими показателями модальных форм глагола.

Авторы статьи в последние годы работают над проблемой лингвистического аннотирования грамматических категорий якутского языка и в течение 2 лет тесно сотрудничают с кандидатом технических наук, доцентом ФТИ СВФУ Ньургуном Анатольевичем Леонтьевым, участвуют в разработке компьютерной программы “Морфологический анализатор якутского языка”.

В данной статье предпринята попытка аннотирования системы наклонений якутского глагола, и в результате исследования за 10 модальными формами якутского глагола закреплены соответствующие условные символы – тэги (синтетические – IND, IMP, ASS, HAB, PREM, ASSUM, COND; аналитические – CONJ, CUNC; синтетико-аналитические – OBL). Это необходимо для того, чтобы морфологический анализатор якутского языка адекватно распознавал все грамматические категории языка, отраженные в электронном корпусе якутского языка. Материал данного исследования может быть актуален в сравнительно-сопоставительных исследованиях алтайских языков.

ЛИТЕРАТУРА

1. Бетлингк О.Н. О языке якутов / Пер. с нем. Рассадин В.И. – Новосибирск: Наука. Сиб. Отд-ние, 1990. – 646 с.

2. Күндэ (Ө.Ө.Уйбаныап). Биһиги саҥабыт. М., 1925.

3. Күндэ (Ө.Ө.Уйбаныап). Саха тыла. Кырамаатыка уонна бырааапсайдык суруйуу. Дьокуускай, 1934.

4. Поппе Н.Н. Учебная грамматика якутского языка. М., Центриздат, 1926.

5. Баабылап Нь.Нь. Саха тылын кырамаатыката. Пэниэтикэ, морполуогуйа. Дьокуускай, 1935.

6. Киргиэлэйэп М.И. Саха тылын кырамаатыката. Ситэтэ суох отто оскуолаҕа үөрэнэр кинигэ. М., 1938.

7. Харитонов Л.Н. Саха тылын грамматиката. САССР госиздата. Якутскай, 1942.

8. Коркина Е.И. Наклонения глагола в якутском языке. М.: Наука, 1970. – 308 с.

9. 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.

10. Рассадин В.И. Очерки по морфологии и словообразованию монгольских языков. – Элиста: Изд-во КалмГУ, 2011 г. – 240 с.

11. Плунгян В. А. Общая морфология. – Москва: Едиториал УРРС, 2003. – 374 с.

12. Тумашова Д.Г. Татарский глагол (Опыт функционально-семантического исследования грамматических категорий). – Казань: изд-ва Казанского университета, 1986.

13. Самсонова Е.М. Наклонение обычно совершаемого действия в якутском языке: особенности семантики и функционирования // Филологические науки. Вопросы теории и практики. 2016. № 5 (59): в 3-х ч. Ч. 3. С. 139–142.

14. Сравнительно-историческая грамматика тюркских языков. Морфология. – Москва: Наука, 1988. – 549 с.

15. Дыбо А.В., Шеймович А.В. Автоматический морфологический анализ для корпусов хакасского и древнетюркского языков // Филология. Научное обозрение Саяно-Алтая. 2014. № 2(08).

16. Дыренкова Н.П. Грамматика хакасского языка. Фонетика и морфология. – Абакан: Хакасское областное национальное издательство, 1948. – 122 с.

17. Грамматика хакасского языка / Под ред. Баскакова Н. А. – М.: Наука, 1975.

УДК 81’32

AN ASSESSMENT OF UNIVERSAL DEPENDENCY ANNOTATION GUIDELINES FOR TURKIC LANGUAGES

F.M. Tyers[1], J. Washington[2], Ç. Çöltekin[3], A. Makazhanov[4]
[1] School of Linguistics, Higher School of Economics, Moscow;
[2] Linguistics Department, Swarthmore College, Swarthmore;
[3] Seminar für Sprachwissenschaft, Universität Tübingen;
[4] National Laboratory Astana, Nazarbayev University, Astana
jonathan.washington@swarthmore.edu

Annotated corpora of three Turkic languages – Turkish, Kazakh, and Uyghur – were released as part of version 2 of the Free/Open-Source Universal Dependencies (UD) syntactic and morphological annotation guidelines. The objective of these guidelines is to provide consistent dependency annotation to facilitate cross-linguistic comparison.

This paper presents the current state of each of the three UD-annotated Turkic corpora, along with an evaluation of the performance of parsers trained on these corpora.

Overall, the UD annotation guidelines for Turkish, Kazakh, and Uyghur are fairly compatible – a testament to the careful design of the guidelines. However, the specific annotation guidelines for each of these languages were developed mostly independently; because of this, differences between the three standards exist. Moving forward with Turkic annotation standards in UD, attempts will be made to reconcile the differences. These differences are overviewed in this paper.

Furthermore, a number of issues in annotation have arisen and have yet to be resolved. Some of these issues require further investigation of the phenomena, and some require consultation within the UD community to determine whether solutions may be determined based on similar phenomena in other languages. A number of these open issues are discussed, including tokenisation (how to deal with words that include an orthographic space, or multiple words that do not include an orthographic space), the difference between core and oblique arguments of verbs, complex predicates (including structures where there is a combination of a non-finite form which governs argument structure and contributes to TAM and a finite-form which contributes to TAM and takes person agreement), multiple derivation (multiple causative or causative–passive combinations), and use of copulas instead of auxiliaries in what appear to be auxiliary constructions.

Keywords: Turkish; Kazakh; Uyghur; treebank; dependency grammar; Universal Dependencies.

ОЦЕНКА КРИТЕРИЕВ МОРФО-СИНТАКСИЧЕСКОЙ РАЗМЕТКИ ДЛЯ ТЮРКСКИХ ЯЗЫКОВ В ПРОЕКТЕ «UNIVERSAL DEPENDENCIES»

F.M. Tyers[1], J. Washington[2], Ç. Çöltekin[3], A. Makazhanov[4]

[1] Школа лингвистики, Высшая школа экономики, Москва;
[2] Департамент лингвистики, Суортмор-колледж, Суортмор;
[3] Школа лингвистики, Тюбингенский университет;
[4] Национальная Лаборатория Астана, Назарбаев Университет, Астана

jonathan.washington@swarthmore.edu

Аннотированные корпусы трех тюркских языков – турецкого, казахского и уйгурского – были выпущены в составе второй версии проекта «Universal Dependencies», предоставляющего свободно распространяемые рекомендаций к универсальной морфо-синтаксической разметке. Целью этих рекомендаций является предоставление единой схемы разметки для упрощения межязыкового анализа.

В настоящей работе описано текущее состояние каждого из трех тюркских корпусов размеченных по принципам UD, а также оценка эффективности синтаксических парсеров, обученных на этих корпусах.

Схемы разметки UD для турецкого, казахского и уйгурского языков во многом совместимы, что свидетельствует о тщательной проработке универсальности принципов UD. Однако конкретные рекомендации по аннотации для каждого из этих языков разрабатывались в основном независимо; из-за этого существуют различия между тремя стандартами. При дальнейшей разработке схем разметки для тюркских языков будут предприняты попытки сгладить данные различия, которые также рассмотрены в данной работе.

Кроме того, возник ряд вопросов по разметке определенных конструкций. Ответы на некоторые из этих вопросов требуют дальнейшего изучения природы соответствующих явлений в языке. В других случаях ответы могут быть получены на основе анализа схожих явлений в других языках; для этого потребуются консультаций с членами сообщества UD. В данной работе обсуждаются ключевые вопросы разметки, в частности: токенизация (считать ли слова, включающие в себя орфографические пробелы отдельными единицами разметки, и наоборот, разбивать ли несколько синтаксических слов, являющихся частью одного орфографического, на отдельные единицы разметки); разница между актантами и сирконстантами; сложные предикаты (включая структуры, где существует комбинация нефинитной формы, которая управляет аргументной структурой и несет временные и

аспектно-модальные функции, и финитной формы, которая также несет временные и аспектно-модальные функции и принимает личное окончание); множественная деривация (комбинации из нескольких каузативов и каузативов-пассивов); использование копулы вместо вспомогательного глагола в конструкциях, напоминающих сочетание главного и вспомогательного глаголов.

Ключевые слова: Турецкий язык; Казахский язык; Уйгурский язык; грамматика зависимостей; проект «Universal Dependencies».

1. Introduction

Universal Dependencies (UD, Nivre et al. 2016) is a Free/Open-Source set of guidelines for syntactic and morphological annotation of corpora, which aims to provide consistent dependency annotation to facilitate cross-linguistic comparison. In addition to the guidelines, annotated corpora are made available under a Free/Open-Source license.

This paper overviews recent work that has gone into making UD annotation guidelines for Turkic languages based on the UD standard. The current status of UD-annotated corpora of Turkic languages is overviewed in section 2. Three separate efforts have resulted in fairly compatible guidelines for Turkish (southwestern Turkic, §2.1), Kazakh (northwestern Turkic §2.2), and Uyghur (southeastern Turkic, §2.3), which is a testament to the careful design of the UD guidelines. However, because the specific annotation guidelines for each of these languages were developed mostly independently, differences between them exist, as described in section 2.5. Moving forward with Turkic annotation standards in UD, attempts will be made to reconcile the existing differences. In addition to efforts with these three languages and annotation standards, annotated corpora of some other Turkic languages have been begun as well (§2.4).

Furthermore, a number of issues in annotation have arisen and have yet to be resolved. Some of these issues require further investigation of the phenomena, and some require consultation within the UD community to determine whether solutions may be determined based on similar phenomena in other languages. A number of these open issues are discussed in section 4, including tokenisation (§4.1), the difference between core and oblique arguments of verbs (§4.2), complex predicates (§4.3), multiple derivation (§4.4), and use of copulas in auxiliary constructions (§4.5). Section 5 wraps up.

2. Current status

In this section we briefly describe treebanks of Turkic languages that have been or are about to be released in UD.

Table 1. **Turkic UD treebanks at a glance**

Treebank	Language	Sentences	Words	Annotation	Genre
Kazakh-UD	Kazakh	1 047	10 032	manual annotation	Wikipedia, fiction
IMST-UD	Turkish	4 660	48 093	semi-auto.	conversion news, social media
Turkish-PUD	Turkish	1 000	16 886	auto./manual	annotation translated news
Turkish-GK	Turkish	2 803	17 800	manual annotation	grammar examples
Uyghur-UD	Uyghur	100	1 662	semi-auto.	conversion fiction

In addition to the released Turkish (§2.1), Kazakh (§2.2), and Uyghur (§2.3) corpora, there has been some work on UD annotation of other Turkic languages (§2.4). Section 2.5 outlines the main differences between the annotation standards of the released corpora.

2.1 Turkish

Turkish is relatively well represented in the UD with two treebanks. The IMST-UD treebank (Sulubacak et al. 2016a) is the result of a semi-automatic conversion of the IMST treebank (Sulubacak et al. 2016b) which, in turn, was based on METU-Sabancı treebank (Oflazer et al. 2003). The second Turkish treebank, Turkish-PUD, in the official UD repository is part of the parallel treebanks released during CoNLL 2017 UD parsing shared task (Zeman et al. 2017). Besides these two treebanks, the treebank reported in Çöltekin 2015 (Turkish-GK) is annotated for the purpose setting UD annotation guidelines for Turkish. To cover a wide range of morphosyntactic phenomena, the Turkish-GK treebank annotates the example sentences from comprehensive grammar book. This treebank follows UD version 1.3 annotation scheme, and currently not converted to version 2.0.

2.2 Kazakh

Kazakh is represented in UD by a single treebank (Makazhanov et al. 2015; Tyers and Washington 2015), which was first released in UD v1.3, and at the moment of writing contains 1109 trees (sentences) and a total of 10894 tokens. The annotation scheme of the treebank defines 16 UD POS tags, 45 "category=value" feature pairs, and 34 dependency relations of which four are language-specific. Tokenisation and morphological processing strategies in the Kazakh UD treebank follow the principles of Turkic lexica as defined by the Apertiun project[1]. One reason for this is to keep the UD corpus compatible with the morphological analysers developed by the Apertium Turkic working group.

Currently the treebank is partially compatible with UD v2.0 standard, with the choice of head direction in some constructions being one of the major discrepancies. The standard requires coordination and some compounds (e.g. names) to be left-headed, while the treebank developers believe that in Kazakh (and other Turkic languages) such constructions should be right-headed due to the placement of morphological locus, which is exclusive to the last (rightmost) element of such constructions. So far this issue has been resolved by an intermediate conversion step, where initially the annotation is performed in a right-headed fashion, and at the time of release a special script flips the heads of the constructions in question.

2.3 Uyghur

In Aili et al. (2016b), a treebank for Uyghur with 20,000 tokens is described. Tokens fit into one of 12 part-of speech categories and there are 137 morphological tags. There are 23 total dependency relations, with adjuncts classified by morphological case. In co-ordination, the conjunction is attached to the following conjunct and the preceding conjunct is attached to the following one (so-called 'head-final' conjunction).

Aili et al. (2016a) present a conversion of the Uyghur dependency treebank Aili et al. (2016b) to Universal Dependencies. They used some default mapping rules to convert the parts of speech and de-

[1] http://wiki.apertium.org/wiki/Turkic_lexicon

pendency relations, and then some limited rules based on the part of speech of the head to distinguish between ambiguous relations (for example mapping `att → {amod, det, nummod}`. The treebank contains surface forms, parts of speech and dependency relations, but no lemmas or morphological features.

2.4 Other

Ageeva and Tyers (2016) present two small treebanks for Tuvan and for Crimean Tatar of approximately 1,000 tokens each for use in testing a method of cross-lingual dependency parsing. They show that it is possible to take advantage of a morphological analyser and a treebank for another language in order to learn an improved delexicalised parser.

2.5 Main differences

At present, there are a number of differences in the dependency annotation standards for Kazakh, Turkish, and Uyghur. Quite a few of these differences are in the morphological annotation (part-of-speech tags and morphological features), but there are a handful of differences in tokenisation (how to approach words that include an orthographic space, or multiple words that do not include an orthographic space) and dependency annotation as well. In general, the Kazakh and Turkish annotation standards are more compatible with one another than either is with Uyghur.

One example of a difference in part-of-speech tagging is how locational pronoun-derived adverbials are represented. In Turkish, words like *nerede* ‘where’ and *nereden* ‘from where’ are labelled as `PRON` (with the appropriate case indicated in the morphology features), and hence usually have the dependency relation of nominal adverbials, `obl`. In Kazakh, the corresponding words *қайда* ‘where’ and *қайда* ‘from where’ are labelled as flat adverbs, or `ADV`, and hence have dependency relations of `advmod`. Which analysis is more appropriate is not clear: the fact that they are pronouns with case suffixes in both languages argues for their annotation as pronouns, while the fact that these pronouns are defective (they can’t take all case suffixes) in each language argues for their analysis as grammaticalised adverbs.

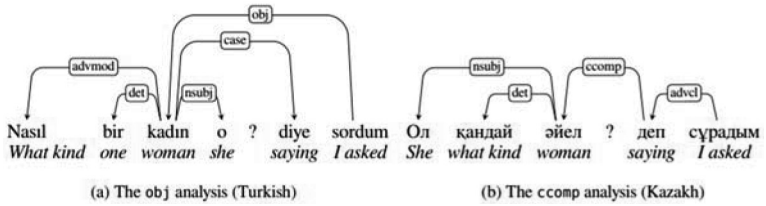


Figure 1. Two alternative analyses for speech with the adverbial "quotative word" use of the verb de- 'say'. Both sentences mean "I asked, 'what kind of woman is she?'" Analysis 1a shows the de- verb form as a case marker to the clause it governs, which is in turn an object of the main verb. In the analysis in 1b, the speech verb is treated as a verbal adverb adjunct to the main verb, with its own clausal complement

The current Uyghur corpus does not have annotation for morphological features, and there are some notable differences between Kazakh and Turkish standards. One of these is the annotation of `Person=3` in Turkish for any nominal–since as they trigger third-person agreement morphology as subjects and possessors. Kazakh does not have this feature. Similarly, Turkish annotates for `Polarity` values (e.g., on verb forms) of both `Pos` 'positive' and `Neg` 'negative', whereas Kazakh only indicates polarity if the value is negative.

Table 2. **Language specific relations: the tick mark (✓) means that the relation is found in the treebank; the − mark means that the relation could apply, but is not applied at present; and the asterisk (*) means that we consider this relation to be an erroneous classification**

Relation	Comments	Kazakh	Turkish	Uyghur
acl:poss	Adnominal modification with possessive	✓	–	–
acl:relcl	Adnominal modification with verbal adjective	✓	–	–
advmod:emph	Adverbial emphasiser (mostly -dA)	–	✓	✓
aux:q	Question word, -mI	–	✓	–
compound:lvc	Light verb	✓	✓	✓
compound:redup	Reduplication compound	–	✓	✓
flat:name	Proper name	✓	–	–

iobj:caus	Causee	✓	–	–
nmod:abl	Oblique in the ablative	*	*	✓
nmod:cau	Causee	*	*	✓
nmod:clas	Noun-noun compound	*	*	✓
nmod:comp	Nominal modifier [mostly ablative]	–	–	✓
nmod:poss	Genitive possessive modifier	✓	✓	✓
nmod:tmod	Time modifier	–	–	✓
obl:own	Owner in -DA	✓	–	–

In terms of tokenisation, the Turkish tokenisation standard always considers denominal adjectives formed with -/lI/ to be a noun followed by an adposition (i.e., two tokens), while in Uyghur these words are treated as lexicalised adjectives (i.e., as one token). The Kazakh treebank varies between these two approaches, currently in a somewhat unprincipled way. Another difference in terms of tokenisation is treatment of the so-called -/ki/ affix, two uses of which are the formation of the attributive locative (-/DAki/ in Turkish, -/DAGI/ in Kazakh) from the locative case form (-/DA/ in both) and the formation of the substative genitive (- /(n)Inki/ in Turkish and -/Nікі/ in Kazakh) from the genitive suffix (-/(n)In/ in Turkish and -/NIң/ in Kazakh). In all three languages, forms with these "compound" affixes are annotated with the appropriate relation (`amod` for attributive locative), but in Kazakh and Uyghur, these forms are not analysed as containing a separate -/ki/ suffix. That is, in Kazakh and Uyghur, there is a single token, while in Turkish, a separate -/ki/ token has a case dependency to the noun.

One difference in annotation of dependency relations is the treatment of the adverbial "quotative word" (Turkish *diye*, Kazakh *әden*, Uyghur دەپ). In all three languages, this form, morphologically speaking, is a verbal adverb (participle) form of the verb *de*- 'say', though in modern Turkish, the -/(y)A/ participle used in *diye* is not productively used as a verbal adverb. In Kazakh and Uyghur treebanks, it is analysed this way–thatis, it receives an `advcl` analysis (dependent on the main clause it comes in) and is labelled a `VERB`, with its relation to the head of its clausal complement labelled as a `ccomp`. This is shown in figure 1b. In the Turkish treebank, however, the speech verb is treated

as an ADP and has a case dependency relationship to the head of the "quoted" phrase, as shown in figure 1a.

There are also a number of differences with how language-specific relations are used. Table 2 presents a summary of the language-specific relations used in each treebank, and the applicability to the other treebanks.

3. Parsing performance

All three treebanks were included in the 2017 CoNLL shared task on Universal Dependency parsing from raw text data (Zeman et al. 2017). The results for the Turkic languages are presented in Table 3. LAS and UAS stand for labelled-attachment score and unlabelled-attachment score respectively, while CLAS stands for content-labelled attachment score (Nivre and Fang 2017).

Table 3. **Parsing performance in the CoNLL shared task. The column Train indicates the number of tokens in the training data, and the column Dev indicates the number of tokens in the development data. Note that there were no separate training sets for Turkish and Turkish-PUD; the latter is only used as a test set for parsers trained on Turkish training data**

Language	Train	Dev	Winning team (LAS)	UAS	LAS	CLAS
Kazakh	0	529	Dumitrescu et al. (2017)	45.72	29.22	25.14
Turkish	38 082	10 011	Dozat et al. (2017)	69.62	62.79	60.01
Turkish-PUD	38 082	10 011	Björkelund et al. (2017)	59.35	38.22	32.32
Uyghur	0	1662	Björkelund et al. (2017)	60.57	43.51	34.07

It is interesting to note the difference between the parsing performance on the Turkish section of the parallel treebank (Turkish-PUD) and on the IMST treebank (the main UD treebank for Turkish). Both of these treebanks have been converted from other treebank formalisms: the METU-Sabancı formalism in the case of the IMST treebank and Google Universal Dependencies in the case of the Turkish section of the parallel treebank.

It is also curious as to why the Kazakh and Uyghur numbers are so different despite the data size being similar (that is, while the Uyghur dev set had three times as many tokens, it didn't have lemmas or any morphology annotation). One explanation could be that there was a lot

more data sparsity when tagging Kazakh as opposed to Uyghur. It's also striking that Björkelund et al. (2017) got slightly better results for Uyghur than Turkish-PUD, despite a much smaller development corpus size. It should be mentioned that in addition to the training and development data, the shared task included at least 10 000 tokens of both Kazakh and Uyghur testing data.

4. Open questions

In this section we discuss certain phenomena in Turkic languages that present challenges during annotation. Some of those phenomena, e.g. multiple derivation, can be fairly well understood, but are difficult to handle adequately given the present UD annotation guidelines. The nature of others may require further investigation within the dependency grammar formalism.

4.1 Tokenisation

One of the guiding principles of Universal Dependencies is about its *lexicalism* (Nivre et al. 2016). That is, *words* are the basic units of annotation. The guidelines explicitly state that *word* here refers to *syntactic words*. It is allowed for orthographic words to be split when it is necessary for the syntactic analysis.

In earlier Turkish dependency parsing/annotation work, on the other hand, the words are split at all *derivation boundaries*, introducing a syntactic word (often called an *inflectional group* in Turkish NLP literature) for each derivational morpheme. For example, the Turkish word *sınırlandırılabilecek* 'that can be limited' can be represented as six syntactic words, delimited by ˆDB 'derivation boundary', as shown in (1).

(1) `sınır+Noun+A3sg+Pnon+Nom`
`ˆDB+Verb+Acquire`
`ˆDB+Verb+Caus`
`ˆDB+Verb+Pass`
`ˆDB+Verb+Able+Pos`
`ˆDB+Adj+AFuttPart`

Although derivation in Turkic languages can be quite productive, as exemplified by (1) above, arguing for necessity of this level of word segmentation is not always practical. Present Turkic UD treebanks seg-

ment the words when parts of the word may have conflicting morphological features and/or parts of the word can participate in different/ conflicting syntactic relations (Çöltekin 2016). In (1) above, none of the syntactic words are necessary since UD morphological features can mark the effect of each morpheme and none of the parts can participate in different syntactic relations–i.e., the parts cannot be modified by other words or ambiguously head other words in a sentence. However, there are also examples where the split is necessary. For example, failing to split the copular suffix in Figure 2 results in two nsubj relations headed by the same word[1]. Furthermore, the same word would be assigned both `Number=Plur` and `Number=Sing` features.

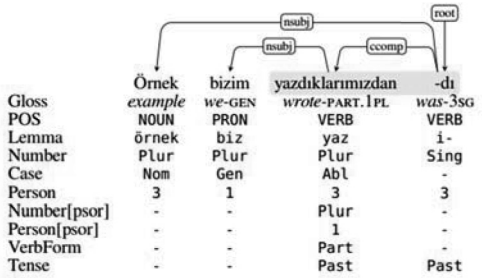


Figure 2. Dependency analysis of the sentence *Örnek bizim yazdıklarımızdandı* 'The example was from the ones we wrote'. Note how on the verbal form *yazdıklarımızdandı* there would be two values for Person, Number and Tense were the copula not split from the non-finite form

How to treat very productive derivational suffixes, which attach to phrases rather than single words, is also a challenge. These include the suffixes -/LI/ 'with', -/sIz/ 'without', and -/LIK/ '-ness, -ed', which appear in most Turkic languages. Very many forms that include these suffixes are lexicalised, for example *evsiz* 'homeless' (lit. 'without house'), *evli* 'married' (lit. 'with house'), *gözlük* 'spectacles' (lit. 'eye-

[1] Note that an analysis of this sentence more in line with the annotation standards for Kazakh would have yazdıklarımızdan as the root, with dı as a cop dependent of it, but this again results in the problem of having two nsubj relations headed by the same word. Issues like this are recognised by v2 of the UD standard (cf., http://universaldependencies.org/v2/copula.html) and affect non-Turkic languages as well.

ness')[1], but in some cases, for example *бір палаталы* 'one chambered, unicameral', it would be advantageous to consider the suffix separately since *бір* 'one' modifies the stem, not the whole form. There are potentially ambiguous examples as well, such as *iki gözlük*, which can technically mean either 'two glasses' or 'for two eyes' (e.g., the value of something–though this usage would be rare)[2], depending on what level the *lük* suffix is interpreted at. One possible solution would be to only split if the word to which the suffix is attached has its own modifiers of a certain type, although this sort of structural difference is difficult to segment in an NLP pipeline.

Another associated issue is related to syntactic words which contain multiple surface words. An example is the Turkish question marker which, when attached to predicates, may also carry some of the morphological features of the predicate. UD currently does not explicitly support syntactic words spanning multiple tokens, though the Kazakh treebank implements some things this way. Some rudimentary solutions exist in the present UD scheme–e.g., the `goeswith` relation or considering a space-separated token a single token–but ad hoc use of these without a standard could cause inconsistencies between Turkic language treebanks. Some issues related to syntactic words containing multiple surface words are discussed further in section 4.3.

Although some general guidelines exist for segmentation of words, there is a need for widely accepted, more concrete rules to ensure consistency among Turkic languages, and even among treebanks of the same language.

4.2 Core and oblique

In the UD v2 standard, the dependency relations `obj`, `iobj`, and `obl` are differentiated in the following way: `obj` is the most core element of a verb that is not its subject (i.e., a direct object), `iobj` is the

[1] It should be mentioned that it's not clear that these examples aren't understood by native speakers of Turkish as compositional and productively formed. Instead, perhaps this interpretation relies on the translation of these words to other languages (a poor criterion!) – it is not necessarily "metaphorical" (at least historically) that evli should mean 'married'.

[2] A reading that might be found in a wider range of real sources might be 'two-division' or 'two-room', based on another meaning of göz. In any case, any interpretation of such forms will depend on the context and whether an established lexicalised meaning exists.

next most core element that isn't a subject or direct object, and oblique is a non-core object. This relies on the notion of a difference between core and non-core elements, or complements versus adjuncts, respectively.

In Turkic languages, there does not seem to be a simple and clear way to delineate complements and adjuncts. No element of a verb phrase is absolutely required to be included in a grammatical utterance, not even the subject. While agreement marking will show the existence of a semantically present subject, even if not included in the sentence, Turkic languages do not mark object or indirect object agreement on the verb. Furthermore, since most of the cases have a very wide range of uses, many phrases can be used in any verb phrase, although with a different interpretation depending on the verb.

It seems clear, at least, that typical "accusative direct objects" (and morphologically unmarked indefinite direct objects) should be annotated with the `obj` relation. However, there is currently only one test that we can use to justify this and other relations: if the element participates in case promotion or demotion when the verb is made passive or causative, we consider it a core argument, to be labelled with `obj` if it seems "more core" and `iobj` if there is another element labelled obj. If the case marking does not change when the verb is made passive or causative, then the element is considered oblique, and receives the `obl` dependency relation. A more apt solution may exist, but has yet to be identified.

4.3 Complex predicates

In this subsection we discuss verbal (i.e., *non-copular*) complex predicates. Such predicates consist of two or more orthographic words that together convey single meaning, which is different from meanings (if any) of those words taken separately. Sometimes it is not at all clear how to classify the relationship between the constituents of complex predicates. For instance, a common Kazakh expression *пайда бол*, meaning appear or be established consists of what appears to be a noun *пайда* ('benefit', when used on its own as a noun) and a verb *бол* ('be' or 'finish'), which in this particular case loses its habitual copular and auxiliary functions. Thus, a verb that normally takes no arguments[1] in

[1] In UD copulas are subordinated to nominal predicates for the sake of cross-linguistic consistency (http://universaldependencies.org/u/overview/syntax.html).

this case governs what appears to be a noun; the question is with what syntactic relation.

Depending on the nature of their constituents, complex predicates in question can be roughly classified into three categories: (i) non-verbal + verbal; (ii) verbal + non-verbal; (iii) verbal + verbal. Assuming that predicates are finite (i.e., non-clausal), in all of these cases the rightmost constituent carries a personal agreement marker (sometimes covert), and in the latter two categories the first verbal constituent is usually non-finite[1] and contributes to TAM. Also, in all of the cases 'particles' and conjunctions may be inserted between the constituents at will – e.g., compare *пайда болды* 'it appeared' and *пайда да болды* 'and it also appeared'.

The first category of complex predicates (non-verbal + verbal) in certain UD treebanks (including Kazakh) is sometimes handled as a special sub-type of compounds, namely a light verb construction. This solution, while possible, relies on meaning, which is undesirable. There are two alternatives: (i) treat such constructions as a single space-separated token (which in some cases is done in the Apertium Kazakh lexicon); (ii) sacrifice the meaning and treat such constructions just as normal verb-argument or nominal-copula relations. Both alternatives have pros and cons. While the first one could be accommodated at the level of morphological analysis and tagging, it is not clear how to handle embedded 'particles' and conjunctions. As for the second alternative it just seems wrong to impose literal (usually absurd) meaning on otherwise meaningful constructions[2].

The second category of complex predicates (verbal + non-verbal) corresponds in Kazakh to negation of finite verbs which appears to consist of two tokens. In this construction, the first element (before the space) is morphologically a verbal noun or adjective ending in -/GAн/ and the last element is either *жоқ* 'non-existent' or *емес* 'not' – e.g., *айтқан жоқпын* 'I did not say'. Currently in the Kazakh treebank this case is handled at the morphological analysis step, and the construc-

[1] The only exception that we are aware of is the non-morphological negation, where (at least in Kazakh) the initial verbal constituent may agree with the subject, e.g. мен олай айтқаным жоқ vs мен олай айтқан жоқпын.

[2] Especially if a non-verbal constituent has no lexical meaning on its own and exists only in this sort of expression – e.g., міз бақпа 'pay no attention' or мiсе тұт 'be satisfied', where міз and мiсе do not exist as lexical items outside of these constructions.

tion is treated as a single token, as shown in figure 3a. It is currently unclear what to do when function words are embedded between the elements of constructions like this. Alternatively this sort of construction could be treated as analytic negation with *жоқ* or *емес* being considered negation words and subordinated to the leading verb with the relation `advmod:neg`, as shown in figure 3b. This approach however leads to non-verbal entities carrying personal agreement markers, which is undesirable. Although, this happens in the current conversion of the Turkish treebank with the question word -/mI/ which can carry person/number agreement and tense, as demonstrated in figure 3c.

(a) Current analysis of Kazakh multi-token negation (b) Alternative proposal (c) Turkish multi-token question word

Figure 3. Some examples of Kazakh multi-token negation (‘I didn’t say’) and the current analysis of Turkish multi-token question word (‘Would I say?’)

The third category of complex predicates (verbal + verbal) includes constructions which appear similar to auxiliary verb constructions, but which are probably best thought of as verbal adverb adjunct of main verb. The trailing finite verb does not contribute to TAM (tense, aspect, and mood – which typical auxiliaries in Kazakh convey), and the meanings of these combinations “feel” lexicalised (though it’s unclear whether this is just due to how they translate to other languages). Some examples include *болып табылады* ‘is found to be’, *болып саналады* ‘is considered’, *аман өтті* ‘mentioned’, *алып келді* ‘brought’, etc. In such constructions the preceding verbs assume a form of -/(I)п/ verbal adverb which can be followed by an auxiliary. The trailing verbs, however, are not always in the closed class of auxiliaries – and the ones that can be auxiliaries do not convey the normally associated auxiliary meaning, such as contributing to TAM. Both verbal elements give a combined meaning to the entire construction, e.g. *ама* ‘name (V)’ + *өт* ‘pass (V)’ combine as *аман өт* ‘mention’. Currently some of these constructions are treated as a single token in the Kazakh treebank due to the fact that they are lexicalised in the morphological analyser used to preprocess text for the treebank, but is designed for use in machine

translation pairs where such lexicalisation is useful. Because of this single-token analysis, the previously mentioned problem of function word embedding exists for these forms. However, other occurrences in the treebank are treated as a main verb with an `advcl` dependent. One disadvantage of this is that it does not match intuitions about the lexical semantics of these constructions, although, again, perhaps these intuitions are based on the translation of these forms to other languages. Another disadvantage of treating the verbs separately is that in some cases it can result in crossing dependencies, as shown in figure 4.

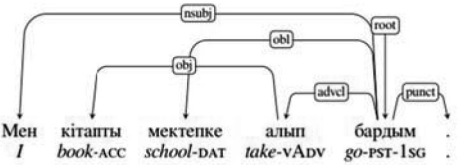


Figure 4. A separate-word analysis of a V+V compound verb in Kazakh, showing overlapping dependencies. The sentence translates to "I brought the book to school." Here, the verb *бар* 'go' has the oblique dependent *мектепке* 'to the school', while the verb *ал* has the direct object *кітапты* 'the book' (accusative). Note that a different order of the words, *Мен кітапты алып мектепке бардым* would have a slightly different meaning, 'I took the book and went to school' or 'Taking the book, I went to school'–and may not entail that the book ended up being brought to school as the depicted sentence does

The following facts all point to the interpretation of these verbs operating as a single, compound unit: that both verbs can contribute to the argument structure of the entire phrase, that the semantics are not always entirely compositional (but each verb usually contributes something), that the phrase seems to represent one event and not two, and that the verbs together share a single TAM reading. One possibility for how to deal with this would be to annotate these constructions with `compound` (or a subtype of the compound relation, e.g. `compound:v`), with the finite verb governing the non-finite one.

Another case that could be considered to fall under this category of complex predicates has not actually occurred in the Kazakh treebank, but is fairly frequent in speech: when a Russian infinitive is followed by the verb *ет* 'do, make'– e.g. *звонить ет* 'make a telephone call', *обжаловать ет* 'appeal to a higher court', etc. Due to the introduction of a foreign word, these constructions could potentially be handled

with a special dep relation that preserves structure but bears no particular grammatical function. Thus, the Russian infinitive could be tagged as X (uncategorised word) and be subordinated to the trailing verb with the relation dep. If a function word is embedded in between, it can be subordinated to the Russian infinitive with the same relation.

4.4 Multiple derivation

Many linguistic phenomena that are commonly expressed by syntactic means – e.g., relations between words – are expressed by morphological features in Turkic languages. In particular, Turkic verbs can be inflected for a range of affixes expressing features like tense, aspect, modality, voice and subject agreement. The UD specifications allow representing most of these features, and the changes in version 2.0 improved this representation considerably. However, in some cases the UD morphology specification is still sub-optimal for expressing some morphological phenomena.

The issue mainly arises when multiple Aspect, Mood and Voice features are present on the same verb. The Turkish examples in (2) include multiple Voice (2a) and Aspect (2b) features on a verb.

```
(2) a. bekle -t   -il  -iyor
       wait  CAUS PASS PROG
       'being stalled (=caused to wait)'
    b. oku  -yuver –iyor
       read RAPID  PROG
       'he/she is reading quickly'
```

In (2a), the verb has both passive voice and causative voice. Similarly in (2b), the affixes indicate that the action is done quickly, and it is in progress, both of which are typically defined as *aspect* (Göksel and Kerslake 2005). While the UD version 2.0 specifications allow marking each of these feature values individually, there is no clear way to mark multiple values for a single feature. In Turkish UD treebank, these words are marked using language-specific feature values. For example, the morphological annotation of (2a) includes Voice=CauPass, and the multiple aspect suffixes in (2b) are indicated by Aspect=ProgRapid.

A related issue is repetition of some of these features. Notably, the Turkish causative marker can be attached to a verb multiple times

without a principled limit, indicating a chain of causation[1]. Similarly, Turkish possibility/ability mood marker can be repeated two times in certain contexts. It is worth noting that this may also arise in non-Turkic languages – see e.g., Ainu in Senuma and Aizawa (2017).

Note that although the above method encodes the relevant information, it makes it difficult for an automated system (e.g., a parser), since the symbol `CauPass` does not clearly indicate the features `Cau` or Pass unless special attention is paid for this non-standard notation. Since some of these combinations are rare[2], it is difficult for a machine learning method to automatically discover that `CauPass` is equivalent to having both features marked individually. Similarly, a researcher, for example, looking for causative verbs in the language using a treebank search tool will likely to be misled by this ad hoc representation.

4.4 Use of copulas with non-finite verb forms

One issue that has arisen recently is how to analyse the use of copulas (as opposed to auxiliary verbs) with non-finite verb forms that occur together with auxiliary verbs. There appear to be cases of this in many Turkic languages. Normal non-finite form + auxiliary forms are straightforwardly dealt with in UD, as in figure 5.

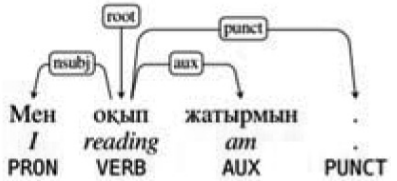


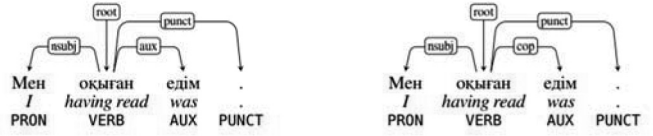
Figure 5. Dependency analysis of an auxiliary construction in Kazakh

In both Kazakh and Turkish, a number of finite verb forms are formed with what appears to be a verbal noun or verbal adjective form,

[1] Although real-life use is often limited, and multiple causative markers often (but not always) indicate emphasis rather than multiple levels of causation. In the Turkish-UD treebank there are no examples of multiple causative, but Turkish-GK includes examples with two causative suffixes, also in combination with the passive suffix.

[2] Of 9113 verbs in Turkish-UD treebank, Voice=CauPass is the most common multiple-feature marking with 115 occurrences. Others include 5 instances of Mood=CndPot, 4 instances of Mood=GenNec and Mood=DesPot, 2 instances of Aspet=DurPerf, and single instances of Aspect=ProgRapid and Mood=NecPot.

followed by normal copula agreement[1]. This includes forms like Turkish *okumuşum* ‘I read (past)’ (with perfect form *okumuştum* ‘I had read’) and *okurum* ‘I read (non-past)’ (with perfect form *okurdum* ‘I would read’) and Kazakh *оқығанмын* ‘I read (past)’ (with perfect form *оқыған едім* ‘I had read’) and *оқырмын* ‘I may read’ (with perfect form *оқыр едім* ‘I would read’). These structures are entirely parallel, although the Kazakh forms have a space between the verb form and the past form of the copula. This construction lends itself to a number of different analyses, as shown in figure 6[2]. There are also, in a smaller set of Turkic languages, “auxiliary” constructions that appear to be composed of a copula along with a form that cannot ever be verbal nouns or verbal adjectives and can only operate as a non-finite form together with an auxiliary. Because these look more like true auxiliary phrases, it’s clearer how they might be treated. One analysis is shown in figure 7.



(a) Copula treated as aux dependent of main verb, which is treated as a converb.

(b) Copula treated as cop dependent of main verb, which is understood to be adjectival or substantive.

Figure 6. Two different analyses of an apparent auxiliary construction in Kazakh that consists of a verbal noun or adjective and a copula, glossed as ‘I had read’. These analyses differ in how they view the copula auxiliary: either as the auxiliary in an auxiliary verb construction or as the copula in a normal copula predicate which happens to have a substantive or attributive verb form as the predicate

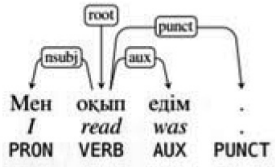


Figure 7. One analysis of an apparent auxiliary construction in Kazakh that consists of a non-finite form and a copula, glossed roughly as ‘I had read’

[1] Note that this is distinct from the obvious verbal nouns in forms like Turkish okumaktayım or Kazakh оқудамын – both meaning “I am reading” (with a long-term or habitual sense), and can be analysed as a verbal noun followed by a locative suffix, followed by a copula with person agreement.

[2] Note that the part of speech of copulas is always AUX in UD.

The set of choices around copula-as-auxiliary constructions includes several options for both dependency relations and morphological annotation, and Turkic languages have different orthographic strategies regarding tokenisation. This issue would be especially good to solve before further annotation.

5. Concluding remarks

We have presented an overview of the current status of the Turkic languages within the Universal Dependencies project and drawn attention to a number of inconsistencies that remain. We hope that this serves to inform and direct future research in the area of Turkic dependency parsing and Turkic language technology in general. After careful analysis we are convinced that the majority of substantial differences in the annotation schemes are a result of conversion from different grammatical traditions as opposed to real significant grammatical differences.

Acknowledgements

The authors would like to thank the UD community for thoughtful discussion and input on a range of issues discussed in this paper. We would also like to thank Deniz Uysal and Tolgonay Kubatova for help with native speaker judgements. The work of Aibek Makazhanov was supported by Nazarbayev University under the research grant №129-2017/022-2017.

REFERENCES

1. Ageeva, Ekaterina and Francis M. Tyers (2016). “Combined morphological and syntactic disambiguation for cross-lingual dependency parsing”. In: Proceedings of TurkLang 2016.

2. Aili, Mairehaba, Weinila Mushajiang, Tuergen Yibulayin, A. Kahaerjiang, and Yan Liu (2016a). “Universal dependencies for Uyghur”. In: Proceedings of WLSI/OIAF4HLT. Osaka, Japan, pp. 44–50.

3. Aili, Mairehaba, Aziguli Xialifu, Maihefureti, and Saimaiti Maimaitimin (2016b). “Building Uyghur Dependency Treebank: Design Principles, Annotation Schema and Tools”. In: International Workshop on Worldwide Language Service Infrastructure, pp. 124–136.

4. Björkelund, Anders, Agnieszka Falenska, Xiang Yu, and Jonas Kuhn (2017). “IMS at the CoNLL 2017 UD Shared Task: CRFs and Perceptrons Meet Neural Networks”. In: Proceedings of the CoNLL 2017 Shared Task:

Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver, Canada: Association for Computational Linguistics, pp. 40–51.

5. Çöltekin, Çağrı (2015). "A grammar-book treebank of Turkish". In: Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14). Ed. by Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski. Warsaw, Poland, pp. 35–49.

6. Çöltekin, Çağrı (2016). "(When) do we need inflectional groups?" In: Proceedings of The First International Conference on Turkic Computational Linguistics. Konya, Turkey.

7. Dozat, Timothy, Peng Qi, and Christopher D. Manning (2017). "Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task". In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver, Canada: Association for Computational Linguistics, pp. 20–30.

8. Dumitrescu, Stefan Daniel, Tiberiu Boroş, and Dan Tufiş (2017). "RACAI's Natural Language Processing pipeline for Universal Dependencies". In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver, Canada: Association for Computational Linguistics, pp. 174–181.

9. Göksel, Aslı and Celia Kerslake (2005). Turkish: A Comprehensive Grammar. London: Routledge.

10. Makazhanov, Aibek, Aitolkyn Sultangazina, Olzhas Makhambetov, and Zhandos Yessenbayev (2015). "Syntactic Annotation of Kazakh: Following the Universal Dependencies Guidelines. A report". In: Proceedings of the 3rd International Conference on Turkic Languages Processing, pp. 338–350.

11. Nivre, Joakim and Chiao-Ting Fang (2017). "Universal Dependency Evaluation". In: Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pp. 86–95.

12. Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Chris Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Dan Zeman (2016). "Universal Dependencies v1: A Multilingual Treebank Collection". In: Proceedings of Language Resources and Evaluation Conference (LREC'16).

13. Oflazer, Kemal, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür (2003). "Building a Turkish treebank". In: Treebanks: Building and Using Parsed Corpora. Ed. by Anne Abeillé. Springer. Chap. 15, pp. 261–277.

14. Senuma, Hajime and Akiko Aizawa (2017). "Toward Universal Dependencies for Ainu". In: Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017). Gothenburg, Sweden, pp. 133–139.

15. Sulubacak, Umut, Memduh Gökırmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit (2016a). “Universal Dependencies for Turkish”. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan, pp. 3444–3454.

16. Sulubacak, Umut, Tuğba Pamay, and Gülşen Eryiğit (2016b). “IMST: A revisited Turkish dependency treebank”. In: Proceedings of the 1st International Conference on Turkic Computational Linguistics (TurCLing). Konya, Turkey.

17. Tyers, Francis Morton and Jonathan North Washington (2015). “Towards a free/open-source dependency treebank for Kazakh”. In: Proceedings of the 3rd International Conference on Turkic Languages Processing, pp. 276–289.

18. Zeman, Daniel, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li (2017). “CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies”. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver, Canada: Association for Computational Linguistics, pp. 1–19.

СОДЕРЖАНИЕ

V МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ
ПО КОМПЬЮТЕРНОЙ ОБРАБОТКЕ
ТЮРКСКИХ ЯЗЫКОВ
«TURKLANG 2017»

Труды конференции

Том 1

В авторской редакции

Подписано в печать 28.12.2017. Формат 60×84 1/16.
Усл. печ. л. 17,4. Тираж 100 экз.

9 785969 004061