МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ИНСТИТУТ ИСТОРИИ ЯЗЫКА И ЛИТЕРАТУРЫ
УФИМСКОГО ФЕДЕРАЛЬНОГО ИССЛЕДОВАТЕЛЬСКОГО ЦЕНТРА
РОССИЙСКОЙ АКАДЕМИИ НАУК
АКАДЕМИЯ НАУК РЕСПУБЛИКИ ТАТАРСТАН
ИНСТИТУТ ПРИКЛАДНАЯ СЕМИОТИКА
МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РЕСПУБЛИКИ КАЗАХСТАН
ИНСТИТУТ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА
ЕВРАЗИЙСКОГО НАЦИОНАЛЬНОГО УНИВЕРСИТЕТА
ИМЕНИ Л.Н. ГУМИЛЁВА
СТАМБУЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ КОМПЬЮТЕРНОЙ ТЕХНИКИ И ИНФОРМАТИКИ

VIII МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ ПО КОМПЬЮТЕРНОЙ ОБРАБОТКЕ ТЮРКСКИХ ЯЗЫКОВ

TURKLANG-2020

(труды конференции)

Уфа - 2020

УДК 004.8+81’32
ББК 81.1

Восьмая Международная конференция по компьютерной обработке тюркских языков «TurkLang-2020». (Труды конференции). Уфа: ИИЯЛ УФИЦ РАН, 2020. – 296 с.

ISBN 978-5-91608-199-2

Ответственность за правильность, точность и корректность цитирования, ссылок, достоверность информации и оригинальность представленных материалов несут их авторы. Мнение авторов может не совпадать с позицией редакционной коллегии.

Сборник содержит материалы Восьмой Международной конференции по компьютерной обработке тюркских языков «TurkLang-2020» (Уфа, Башкортостан, Россия, 18–20 октября 2020 г.). Для научных работников, преподавателей, аспирантов и студентов, специализирующихся в области компьютерной лингвистики и ее приложений.

ISBN 978-5-91608-199-2

ПРЕДИСЛОВИЕ

Восьмая Международная конференция по компьютерной обработке тюркских языков TurkLang-2020 впервые прошла в дистанционной форме с использованием современных информационных технологий в г. Уфе на базе Института истории, языка и литературы УФИЦ РАН. Предыдущие конференции прошли в Астане (2013), Стамбуле (2014), Казани (2015, 2017), Бишкеке (2016), Ташкенте (2018), Симферополе (2019). География участников конференции показывает, что тематика конференции продолжает оставаться актуальной.

Целью серии Международных конференций TurkLang является создание пространства совместных компьютерных лингвистических исследований для тюркских языков.

Конференция, впервые проведенная полностью на открытой онлайн-площадке, объединила ученых из Российской Федерации (Абакан, Йошкар-Ола, Казань, Кызыл, Москва, Ростов-на-Дону, Омск, Санкт-Петербург, Симферополь, Томск, Уфа, Чебоксары, Якутск), Кыргызстана (Бишкек, Ош), Казахстана (Алматы, Нур-Султан), Азербайджана (Баку), Турции (Стамбул), Венгрии (Панония), Узбекистана (Ташкент, Фергана, Карши), Молдовы (Гагаузия), США (Нью-Йорк), Гонконг и прошла в атмосфере конструктивного обсуждения представленных научных докладов. В выступлениях участников были представлены качественно новые результаты, связанные с разработкой компьютерных лингвистических приложений для тюркских, монгольских и угро-финских языков.

В сборник трудов включены представленные оргкомитету статьи участников VIII Международной конференции по компьютерной обработке тюркских языков «TurkLang-2020». К сожалению, разразившаяся пандемия не позволила собрать все оформленные для печати материалы докладчиков, но по мере представления, они будут размещены на сайтах оргкомитета.

Участники конференции особо отмечают важность и ценность проведения Международной конференции TurkLang

как наиболее актуального и действенного механизма сохранения, изучения и развития тюркских языков и культур в эпоху глобализации и всеобщей компьютеризации. Было особо отмечено, что конференция TurkLang становится привлекательной открытой площадкой для обсуждения аналогичных проблем и задач также и для других малоресурсных языков.

С целью обеспечения эффективности и продуктивности конференции TurkLang, повышения ее привлекательности и пользы для всего тюркского мира, включая исследователей тюркских языков и разработчиков компьютерных средств их поддержки в инфокоммуникационном пространстве, и ее развития как междисциплинарной дискуссионной и технологической площадки, считаем целесообразным проведение круглых столов, семинаров (в том числе в он-лайн режиме) и обсуждение участниками конференции совместных проектов по компьютерной обработке тюркских языков, принятие мер к тому, чтобы сайт TurkLang стал одним из центров информационного обмена по вопросам компьютерной лингвистики тюркских языков.

Директор Ордена Знак Почета
Института истории, языка и литературы,
д.г.н., профессор А.В.Псянчин

ПРОГРАММНЫЙ КОМИТЕТ

Сулейманов Джавдет Шевкетович (Казань, Республика Татарстан, РФ) – сопредседатель

Шарипбаев Алтынбек Амирович (Нур-Султан, Казахстан) – сопредседатель

Ешреф Адалы (Стамбул, Турция) – сопредседатель

Псянчин Айбулат Валиевич (Уфа, Республика Башкортостан, РФ) – сопредседатель

Сиразитдинов Зиннур Амирович (Уфа, Республика Башкортостан, РФ) – заместитель председателя

Жубанов Аскар Кудайбергенович (Алматы, Казахстан)

Абдурахмонова Нилуфар (Ташкент, Узбекистан)

Алтынбек Гулила (Урумчи, Китай)

Гатиатуллин Айрат Рафизович (Казань, Республика Татарстан, РФ)

Дыбо Анна Владимировна (Москва, РФ)

Желтов Валериан Павлович (Чебоксары, Республика Чувашия, РФ)

Исраилова Нелла Амантаевна (Бишкек, Кыргызстан)

Кубединова Ленара Шакировна (Симферополь, Республика Крым, РФ)

Мамедова Масума Гусейновна (Баку, Азербайджан)

Офлазер Кемаль (Доха, Катар)

Садыков Ташполот (Бишкек, Кыргызстан)

Салчак Аэлита Яковлевна (Кызыл, Республика Тыва, РФ)

Сулайманов Мухаммад-али (Симферополь, Республика Крым, РФ)

Татевосов Сергей Георгиевич (Москва, РФ)

Торотоев Гаврил Григорьевич (Якутск, Республика Саха (Якутия), РФ)

Тукеев Уалишер Ануарбекович (Алматы, Казахстан)

Ергеш Бану (Нур-Султан, Казахстан)

ОРГАНИЗАЦИОННЫЙ КОМИТЕТ

Председатель организационного комитета:

Псянчин А.В., заместитель председателя УФИЦ РАН, директор ИИЯЛ УФИЦ РАН (Уфа, Республика Башкортостан, РФ)

Члены организационного комитета:

Хисамитдинова Ф.Г. (Уфа, Республика Башкортостан, РФ)

Сиразитдинов З. А. (Уфа, Республика Башкортостан, РФ)

Бускунбаева Л.А. (Уфа, Республика Башкортостан, РФ)

Ишмухаметова А.Ш. (Уфа, Республика Башкортостан, РФ)

Гатиатуллин А.Р. (Казань, Республика Татарстан, РФ)

СЕКЦИЯ 1
ЭЛЕКТРОННЫЕ КОРПУСЫ ТЮРКСКИХ ЯЗЫКОВ

Pirmanova K.K., Ongarbaeva M.S.
Institute of Linguistics named after A.Baitursynov,
Kazakhstan, Almaty

ISSUES OF CREATING NATIONAL LANGUAGE CORPORA

Abstract. The article is devoted to the main issues of creating National Language Corpuses. The main features of the National Language Corpus are touched upon. Also, the problems of development and creation of the Kazakh national language corpus, its variants that are in the public domain are considered.

Keywords*: national language corpus, national corpus of the Kazakh language, variants of the corpus of the Kazakh language*.

Пирманова К.К., Онгарбаева М.С.
Институт языкознания им. А.Байтурсынова,
Казахстан, Алматы

О НАЦИОНАЛЬНЫХ КОРПУСАХ КАЗАХСКОГО ЯЗЫКА

Аннотация. Статья посвящена вопросам создания национальных языковых корпусов. Затронуты основные возможности национального языкового корпуса. Также рассмотрены существующие в открытом доступе варианты национального корпуса казахского языка, проводится сравнение и анализ их функциональных возможноситей, ставится задача разработки и создания полного казахского национального языкового корпуса.

Ключевые слова*: национальный языковой корпус, национальный корпус казахского языка, варианты корпусов казахского языка.*

Общепринято считать, что корпус текстов — это определенным образом организованное множество, элементами которого являются тексты. Организация корпуса может быть

самая разная в зависимости от прагматических целей его создателя или пользователя. Тексты, являющиеся составляющими элементами корпуса, могут представлять собой целое оригинальное словесное произведение или какую-либо его часть [8].

В казахском языкознании вопрос создания национального корпуса был поднят в начале XXI века и на сегодня имеются несколько вариантов национального корпуса. Сравнительный анализ корпусов имеет как теоретическое, так и практическое значение по нескольким причинам: во-первых, он позволяет проверить и описать систематические различия и сходства в коммуникативном использовании языка, во-вторых, облегчает выявление кросслингвистических исследований, включая синхронную типологию и функционально-коммуникативные аспекты языкового сравнения [2].

В отделе прикладной лингвистики Института языкознания имени А.Байтурсынова работы по созданию национального корпуса начались в 2009 г. в соответствии с требованиями информационных технологий. По инициативе профессора А.К.Жубанова были исследованы и изучены методы и технологии лингвистических корпусов.

В дальнейшем были разработаны программы по внедрению лингвистических разметок и накоплен значительный опыт. В то же время впервые был разработан автоматический морфологический анализатор. Программу создали программист института Языкознания имени А.Байтурсынова Д.Токмырзаев и бывший научный сотрудник Института информатики К.Койбагаров.

На сегодня в разработке новых лингвистических функционалов и в наполнении базы данных корпуса активно участвуют А.Жанабекова, К.Пирманова, Б.Карбозова, А. Кожахметова, Г.Тлегенова, Е.Бесиров. Особо следует отметить роль профессора А. Жунисбека, основоположника казахской экспериментальной фонетики, разработчика ряда вопросов синтеза и анализа казахской. Его трехступенчатое руководство активно применяется при фонетической разметке казахского корпуса.

Разработка программного обеспечения национального корпуса осуществлялась в инструментальной среде Visual Studio 2010 на языке программирования C#. Морфологические разметки проводились в базе MSSQLServer 2008. Основной текст — это совокупность текстов каждого автора или стиля (жанра) в формате MSWORD, которые имеют метатеги, например:

<Жанр>Проза</Жанр>

<Автор>Әбіш</ Кекілбаев</Автор>.....[4].

Следует отметить, в корпусе казахского языка преобладают тексты художественной литературы. Рассматриваемый корпус также позволяет осуществлять поиски по выбранным пользователем стилям, например: научный стиль, публицистика, художественная литература, среди них: Әбдіжәміл Нұрпейісов, “Соңғы парыз” (1921), Әбдіжәміл Нұрпейісов, “Қаң мен тер” (1973), Мұхтар Әуезов, “Қорғансыздың күні”(1921), Ілияс Есенберлин, “Көшпенділер” (1973), Мұхтар Мағауин, “Шыңғыс хан” (2011), Жүсіпбек Аймауытов, “Ақбілек” (1927), Жүсіпбек Аймауытов, “Күнекейдің жазығы” (1928), Мұхтар Әуезов, “Шығармалар” (1925–1929), 20-ти томник, Ғабит Мүсірепов, “Ұлпан” (1974), Мұхтар Әуезов, “Іздер” (1932), Қабдеш Жұмаділов, “Бір қаланың тұрғындары”, (1982) итд.

Накопленный сотрудниками опыт разработки казахского корпуса был обобщен и издан в 2017 г. в виде монографии «Корпусная лингвистика» [3].

Разработанный сотрудниками Института языкознания Национальный корпус казахского языка – это уникальный инновационный справочник о языке, превышающий возможности всех взятых вместе словарей и грамматик, созданных трудом многих поколений лингвистов-лексикографов. Отметим, что в последнее время наблюдается тенденция к интеграции словарей и корпусов. Словарные статьи, в которых осуществлена разметка и каждый вход которого имеет ссылку на представительный корпус, являются мощным инструментом для лингвистического исследования [5]. Опыт создания таких электронных словарей с прямым выходом

в корпус уже имеется, в частности, словари, созданные на основе НКРЯ [9].

Еще одним вариантом казахского национального корпуса является Алматинский корпус. Работа над проектом Алматинского корпуса была начата в мае 2012 г. при поддержке ректора КазНУ им. аль-Фараби Г.М. Мутанова и проводилась силами кафедры общего языкознания и иностранной филологии факультета филологии, литературоведения и мировых языков университета с участием сотрудников факультета филологии НИУ ВШЭ (Москва). В 2013 г. объем корпуса составлял 650 тыс. словоупотреблений [6].

Создание рассматриваемого корпуса казахского языка включает обработку естественного языка (natural language process (NLP); лексикографическую обработку, токенизацию, лемматизацию, морфологический анализ и другие с целью разработки автоматизированного извлечения информации, [6].

Следующим вариантом корпуса казахского языка является пилотный мегапроект национального корпуса казахского языка “ҚАНАТҚАҚТЫ”. [10].

В.А. Плунгян считает слово «национальный» термином, отражающим скорее семантику английского слова «national», чем русского слова «нация». Впервые это определение появилось в названии Британского национального корпуса (British National Corpus, BNC), созданного в 1990-е годы в Великобритании специалистами лексикографами; это не самый первый электронный корпус, созданный в мире, но один из лучших, крупнейших и наиболее известных. Для британцев слово «национальный» означало в первую очередь «характеризующий британский национальный вариант английского языка» (в отличие от американского, австралийского и т. п.), но поскольку этот корпус очень быстро стал практически эталоном корпуса вообще, то значение слова «национальный» незаметно изменилось. Национальным корпусом стали называть просто самый большой и представительный корпус, характеризующий язык данной страны в целом [7].

Корпусная лингвистика формируется как особый раздел языкознания, она позволит многим специалистам по казахскому языку использовать крупномасштабные экспериментальные материалы, находить необходимые языковые данные и вносить соответствующие изменения. Все это способствует новому

взгляду на эмпирические подходы к достоверности исследований казахского языка и внедрению наиболее важных языковых материалов в области науки.

ЛИТЕРАТУРА

1. Алматинский корпус казахского языка. Информационный ресурс. URL: http://web-corpora.net /KazakhCorpus /search/?interface_language=ru (дата доступа: 8.07.2020).
2. Гвишиани Н.Б. Практикум по корпусной лингвистике / Н.Б. Гвишиани // English on Computer: A Tutorial in Corpus Linguistics. – М.: Высшая школа, 2008. – 191 с.
3. Жұбанов А.Қ., Жаңабекова А.Ә. Корпустық лингвистика /А.Қ. Жұбанов, А.Ә. Жаңабекова. – Алматы: "Қазақ тілі" баспасы, 2017. – 336 б.
4. Жұбанов А.Қ. Компьютерлік лингвистикаға кіріспе //Монография. – Алматы, 2013. – 204 б.
5. Кустова Г.И. Семантическая разметка в электронных корпусах и электронных словарях // Труды международной конференции «Корпусная лингвистика–2011». – СПб., 2011. С. 234–242.
6. Мадиева Г.Б. Методы и приемы составления корпуса казахского языка. Сборник материалов международного научно-методического семинара, 26 ноября, 2015. – С. 3-12.
7. Плунгян В.А./ Зачем/ мы делаем Национальный корпус русского языка?/ В.А. Плунгян // Отечественные записки. – 2005. – № 2. – С. 296–308.
8. Рыков В. В. Прагматически ориентированный корпус текстов. URL: http://rykov-cl.narod.ru/t.htm.
9. Словари национального корпуса русского языка. Информационный ресурс. URL: http://dict.ruslang.ru/ (дата доступа: 8.07.2020).
10. Қанатқақты. Информационный ресурс. URL:http://89.250.84.132 /Findsystem/Findsystem/ (дата доступа: 8.07.2020).

Salchak A.Ya., Ondar V.S., Oorzhak B.Ch., Hertek A.B.
Tuva State University,
Russia, Tuva, Kyzyl

ACTIVITIES OF THE SCIENTIFIC AND EDUCATIONAL CENTER "TURKOLOGY" IN THE SPHERE OF CORPUS LINGUISTICS AND INFORMATION TECHNOLOGIES IN HUMAN SCIENCES

Abstract. The article describes the activities of the scientific and educational center "Turkology" of Tuvan State University in the field of corpus linguistics and information technology. A description of the results obtained during the implementation of research projects, an overview of the prepared databases are given, the work on the creation of the Electronic Corpus of the Tuvan language is highlighted.

Keywords: *corpus linguistics, database, tuvan language, corpus of texts.*

Салчак А. Я., Ондар В.С., Ооржак Б.Ч., Хертек А.Б.
Тувинский государственный университет,
Россия, Тува, Кызыл

ДЕЯТЕЛЬНОСТЬ НОЦ «ТЮРКОЛОГИЯ» В СФЕРЕ КОРПУСНОЙ ЛИНГВИСТИКИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ В ГУМАНИТАРНЫХ НАУКАХ

Аннотация. В статье описывается деятельность НОЦ «Тюркология» Тувинского государственного университета в сфере корпусной лингвистики и информационных технологий. Дается описание результатов, полученных в ходе выполнения исследовательских проектов, обзор подготовленных баз данных, освещается работа по созданию Электронного корпуса тувинского языка.

Ключевые слова: *корпусная лингвистика, база данных, тувинский язык, корпус текстов.*

Научно-образовательный центр «Тюркология» (НОЦ «Тюркология») Тувинского государственного университета был открыт в 2009 г. в целях развития и укрепления научной и научно-образовательной работы университета, преемственности научных, научно-педагогических кадров и закрепления молодежи в сфере науки, образования, поддержки государственного статуса тувинского языка как одного из государственных языков наряду с русским в Республике Тыва.

Научно-исследовательская работа НОЦ «Тюркология» включает изучение актуальных вопросов тувинского языкознания, разработку методических вопросов преподавания тувинского языка, вопросы ареальной лингвистики и языковых контактов. И одним из приоритетных направлений деятельности Центра является корпусная лингвистика.

Первым большим проектом по созданию корпуса текстов на тувинском языке явился проект «Электронный корпус текстов тувинского языка» (рук. Салчак А.Я., РГНФ, 2011-13 гг.). Созданы базы текстов художественной литературы советского и современного периодов, базы по произведениям фольклора разных жанров, база официально-деловых документов, база публицистических текстов, база диалектного корпуса. Разработаны пробные программы по поиску морфем, словоформ, аналитических скреп в тексте тувинского языка. Функционирует сайт «Электронный корпус текстов тувинского языка» (ЭКТТЯ) [1]. Результаты проекта используются в корпусных исследованиях, практике преподавания вузовских курсов по тувинской филологии.

Данная работа была продолжена в проекте «Создание базы данных лексического фонда тувинского языка» (рук. Ооржак Б.Ч., РГНФ, 2016-17 гг.). В результате проекта были созданы базы данных по тематическим группам имен существительных, лексико-семантическим группам имен прилагательных, глаголов, по лексико-семантическим разрядам местоимений и наречий. Определена лексическая сочетаемость выделенных групп.

В рамках проекта по Государственному заданию Министерства науки и высшего образования «Системные

изменения в языковой картине мира тувинцев: традиции и современность» (рук. Ооржак Б.Ч., 2017-19 гг.) пополнялись базы данных, были созданы диалектный и фольклорный подкорпусы, базы параллельных тувинско-русских, русско-тувинских и монгольско-тувинских текстов, разработаны интерактивные лингвистические карты по диалектным явлениям в тувинском языке. На основе материалов Электронного корпуса были защищены диссертации: на соискание ученой степени доктора филологических наук Бавуу-Сюрюн М.В. «История формирования диалектов и говоров тувинского языка» (2018); Ооржак Б.Ч. «Система грамматической модальности в тувинском языке (в сопоставлении с тюркскими языками Сибири)» (2019); на соискание ученой степени кандидата физико-математических наук Монгуш Ч. М. «Разработка метода и средств фрагментации и дефрагментации формальных контекстов» (2020). Подготовлены диссертации: на соискание ученой степени доктора филологических наук Хертек А.Б. по теме «Грамматические категории имени в тувинском языке (в сопоставлении с тюркскими языками Сибири)»; на соискание ученой степени кандидата филологических наук на соискание ученой степени кандидата филологических наук Ондар М. В. «Особенности языка тувинского героического эпоса»; Ооржак С. С. «Лексика верований в тувинском языке», Кошкендей И.М. «Особенности языка тувинских народных песен и припевок», Монгуш А.А. «Концепт *Родина* в стихотворных произведениях тувинских писателей».

Базы «Электронного корпуса текстов тувинского языка» послужили основой для создания научно-методического и образовательного контента «Тываның чогаалчылары – Писатели Тувы» [2], который содержит тексты произведений тувинских писателей 1970-90-х годов в электронном виде, автобиографические сведения о писателях, критические статьи. В настоящее время материал сайта широко используется учителями тувинского языка и литературы, который сегодня включает автобиографии и сведения о деятельности тувинских писателей, тексты произведений.

За период работы над проектами получены более 20 свидетельств РИД на базы данных и программы:

- база данных «Словоизменение имени существительного в тувинском языке» (Хертек А.Б., Ооржак Б.Ч., Далаа С.М.). Свидетельство № 2017620074 от 18 января 2017 г.
- база данных «Формы изъявительного наклонения глагола в тувинском языке» (Ооржак Б. Ч., Хертек А. Б., Далаа С. М.). Свидетельство № 2017620139 от 3 февраля 2017 г.
- база данных «Гидронимия Тувы» (Далаа С.М., Хертек А.Б., Ондар Ш.В.). Свидетельство № 2017620954 от 24 августа 2017 г.).
- программа ЭВМ «Гидронимия Тувы» (Далаа С.М.). Свидетельство № 2017619394 от 24 августа 2017 г.
- база данных «Тувинские героические сказания» (Ондар М.В., Бавуу-Сюрюн М.В., Далаа С.М., Монгуш Ч.М.). Свидетельство № 2017620090 от 19 января 2017 г.
- программа для ЭВМ «Клише и стандарты в текстах тувинских героических сказаниях» (Ондар М.В., Бавуу-Сюрюн М.В., Далаа С.М., Монгуш Ч.М.). Свидетельство №2017620024 от 10 января 2017 г.
- база данных «Средства выражения пространственных отношений в тувинском языке» (Далаа С.М., Хертек А.Б., Ооржак Б.Ч., Ондар В.С.). Свидетельство №2018620755 от 25 мая 2018 г.
- база данных «Антропонимы в тувинском героическом эпосе» (Далаа С.М., Бавуу-Сюрюн М.В., Ондар М.В.). Свидетельство № 2018621622 от 18 октября 2018 г.
- программа для ЭВМ «FCACorpus концептуального моделирования тувинских текстов методами анализа формальных понятий» (Монгуш Ч.М.). Свидетельство № 2018618907от 23 июля 2018 г.
- программа для ЭВМ «Программа формирования контекстов в корпусе тувинского языка» (Монгуш Ч.М., Быкова В.В.). Свидетельство №2018618908 от 23 июля 2018 г.
- база данных «Лексика животноводства в языке цэнгэльских тувинцев» (Далаа С.М., Баярсайхан Б.). Свидетельство № 2019621344, от 22 июля 2019 г.

- база данных «Лексика похоронно-поминального обряда в тувинском языке» (Далаа С.М., Бавуу-Сюрюн М.В., Ооржак С.С.).

Направление деятельности Центра в сфере корпусной лингвистики и информационных технологий в гуманитарных науках продолжается в сотрудничестве с кафедрой информатики физико-математического факультета Тувинского госуниверситета в составе консорциума с вузами России, Германии, Греции, Кипра и Монголии. Университет стал обладателем гранта международного конкурса Европейского Союза Erasmus + на улучшение образовательных программ в области искусства и гуманитарных наук с помощью европейских методов и инструментов. Проект направлен на подготовку магистерской программы, обеспечивающей подготовку профессиональных кадров, владеющих новыми цифровыми технологиями в области гуманитарных наук. Целью этой работы является продвижение общих ценностей и более тесное взаимопонимание между разными людьми и культурами, поскольку основное внимание предлагаемой магистерской программы уделяется цифровизации искусства и гуманитарных наук. В рамках проекта планируется создание цифровой лаборатории (онлайн-платформы проекта), которая будет способствовать международному сотрудничеству исследователей, преподавателей и студентов в рамках небольших проектов по оцифровке, сохранению, представлению и продвижению культурного наследия. Гуманитарный аспект новой магистерской программы будет усилен методами и инструментами STEM (Science, technology, engineering, mathematics) в соответствии с политическими рекомендациями Европейского Союза по усилению STEM-образования и планом действий по цифровому образованию. Это будет способствовать продвижению и облегчению развития цифровых навыков учителей, внедрению инновационных цифровых инструментов для эффективного обучения, обеспечению согласованного межпредметного подхода в рамках многопрофильной системы.

В начале 2020 г. была также начата работа над общеуниверситетским проектом по созданию доступного

образовательного и информационного электронного ресурса, содержащего основные сведения о Туве; создание баз данных как основы проведения фундаментальных и прикладных научных исследований с применением информационных технологий. Данный проект предполагает обеспечение учебного процесса университета дополнительными учебными материалами, которые включают в себя региональный и этнокультурный компонент изучаемых дисциплин, оригинальную информацию о новейших результатах научных исследований и разработок коллектива университета; составление баз данных и компьютерных программ для обработки и поиска информации; создание, дальнейшую поддержку и развитие единого образовательного и информационного портала университета.

Источники

1. Электронный корпус текстов тувинского языка. URL:http: http://www.tuvancorpus.ru/ (дата обращения: 15.09.2020).

2. Электронный ресурс. URL:http://pisateli-tuvy.ru/ (дата обращения: 15.09.2020).

Sirazitdinov Z.A.
Institute of History, Language and Literature, UFRC, RAS,
Russia, Bashkortostan, Ufa

TO THE QUESTION ABOUT THE NATIONAL CORPUS BASHKIR LANGUAGE

Abstract. The article examines corpus projects on the Bashkir language existing on the Internet, reveals the characteristics and principles of their development. The criteria of the volume, representativeness, balance and completeness of the national corpus, which are the main delimiters that distinguish it from ordinary corpuses, are singled out and substantiated. The author puts forward the task of creating a real national corpus of the Bashkir language on the basis of the considered corpus projects, which includes, in addition to prose, journalistic and folklore texts, also texts of artistic drama, educational and scientific, official business texts, subcorpuses of free spoken language and literary speech

Keywords: *corpus linguistics, Bashkir language, national corpus criteria, corpus representativeness, corpus completeness, national corpus collective, corpus balance.*

Сиразитдинов З.А.
Институт истории, языка и литературы УФИЦ РАН,
Россия, Башкортостан, Уфа

К ВОПРОСУ О НАЦИОНАЛЬНОМ КОРПУСЕ БАШКИРСКОГО ЯЗЫКА

Аннотация. В статье рассматриваются существующие в сети интернет корпусные проекты по башкирскому языку, раскрываются характеристики и принципы их разработки. Выделяются и обосновываются критерии объема, репрезентативности, сбалансированности и полноты национального корпуса, являющиеся основными разграничителями отличающими его от обычных корпусов. Автором выдвигается задача создания на основе рассмотренных

корпусных проектов реального национального корпуса башкирского языка, включающего кроме прозаических, публицистических и фольклорных текстов также тексты художественной драматургии, учебных и научных, официально-деловых текстов, подкорпусов произвольной разговорной речи и литературной речи.

Ключевые слова: *корпусная лингвистика, башкирский язык, критерии национального корпуса, репрезентативность корпуса, полнота корпуса, коллектив национального корпуса, сбалансированность корпуса.*

Сейчас все крупные языки обзавелись своими корпусами, созданы и ведутся корпусные разработки и по тюркским языкам: турецкому, казахскому, кыргызскому, башкирскому, татарскому, тувинскому, хакасскому, шорскому и др.

Сегодня по башкирскому языку в сети Интернет существуют 6 корпусных разработок [1; 6; 7; 8; 11; 18], ведется работа по созданию аудиокорпуса диалектных текстов. Три из этих функционирующих корпусов являются продуктами лаборатории лингвистики и информационных технологий Института истории, языка и литературы Уфимского федерального исследовательского центра Российской академии нук (УФИЦ РАН). Это корпусные проекты прозаических, публицистических и фольклорных текстов, они размещены на портале Машинного фонда башкирского языка [mfbl2.ru].

1. Проект корпуса прозаических текстов. На сегодня этот корпус содержит тексты художественных произведений 163 прозаиков общим объемом порядка 16 миллионов словоупотреблений. Паспортизация текстов включает: название текста, ФИО автора, объем текста, время создания (если автором указано время создания, то указывается эта дата, если нет, то дата издания (например: “Мөхәббәт һәм нәфрәт”, 1964).

2. Проект корпуса публицистических текстов. Корпус содержит тексты республиканских газет “Башҡортостан” (2016-2017), “Йәшлек” (2010, 2016-2017), “Киске Өфө” (2009-2011); районных газет “Ашҡаҙар” (2009), “Табын” (2000, 2009, 2012-2013); журналов “Ағиҙел” (2010-2013), “Башҡортостан ҡыҙы”

(2016-2017). Объем корпуса составляет 12 миллионов словоупотреблений, система экстралингвистических разметок данного корпуса включает название, автора, дату публикации, тематику и жанр статьи. Выделены следующие темитики и жанры публицистического стиля:

- тематика текстов: политическая и социальная жизнь (политика, право, философия); экономика (производство, строительство, бизнес, финансы, коммерция); сельское хозяйство; искусство, культура и литература; наука и техника; образование; природа, путешествие; частная жизнь; спорт; религия; психология; медицина; красота и здоровье;
- жанры текстов: интервью, беседа; статья, очерк, репортаж, обозрение; советы; письма; обзор печати (новости из других источников); поздравления; художественно-публицистические жанры (эссе, фельетон, рассказ, стихи, эпиграммы); рецензия.

3) Проект корпуса фольклорных текстов. Данный корпус содержит фольклорные тексты из многотомных изданий фольклора башкирского народа и неизданных материалов из архивного фонда УФИЦ РАН общим объемом в 855870 словоупотребления.

Система экстралингвистических разметок фольклорного корпуса включает: название текста, жанр, жанровую разновидность/тематику, источник. Определены следующие жанры: эпосы, сказки, легенды, предания, песни, баиты, такмаки, обрядовый фольклор, мунажаты, поговорки, пословицы, загадки, поверья, скороговорки, приметы, заклятья, проклятья, клятвы.

Лаборатория лингвистики и информационных технологий также ведет разработку аудиокорпуса говоров башкирского языка. На данном этапе идет сбор, обработка и транскрибирование полевых материалов. Запись осуществляется на цифровой диктофон в несжатом формате (wav, 16бит/22kHz – 16бит/48kHz). Экстралингвистическая разметка материала включает данные об информаторе: пол, образование, возраст, язык обучения, язык общения в семье, идентификация национальности, имя, отчество, фамилия, место последнего долгого проживания, время проживания до последнего места

проживания, время записи. Исследователями определены и описаны принципы транскрибирования полевых материалов [16].

Существуют также корпусные проекты, разработанные другими исполнителями. Это следующие три корпусных проекта:

Башкирский поэтический корпус [1]. Поэтический корпус включает тексты стихотворных произведений башкирских поэтов XX и начала XXI века, имеет объём в 1,8 млн. словоупотреблений. Разработан сотрудниками Башкирского государственного университета совместно с Институтом языкознания РАН. Тексты в корпусе снабжены наряду с морфологической разметкой, также специальной стиховедческой разметкой, позволяющей осуществлять поиск в строках, написанных определённым метром.

Устный корпус башкирского языка дер. Рахметово и с. Баимово [18]. Является совместной разработкой сотрудников НИУ ВШЭ, СПбГУ, Института лингвистических исследований РАН при поддержке Института истории, языка и литературы УФИЦ РАН. Корпус содержит устные монологические тексты общим объемом в 25000 слов.

Национальный корпус башкирского языка [11]. Разработчиками являются сотрудники Школы лингвистики НИУ ВШЭ, волонтеры республики, которые обеспечили электронными текстами. Объем корпуса 20 – миллионов словоупотреблений, включает тексты художественных произведений и разнородные тексты из башкироязычных сайтов. К сожалению, данный корпус не соответствует своему названию и порождает путаницу.

Актуальность разработки корпусов, создания общих унифицированных принципов представления в них языковых данных, важность самих теоретических и практических лингвистических исследований на базе этих данных привели к созданию самого нового направления в языкознании — корпусной лингвистики [5, с. 3]. В корпусной лингвистике также определены типы языковых корпусов и общие методологии их составления. Составители корпусов выделяют

понятия корпуса и национального корпуса. Возникает вопрос “что есть национальный корпус языка и чем он отличается от простого корпуса?” В корпусной лингвистике разработчиками даются схожие определения понятия национального корпуса языка, выделяются критерии, отличающие его от просто корпуса.

«Национальный лингвистический корпус – огромная коллекция устных и письменных текстов различных жанров, стилей, региональных и социальных вариантов, представленных в языке и интересных для изучения языка» [17, с. 100].

«Национальный корпус – это собрание текстов в электронной форме, представляющих данный язык на определенном этапе его существования, отображающий данный язык во всем многообразии жанров, стилей, социальных и территориальных диалектов и т.п. Корпус должен быть представительным, т.е. содержать по возможности все типы письменных и устных текстов» [4, с. 43].

Составители Национального корпуса русского языка также указывают, что национальный корпус имеет две важные особенности, одним из которых является содержание в нем всех типов письменных и устных текстов, представленных в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т.п.) [12].

Анализ вышерассмотренных определений показывает, что любой корпус, который не содержит устную или письменную составляющую или не представляет все стили и жанры языка, не может претендовать на звание национального корпуса, является просто корпусом.

Созданные корпусы башкирского языка являются базой для создания реального Национального корпуса башкирского языка.

В рамках реального Национального корпуса башкирского языка предстоит работа по охвату таких стилей и жанров языка как учебный и научный (включение школьных, сузовских и вузовских учебников, гуманитарных научных публикаций, материалов из энциклопедий), официально-деловой (текстов

законов и указов, документов), художественной драматургии и др.

В корпусной лингвистике определен минимальный объем корпуса в словоупотреблениях, когда он может считаться национальным. Большинство создателей и исследователей считают, что объем национального корпуса должен превышать 100 млн. словоупотреблений [5; 3]. Такие требования большого объема для национального корпуса основываются на мнении Синклера [21] о том, что сколько-нибудь значащие статистические лингвистические данные и результаты можно получать только на достаточно большом объеме материала.

Следующим из важных требований к национальному корпусу является репрезентативность. В корпусной лингвистике репрезентативность понимается как представление в корпусе всех видов речевых и письменных материалов. В частности, для письменных — текстов различных периодов, жанров, стилей, авторов и т. п. [5, с. 5; 4, с. 47]. Репрезентативность должна обеспечить национальному корпусу возможность получать представления о языке в целом.

Репрезентативность тесно связана со сбалансированностью корпуса, которая достигается пропорциональным представлением разнообразных лингвистических материалов по стилю, жанру и тематике [10, с. 248]. Данное понятие как основное в корпусной лингвистике, тем не менее не является тривиальным. Так, по анализам Д. Байбера, корпус, составленный на базе пропорционального охвата типов речи и языковых стилей, должен был бы содержать около 90% обычной разговорной речи, 3% писем и замечаний и 7% опубликованных и неопубликованных текстов классических стилей и жанров [19, с. 20]. Учитывая трудоемкость составления корпусов устной речи, в настоящее время составление корпуса, реально отражающего язык в представленных выше пропорциях, представляется, мягко говоря, весьма проблематичным.

Однако определение репрезентативности и сбалансированности к национальному корпусу как содержание всех типов письменных и устных текстов, и их разных жанров, по возможности пропорционально их доли в языке,

декларируется как основное требование и в Национальном корпусе русского языка [12]. Сегодня составители национальных корпусов, не касаясь проблемы соотношения устных и письменных видов речи, которая не решится в ближайшее время, предлагают разные пропорции включения в устную и письменную части корпусов стилевых и жанровых типов языковых материалов, в том числе с опорой на результаты социолингвистического анкетирования [15, с. 41.]

Вопросы репрезентативности и сбалансированности будущего Национального корпуса башкирского языка с учетом лингвистических и экстралингвистических факторов должны стать объектом дальнейшего совместного обсуждения широкого круга филологов республики. На начальном этапе нами для текстовой части национального корпуса предлагается следующая пропорция: проза – 40%, публицистика – 40%, учебная и научная литература – 5%, фольклор – 5%, поэзия – 5%, официально деловой стиль – 5%.

Проблема репрезентативности, как важнейшая для национального корпуса, должна учитываться на всех этапах реализации и эксплуатации [5, С,5]. Строгое же соблюдение сбалансированности нами ставится как задача финальная, а не соблюдаемая на начальных этапах построения Национального корпуса башкирского языка, как и удовлетворения национального корпуса требованию полноты.

В корпусной лингвистике удовлетворение национального корпуса требованию полноты является следующим важным критерием, отличающим его от простых лингвистических корпусов. Полнота корпуса достигается тогда, когда корпус находит примеры употребления для всех слов и грамматических конструкций языка за исключением тех, которые: принадлежат к сугубо индивидуальному словоупотреблению; являются ошибкой, ненормативным употреблением; представляют собой анахронизм, явно устаревшее словоупотребление; являются не ассимилированным заимствованием [9].

Структурированность представления данных также является одним из важных требований к национальному корпусу. Какие информативные единицы будут закладываться в базу корпуса, что может пользователь искать и находить в корпусе? В существующих корпусных проектах башкирского языка заложена морфологическая информация, которая включает а) частеречную характеристику; б) совокупность морфологических словоизменительных признаков. В

национальном корпусе должна присутствовать информация по семантике слова, работа над которой ведется. Несомненно должна быть представлена и синтаксическая информация на основе синтаксической разметки текстов. Поиски в этом направлении начаты.

Считаем, что в национальном корпусе должна быть представлена информация по лексико-грамматической омонимии и осуществлена возможность автоматического снятия этих неоднозначностей. Данная проблема требует проведения практических исследований на основе корпусных проектов и теоретического обобщения.

Создание простых корпусов типа текстов определенного писателя, исторических документов, речевых данных социальных групп в синхронии и других может быть одномоментным решением: создали, запустили в сети Интернет для всеобщего пользования и больше к нему не возвращаются. Но в случае с национальным корпусом такой подход невозможен. Соблюдение сбалансированности и полноты национального корпуса не может быть одномоментным решением. Национальный корпус изменяется вместе с самим языком и требует постоянного участия коллектива опытных лексикографов на протяжении всего времени существования корпуса. Поэтому одним из главных требований к национальному корпусу является наличие постоянного лингвистического сопровождения компетентными специалистами-филологами. Так, для создания национального корпуса чешского языка в 1994 году при Университете Карлова был создан Институт Чешского Национального Корпуса [22].

Отметим, что наличие национального корпуса является определяющим признаком уровня технического и экономического развития государства [14]. Задача его создания включена в Государственную программу “Сохранение и развитие государственных языков Республики Башкортостан и языков народов Республики Башкортостан” на 2019-2024 гг. Поэтому мы надеемся на то, что фронтальная работа по созданию Национального корпуса башкирского языка с вышеизложенными концептуальными положениями будет начата в ближайшее время.

ЛИТЕРАТУРА

1. Башкирский поэтический корпус. [Электронный ресурс]. URL:http://web-corpora.net/bashcorpus/search (дата обращения: 10.11.2020).

2. Бектаев К. Б. Статистико-информационная типология тюркского текста. Алма-Ата: Наука, 1978. – 167 с.

3. Богоявленская Ю. В. Репрезентативность лингвистического корпуса: метод верификации достоверности полученных данных // Политическая лингвистика, 2016. – С.163–166.

4. Волоснова Ю. А. Корпусная лингвистика: проблемы и перспективы // Лесной вестник. Филология, 7/2006, – С.43–49.

5. Захаров В. П. Корпусная лингвистика: Учебно-методическое пособие. – СПб., 2005. – 48 с.

6. 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.

7. Бускунбаева Л.А., Сиразитдинов З.А., Ишмухаметова А.Ш., Ибрагимова А.Д., Мигранова Л.Г. Корпус текстов периодической печати на башкирском языке // Актуальные проблемы диалектологии языков народов России. Материалы XII региональной конференции. Уфа: Институт истории, языка и литературы Уфимского научного центра РАН. 2012. С. 139-141.

8. 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). 2014. С. 86-89.

9. Корпус русского литературного языка. [Электронный ресурс]. URL:http://narusco.ru/project.htm (дата обращения: 10.11.2020).

10. Мазынская С.В. Сопоставительный анализ англоязычных корпусов текстов, доступных онлайн // Карповские научные чтения : сб. науч. ст. : в 2 ч. / Белорус. гос. ун-т ; ред. кол.: А.И. Головня [и др.]. – Минск, ИВЦ Минфина, 2017. – Ч. 1. – С. 284-286.

11. Национальный корпус башкирского языка [Электронный ресурс]. URL: https://bashcorpus.ru/#about_corpora. (дата обращения: 17.09.2020).

12. Национальный корпус русского языка. [Электронный ресурс]. URL: http://www.ruscorpora.ru (дата обращения: 17.09.2020).

13. Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении, 2008, № 16 (2), – С. 7–20.

14. Плунгян В.А. Насущная практическая задача для описания языка — это создание его полного электронного корпуса. [Электронный ресурс]. URL:https://www.kommersant.ru /doc/2718287

15. Резанова З.И. Лингвистический корпус «Томский региональный текст»: типологически релевантные параметры сбалансированности и репрезентативности// Вестник Томского государственного университета. Филология. 2015. №1 (33 С.38-50.

16. Сиразитдинов З.А., Бускунбаева Л.А., Садыков Т.С. О принципах создания фонетических корпусов тюркских языков: на примере материалов говоров башкирского языка//Alatoo Academic Studies. 2019. № 3. P.96-105.

17. Сысоев П. В. Лингвистический корпус в методике обучения иностранным языкам // Язык и культура. 2010. № 1. – С. 99–111.

18. Устный корпус башкирского языка дер. Рахметово и с. Баимово [Электронный ресурс]. URL:https://linghub.ru /oral_bashkir_corpus/) (дата обращения: 10.11.2020).

19. Biber D.: Reprezentativnost v projektu korpusu. Studie z korpusové lingvistiky. Acta Universitatis Carolinae. Philologica 3-4. Praha: Univerzita Karlova – Nakladatelství Karolinum 2000, – С. 107–136.

20. Leech G. Computers in English language research / G. Leech, A. Beale //Cambridge language teaching surveys. – Cambridge: Cambridge University Press, 1985 – Vol. 17(3). – P. 5–18.

21. Sinclair J. Corpus. Concordance and Collocation/Oxford: Oxford University Press, 1991. – 137 p.

22. Чешский национальный корпус ČNK [Электронный ресурс]. URL: http://czechkorpus.blogspot.com/. (дата обращения: 17.09.2020).

Sirazitdinov Z.A., Shamsutdinova G.G., Buskunbaeva L.A.
Institute of History, Language and Literature,
Ufa Research Center RAS, Russia, Bashkortostan, Ufa

ABOUT DEVELOPMENT OF THE AUDIO CORPUS OF THE EASTERN DIALECT OF THE BASHKIR LANGUAGE

Abstract. The article discusses the principles of creating a corpus of audio materials of the Eastern dialect of the Bashkir language. The task is to representatively collect field materials taking into account the age, gender, level of education, language of communication, national identity of informants and the variety of topics for conversation.

The corpus is intended to present audio recordings, transcripts of audio files in the form of transcribed texts, their literary version and Russian translation of the dialect text.

The authors propose transcription in a "semi-orthographic record" that is widely used in speech corporas. The developed system of graphic transmission of dialect features is based on the use of special characters and letter combinations. Dividing the informant's thematic monologue into phrases and syntagmas, transcribing and syncing the transcript with the audio file is performed in the Elan annotation program. Prosodic speech characteristics, such as accent, intonation, and tone, are not taken into account at this stage of transcribing dialect material.

Keywords: *Bashkir language, Turkic languages, linguistic corpus, corpus linguistics, dialectology, expedition material.*

Сиразитдинов З.А., Шамсутдинова Г.Г., Бускунбаева Л.А.
Институт истории, языка и литературы УФИЦ РАН,
Россия, Башкортостан, Уфа

О РАЗРАБОТКЕ АУДИОКОРПУСА ВОСТОЧНОГО ДИАЛЕКТА БАШКИРСКОГО ЯЗЫКА

Аннотация. В статье рассматриваются принципы создания корпуса аудиоматериалов восточного диалекта башкирского языка. Ставится задача репрезентативного сбора полевых материалов с учетом возрастной, гендерной принадлежности,

уровня образования, языка общения, национальной идентичности информантов и разнообразия тем для беседы.

В корпусе предполагается представление аудиозаписи, расшифровки звуковых файлов в виде транскрибированных текстов, их литературный вариант и русский перевод диалектного текста.

Авторами предлагается транскрипция в «полуорфографической записи», широко распространенной в речевых корпусах. Разработанная система графической передачи диалектных особенностей основана на использовании специальных знаков и буквосочетаний. Членение тематического монолога информанта на фразы и синтагмы, транскрибирование и синхронизация транскрипта с аудиофайлом осуществляется в программе аннотирования ELAN.

Ключевые слова: *башкирский язык, тюркские языки, лингвистический корпус, корпусная лингвистика, диалектология, экспедиционный материал.*

Введение. Ареалом распространения восточного диалекта башкирского языка являются Абзелиловский, Баймакский, Белокатайский, Белорецкий, Бурзянский, Кигинский, Мечетлинский, Салаватский, Учалинский районы Республики Башкортостан, а также прилегающие к ним отдельные районы Челябинской, Курганской и Свердловской областей.

Восточный диалект башкирского языка не раз становился объектом исследования диалектологов. Говоры диалекта изучены методами лингвистической географии, описаны монографически, изданы многочисленные научные исследования, составлены словари.

Изданный в 2005 г. «Диалектологический атлас башкирского языка» включает в себя уникальный материал, собранный диалектологами ИИЯЛ УНЦ РАН в 1973–1983 гг. и показывающий территориальные распределения фонетических, лексических и грамматических явлений по 250 опорным пунктам Республики Башкортостан и сопредельных областей [2]. Данный атлас также содержит значительный материал по восточному диалекту.

Лексикографические диалектные материалы, диалектологический атлас башкирского языка и транскрибированные тексты из образцов речи представлены в виде диалектологической базы данных в Машинном фонде башкирского языка [4].

Все комплексные исследования по данному диалекту (по другим диалектам ситуация схожая) проводились до 80-х гг. XX столетия. К сожалению, с 80-х гг. XX в. и по настоящее время диалекты остаются в стороне от фронтальных полевых исследований. Ни корпуса по материалам восточного диалекта, ни самих оцифрованных материалов для создания корпуса не существует.

Данный проект призван начать сбор и обработку обширного материала с охватом широкого круга информантов по поло-возрастным и другим социальным группам с перспективой создания представительного корпуса, включающего в себя аудиоматериалы и их транскрипции по говорам восточного диалекта башкирского языка.

Принципы сбора полевых материалов и разработка метаразметки.

1) Сбор диалектного материала производится по всем 6 говорам восточного диалекта башкирского языка: айский, сальютский, аргаяшский, миасский, кизильский и учалинский.

2) По каждому населенному пункту предполагается осуществить запись по гендерному признаку, по возрастным группам. Выделяются следующие возрастные группы:

– дошкольный и начальный класс (до 11 лет);

– средний школьный (от 11–15 лет);

– старший и студенческий возраст (16–25 лет);

– средний возраст (25–45 лет);

– старший возраст (45–65 лет);

– пожилой возраст (от 65 лет).

Учитывается образование информанта: начальное, среднее, высшее.

3) Аудиозапись сопровождается информацией, которая составляет экстралингвистическую разметку аудиофайла:

– гендер: мужской, женский;

– образование: начальное, среднее (среднее школьное или суз), высшее;

– возраст;

– язык обучения: башкирский, русский, татарский, чувашский;

– язык общения в семье: башкирский, русский, татарский;

– национальность информанта;

– имя, отчество, фамилия;

– место проживания;

– место последнего долгого проживания до переезда в данное место (в случае переезда);

– время проживания до последнего места проживания (в случае переезда);

– время записи.

4) Выделяется тип общения: монолог, диалог, полилог.

5) Для записи информантов определены следующие 15 тем:

– свадьба, свадебные обычаи;

– блюда (повседневные и праздничные);

– домашние животные (какие держат и как содержат);

– система родства (дети и близкие родственники);

– приусадебное хозяйство (огород, сад);

– дом, постройки (когда и кем построен, какая крыша, рамы);

– топонимия в окрестностях поселения;

– история села, школы, рода;

– повседневная жизнь (работа, школа);

– времена года, погода;

– малые формы фольклора (частушки, пословицы, поговорки, сказки);

– поездка в райцентр (по каким делам, каким транспортом пользуются);

– игры детей;

– друзья (кто они, где они живут);

– животный мир около поселения (какие птицы и звери обитают).

6) Запись осуществляется на цифровой диктофон в несжатом формате PCM (.WAV), при отсутствии посторонних

звуков (16бит/22kHz – 16бит/48kHz). Первичная обработка аудиозаписей (очистка от посторонних шумов и длительных пауз) и паспортизация производятся в программе Sound Forge.

7) Паспортизация файлов и экстралингвистическая разметка осуществляются в базе данных Access.

Принципы транскрибирования и лингвистической разметки. Членение аудиофайла на коммуникативные эпизоды (фразы), транскрибирование и синхронизация транскрипта с аудиофайлом осуществляются в программе аннотирования ELAN. Работа осуществляется в трех уровнях: транскрибирование аудиофайла, представление в литературной форме и перевод на русский язык. Транскрипция выполняется в «полуорфографической записи», широко распространенной транскрипции в речевых корпусных разработках [3]. Эта транскрипция близка к фонематической, максимально приближена к орфографии современного башкирского языка. Была разработана единая система графической передачи диалектных особенностей с использованием специальных знаков и буквосочетаний для обозначения тех звуков, которые невозможно передать при помощи башкирской орфографии [1].

Единицей описания в корпусе являются не только слова, но и знаки сегментирования, символы, обозначающие паралингвистические элементы речи, такие как смех, кашель, вздохи, стоны, причмокивание, плач и другие хезитационные явления, сопровождающие живую речь.

Особое внимание уделяется и непреднамеренным остановкам информационного потока в процессе коммуникации, которые обусловлены целым рядом факторов как индивидуальных, психологических, так и физиологических.

В процессе транскрибирования аудио-файлов вышеуказанные хезитационные явления учитываются и передаются специальными знаками.

Просодические характеристики речи, такие как ударение, интонация, тон, на данном этапе транскрибирования диалектного материала не учитывались.

Если в письменной речи структурно-смысловое членение высказывания осуществляется с помощью пунктуационных

средств, то в устной речи звуковой поток членится на синтагмы (/) (смысловое целое, отделенное небольшой паузой) и фразы (//) (законченное целое, которое может состоять из группы синтагм, но может состоять и из одной синтагмы и которое нормально характеризуется конечным понижением тона).

Специальные знаки транскрибирования:

/ – знак сегментирования синтагмы;

// – знак сегментирования повествовательного высказывания (членение на фразы и синтагмы осуществляется с учетом интонационно-синтаксических характеристик отрезков звуковой цепи);

! ? – знак сегментирования вопросительных и восклицательных фраз;

(..) – знак длительной паузы в потоке речи информатора;

( ) – знак паузы хезитации, если она заполнена некоторыми звуками, соответствующие буквенные символы ставятся внутри скобок: напр., э (э-э);

(//) – обрыв высказывания перед последующим словом;

[ ] – самокоррекция (неправильное слово или грамматическая форма, которые в последующем исправляются информантом);

.. – обрыв слова перед последующим словом;

$ Н – не поддающееся расшифровке или неуверенно расшифрованное слово или словосочетание;

$ С – паралингвистический элемент речи – смех;

$ К – паралингвистический элемент речи – кашель;

$ В – паралингвистический элемент речи – вздох, стон, причмокивание и др.

... # – переключение языковых кодов (с башкирского на русский);

+ – для указания таких языковых явлений, как элизия и ассимиляция (+баралманы – бара алманы, +ҡайтыб*ара – ҡайтып бара/

Заключение. На сегодняшний день обработаны аудиофайлы 52 информантов по Бурзянскому, Учалинскому и Дуванскому районам Республики Башкортостан. Предстоит сбор полевых материалов по населенным пунктам остальных

районов, которые были определены как опорные для построения Диалектологического атласа башкирского языка.

В дальнейшем планируется разработка корпуса-менеджера для осуществления поисковых запросов по транслитерации и литературному эквиваленту.

Корпус восточного диалекта башкирского языка предоставит возможность свободного доступа лингвистов различной специализации к первичному диалектному материалу, выступит в качестве бесценного источника для изучения отдельного говора, установления территории распространения того или иного языкового явления, станет базой для изучения исторического развития и становления литературного языка, социолингвистического анализа, сравнительно-сопоставительных исследований языков.

ЛИТЕРАТУРА

1. Бускунбаева Л.А., Сиразитдинов З.А. Разработка аудиокорпуса восточного диалекта башкирского языка: проблемы и перспективы // Известия Уфимского научного центра РАН. 2020. № 2. с. 90–97.

2. Диалектологический атлас башкирского языка / Под ред. Ф.Г. Хисамитдиновой. Уфа, 2005.

3. Кибрик А.А., Подлесская В.И. К созданию корпусов устной русской речи: принципы транскрибирования // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2003. № 10. С. 5–13.

4. Сиразитдинов З.А., Бускунбаева Л.А., Каримова Р.Н. Диалектологическая основа Машинного фонда башкирского языка как инструмент исследования континуума башкирского диалекта. Oriental Studies. 2016; №9 (1). С.: 212-219.

Ubaleht I.P.
Omsk State Technical University,
Russia, Omsk

THE CREATION OF SPEECH CORPUS OF THE DIALECTS OF THE SIBERIAN TATARS

Abstract. In this paper we present the first results of the creation of speech corpus of the dialects of the Siberian Tatars. The collected speech data were published, are accessible to the public, and are licensed under CC BY. This corpus will be used to the Lexeme system. Lexeme – is new application for managing speech corpora for under-resourced languages. Now, the Lexeme system is under development.

Keywords: *The Tatar Language, The Dialects of the Siberian Tatars, Open Speech Corpora, The Under-resourced Languages.*

Убалехт И. П.
Омский государственный университет,
Россия, Омск

СОЗДАНИЕ РЕЧЕВОГО КОРПУСА ДИАЛЕКТОВ СИБИРСКИХ ТАТАР

Аннотация. В данной статье показаны первые результаты работы по созданию речевого корпуса диалектов сибирских татар. В настоящее время собранные речевые данные опубликованы в соответствии с лицензией Creative Commons Attribution 4.0 (CC BY 4.0) и находятся в открытом доступе. Данный речевой корпус планируется использовать с приложением "Лексема". "Лексема" – это новое приложение для работы с речевыми корпусами малоресурсных языков. В настоящее время приложение "Лексема" находится в стадии разработки.

Ключевые слова: *татарский язык, диалекты сибирских татар, открытые речевые корпуса, малоресурсные языки.*

1. Introduction

Nowadays, the field of the applied scientific research which includes the tasks of automation of activity linguists and other specialists who works with speech data is being dynamically developed. There are following tasks in this applied field: the organization of audio data structured storage, the speech segmentation into phrases, words and phonemes, the creation of multi-level speech data annotations, the organi-zation of queries to speech databases, the statistical analyseof speech data and other tasks. Solving these above mentioned tasks is essential for well-resourced languages as well as for under-resourced languages. Specialized software is being used for solving above mentioned tasks.

At present, there are enough software solutions which allow working with speech corpora. There are following stand-alone applications: IrcamCorpusTools[1], EXMARaLDA[2], LaBB-CAT [3] and newer systems such as SPPAS [4]. The following modern software solutions are based on a client-server model: the EMU Speech Database Management System [5] and ISCAN [6]. For under-resourced languages LingSync and the Online Linguistic Database [7] can be note.However, the need to develop new solutions in this area, especial for under-resourced languages, remains animportant challenge. The briefly review of our solution in Section 2. In the past few years, many software instruments and corpora for the Tatar language have been appeared [8,9] but Tatar dialects remains are poorly resourced and documented.In Section 3, we described our first results of the creation of speech corpus of the dialects of the Siberian Tatars.

2. The Lexeme System

Lexeme is a new system provides following features: the storage of audio data, data processing, representing of speech information to users. This system will have special features for the documentation and revitalization of endangered languages. Lexeme is based on the following key principles:

- Openness and transparency (all code and data (including primary audio data) will be accessible on GitHub and licensed under CC BY)

- Universality (the system will consist of independent levels, users can use artifacts irrespective to level for own projects)
- Targeted at different users (linguists, computational linguists, speakers of endangered languages, language activists).

Nowadays, the Lexeme system is under development. We describe current status of the creation of one of the corpora for the Lexeme system in Section 3. We created two more corpora for the Lexeme system: corpus of Siberian Ingrian Finnish (speech data of this corpusare available on GitHub and licensed under CC BY 4.0) [10] and corpus of the dialects of the Siberian Estonians (speech data of this corpus have not yet been published).

3. Speech Corpus of the Dialects of the Siberian Tatars

3.1. Language Context

The dialects of the Siberian Tatars are dialects used by the Tatars who living in the Tyumen, Omsk, Novosibirsk, Tomsk and Kemerovo region. These dialects are relatively well-studied [11,12], around 100,000 people are spoken in these dialects but nonetheless these dialects arepoorly resourced.

The language of Siberian Tatars has three dialects: Tobol-Irtysh, Tom and Baraba. Tobol-Irtysh dialect consists of following subdialects: Tyumen, Tobol, Zabolotny, Tevriz and Tara. The speech data of our first expedition were recorded in Tevriz subdialect area.

3.2. The Current Status of the Creation of Speech Corpus of the Dialects of Siberian Tatars

Our first expedition was undertaken to Siberian Tatar village Ilchebaga (Ust-Ishimsky district, Omsk oblast, Russia) in 2020. We recorded the speech data of 10 speakers in this first expedition. These primary speech data already were published and are accessible to the public [13]. These speech data are available on GitHub and licensed under CC BY 4.0.

Amount of primary audio data and characteristics of speakers are shown in Table 1. We started creating the Siberian Tatar speech corpus based on this data. We plan to collect speech material of all dialects and accents of the Siberian Tatars for this speech corpus.In

2020, we couldn't record more speech data because of the coronavirus COVID-19 pandemic.

Table 1.

Speech data from expedition to Ilchebaga: breakdown by speaker

The code of the speaker and gender	The Year of Birth	The current place of residence	The place of the birth	Birthplaces of parents or/and ethnic of parents	Speech data (duration, in minutes)
AVN-69 (M)	1969	Ilchebaga	Ilchebaga	Both parents:Siberian Tatars	2,5
GMG-67 (M)	1967	Ilchebaga	Ilchebaga	Both parents:Volga Tatars	2,5
GNSh-29 (F)	1929	Ilchebaga	Ilchebaga	Three grandparents: Volga Tatars, one grandparent:Siberian Tatars	24,5
KMM-63 (M)	1963	Ilchebaga	Tavinsk	Both parents: Tavinsk (Siberian Tatars)	49
MKhU-50 (F)	1950	Ilchebaga	No data	Father: Tavinsk, mother: Tebendya, both Siberian Tatars	32
MRCh-60 (M)	1960	Ilchebaga	No data	Three grandparents: Erbagul, one grandparent: Ilchebaga	34
NGA-45 (F)	1945	Ilchebaga	Yarkovo	Father: Volga Tatars, mother: Siberian Tatars	12

NIA-53 (M)	1953	Ust-Ishim	Kuchum	Father: Kuchm (Siberian Tatars), mother: Volga Tatars	9
SGL-61 (M)	1961	Ilche-baga	No data	No data	5
Anonym Speaker (M)	No data	Ilche-baga	No data	Mother: Siberian Tatars, father: Bukharian Tatars	4

4. Conclusion

In this paper, we have presented our current results of the creation of speech corpus of the dialects of the Siberian Tatar. For the first time, the speech data from Tevriz subdialect area were published and are accessible to the public. Furthermore, we briefly reviewed key principles of the Lexeme system.

REFERENCES

1. Veaux, C., Beller, G., Rodet, X.: Ircam Corpus Tools: an extensible platform for speech corpora exploitation. In: LREC, pp. 1-7. Marrakech, Morocco (2008).
2. Schmidt, T., Wörner, K.: EXMARaLDA–Creating, analysing and sharing spoken language corpora for pragmatic research. Pragmatics, 19(4), 565-582 (2009).
3. Fromont, R., Hay, J.: LaBB-CAT: An annotation store. In: Proceedings of the Australasian Language Technology Association Workshop 2012, pp. 113-117. (2012).
4. Bigi, B.: SPPAS-multi-lingual approaches to the automatic annotation of speech. The Phonetician 111(112), 54-69 (2015).
5. Winkelmann, R., Harrington, J., Jänsch, K.: EMU-SDMS: Advanced speech database management and analysis in R. Computer Speech & Language 45, 392-410 (2017).
6. McAuliffe, M., Coles, A., Goodale, M., Mihuc, S., Wagner, M., Stuart-Smith, J., &Sonderegger, M.: ISCAN: A system for

integrated phonetic analyses across speech corpora. 1322-1326 (2019).

7. Dunham, J., Cook, G., and Horner. J.: LingSync& the Online Linguistic Database: New models for the collection and management of data for language communities, linguists and language learners. In: Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages, pp. 24-33. USA (2014).

8. Suleymanov, D., Nevzorova, O., Gatiatullin, A., Gilmullin, R., &Khakimov, B.: National corpus of the Tatar language "Tugan Tel": grammatical annotation and implementation. Procedia-Social and Behavioral Sciences, vol. 95, pp. 68-74. (2013).

9. Khusainov, A.: Tekhnologiya avtomatizatcii sozdaniya i otcenki kachestva programmnykh sredstv analiza rechi s uchetom osobennostey maloresursnykh yazykov. Diss. Kazan (2014).

10. Corpus of the Siberian Ingrian Finnish language, https://github.com/ubaleht/SiberianIngrianFinnish, last accessed 2020/10/12.

11. Akhatov, G.: Dialekt Zapadnosibirskikh Tatar. Bashkirskii gos. universitet, Ufa (1963).

12. Tumasheva, D.: Slovar Dialektov sibirskikh Tatar.Izdatelstvo Kazanskogo universiteta, Kazan (1992).

13. Corpus of the dialects of the Siberian Tatars, https://github.com/ubaleht/SiberianTatar, last accessed 2020/10/12.

Khusainov A.F.
Institute of Applied Semiotics of the AS of the RT,
Russia, Tatarstan, Kazan

TOOL FOR DISTRIBUTED CREATION OF ANNOTATED SPEECH BODIES

Abstract. In this paper we present the results of our work on ensuring the technological possibility of creating the first annotated corpus of spontaneous speech for the Tatar language. We developed the necessary distributed client-server system and will describe the formation of the architecture of the software, the development of a database, the implementation of the necessary algorithms for the distributed collection and annotation of speech fragments, as well as the introduction of methods for automatic analysis of the audio signal. The developed software can be accessed via collect.speech.tatar web-site. As the first project implemented on the basis of this site, we have chosen to create the first Tatar broadcast speech corpus. The source of the audio includes one month of TV broadcasts of "TNV Planeta" channel, which give us in total more than 700 hours of speech.

Keywords*: neural machine translation, multilingual datasets, Turkic languages*

Хусаинов А.Ф.
Институт прикладной семиотики АН РТ,
Россия, Татарстан, Казань

ИНСТРУМЕНТ ДЛЯ РАСПРЕДЕЛЕННОГО СОЗДАНИЯ АННОТИРОВАННЫХ РЕЧЕВЫХ КОРПУСОВ

Аннотация. В данной статье представляем результаты нашей работы по обеспечению технологической возможности создания первого аннотированного корпуса спонтанной речи

для татарского языка. Нами описывается формирование архитектуры программного обеспечения, структура базы данных, необходимые алгоритмы распределенного сбора и аннотирования фрагментов речи для разработанной распределенной системы клиент-сервера. В статье также затрагиваются вопросы внедрения этих методов для автоматического анализа звукового сигнала. Доступ к разработанной программе можно получить через сайт collect.speech.tatar. В качестве первого проекта, реализованного на базе этого сайта, нами создан первый корпус татарской трансляции речи. Источником аудиозаписи являются телетрансляции телеканала «ТНВ-Планета» за один месяц, которые дают нам в общей сложности более 700 часов выступления.

Ключевые слова: *нейронный машинный перевод, многоязычные наборы данных, тюркские языки.*

1. Введение

Системы автоматического распознавания речи используются для взаимодействия с большинством современных электронных устройств и приложений, таких как, например, интеллектуальные мобильные ассистенты (Siri, Cortana, GoogleVoice), умные колонки (AmazonEcho, Яндекс Колонка). Для построения подобных систем используются, в том числе, большие объемы аудио фрагментов с указанием произнесенного в них текста (размеченные аудио корпуса). Для надежной работы итоговой системы объемы таких корпусов должны составлять несколько тысяч часов речи. Создание корпуса такого объема невозможно в лабораторных условиях.

Создание распределенной системы пополнения и разметки аудио корпуса татарского языка станет важнейшим фактором, обеспечивающим возможность создания подобного корпуса для татарского языка и способствующим дальнейшему развитию современных технологий для татарского языка. В перспективе это позволит поддержать тренд на более активное использование татарского языка в современном информационном пространстве, будет способствовать созданию

необходимых лингвистических ресурсов и их использованию в самых современных технологиях искусственного интеллекта для татарского языка. Работа над созданием первого корпуса спонтанной татарской речи также будет способствовать развитию целого ряда междисциплинарных исследований: от изучения фонетики современного татарского языка до построения нейросетевых акустических моделей татарских фонем.

Таким образом, проводимые исследования и разработки будут направлены как на изучение, так и на сохранение и развитие татарского языка в современных цифровых условиях.

Во втором разделе данной статьи представлен обзор исследований в области речевых технологий для татарского языка, раздел 3 содержит описание созданных инструментальных средств, раздел 4 - перспективы их использования для построения корпуса эфирной татарской речи.

2. Обзор

Область автоматического анализа татарской речи развивается в Институте прикладной семиотики АН РТ с 2010-х годов. За это время поэтапно (с возрастанием сложности стоящих задач) были успешно разработаны системы распознавания изолированно произнесенных команд, идентификации речи говорящего, распознавания слитной читаемой речи [1, 2]. Разработка системы распознавания слитной читаемой речи потребовала создания первого для татарского языка многодикторного речевого корпуса [3]. Он формировался на протяжении более 7 лет и на текущий момент состоит из записей 500 дикторов разного пола и возрастов. Запись осуществлялась на специально настроенном компьютере, в специализированном программном обеспечении и с использованием профессиональной звукозаписывающей аппаратуры; соблюдалось выполнение строгого регламента записи, контролировались условия окружающего шума, а также правильное произношение диктором всех фраз. Созданный корпус был использован для построения системы распознавания так называемой подготовленной читаемой речи: такая речь не

содержит экстралингвизмов, отсутствуют запинания, повторы слов/частей слова, заикания, заполненные паузы и т.д.

Следующим этапом развития речевых технологий для татарского языка является построение системы распознавания спонтанной речи. Данная задача требует подготовки специализированного набора алгоритмов, и, что не менее важно, специализированного корпуса речи. Его объем должен составлять тысячи часов, соответственно, он не может быть создан по той же технологии, что и предыдущий речевой корпус. Поэтому было решено разработать необходимые программные средства, которые бы обеспечили технологическую возможность создания первого аннотированного корпуса спонтанной речи для татарского языка за счет распределенной клиент-серверной архитектуры.

3. Разработанные программные средства для создания аннотированного речевого корпуса

Исходя из стоявшей цели обеспечить возможность создания первого аннотированного корпуса спонтанной речи для татарского языка были поставлены и решены следующие задачи:

1. формирование архитектуры разрабатываемых программных средств;
2. разработка базы данных;
3. реализация необходимых алгоритмов для распределенного сбора и аннотирования речевых фрагментов;
4. внедрение методов автоматического анализа аудио сигнала.

Разработанная система представляет собой клиент-серверную архитектуру на базе ASP.NetCore [4], React.js [5] и построена на основе подхода предметно-ориентированного проектирования (DomainDrivenDesign).

Весь программный комплекс был разделен на следующие слои:

1. Слой инфраструктуры: главной задачей слоя является обеспечение других уровней интерфейсом доступа к данным.

Для непосредственной работы с БД использовался Entity Framework Core.

2. Слой предметной области: основной уровень, реализующий логику предметной области работы с аудио-коллекцией. Он включает в себя функции взаимодействия с абстракциями уровня инфраструктуры.

3. Слой приложения: уровень приложения обеспечивает выполнение запрошенных функций приложения за счет использования объектов и сервисов уровня предметной области. Используются специализированные «обертки»над различными типами объектов для получения и возврата данных на уровень представления.

4. Распределенный сервисный слой: этот уровень используется для обслуживания функций приложения через RESTAPI-интерфейс. Он не реализует никакой логики предметной области, а только транслирует HTTP-запросы. На этом уровне также реализуется функционал авторизации, кешированияи т.д.

5. Слой представления: ASP.NET Core приложение рассматривается как уровень представления.

6. Клиентское приложение: веб-приложение, с которым непосредственно взаимодействуют пользователи.

В качестве СУБД была выбрана PostgreSQL. Помимо отношений, необходимых для функционирования базовых функций сайта (система ролей, система проектов, логирование и т.д.), были созданы следующие таблицы: Audio Segments, Audio Segment Annotations, Audio Segment Annotation Validations, рис. 1.

Для конечных пользователей сайт предоставляет следующую базовую функциональность:

1. возможность загрузки новых аудиофайлов в базу данных;

2. автоматический анализ загруженных файлов; разбиение файлов на фрагменты, разделенные сегментами, не содержащими речи;

3. веб-форма для краудсорсинговой разметки загруженных фрагментов: возможность прослушивания и аннотирования фрагмента;

4. веб-форма для экспертной валидации разметки фрагментов: прослушивание записи, выбор правильного варианта разметки из созданных разными пользователями или создание собственной версии разметки;

5. контроль статуса разметки фрагментов администратором проекта.

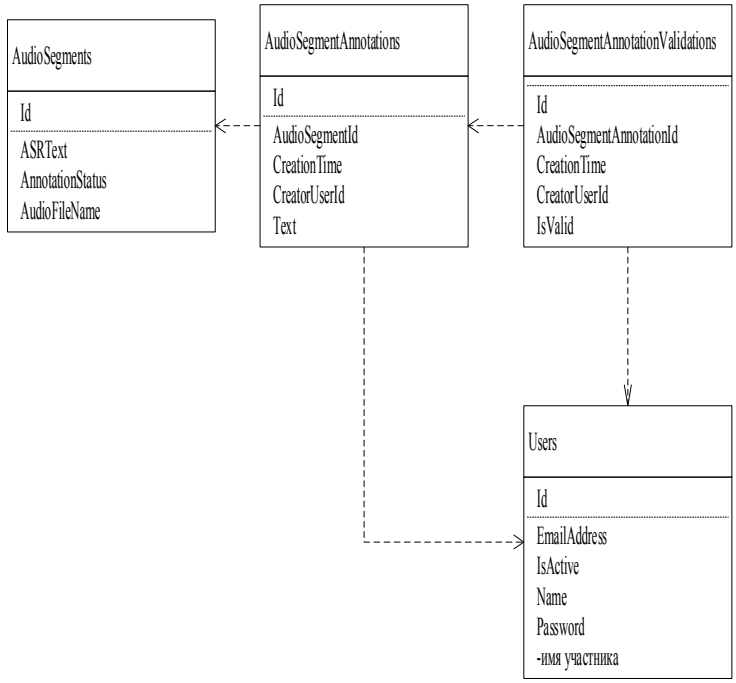


Рис. 1. Схема отношений БД

Для реализации перечисленных функций были разработаны следующие веб-страницы:

- режим просмотра статуса сегментов (возможные значения: «не размечен», «размечен», «проверен») и загрузки/удаления аудио-фрагментов (рис. 2);
- режим разметки: с возможностью прослушивания и аннотирования записи (рис. 3);

- режим валидации: эксперт видит аннотации текущей записи и принимает решение о выборе одной из них или создании нового варианта разметки (рис. 4).

Рис. 2. Режим просмотра и загрузки сегментов

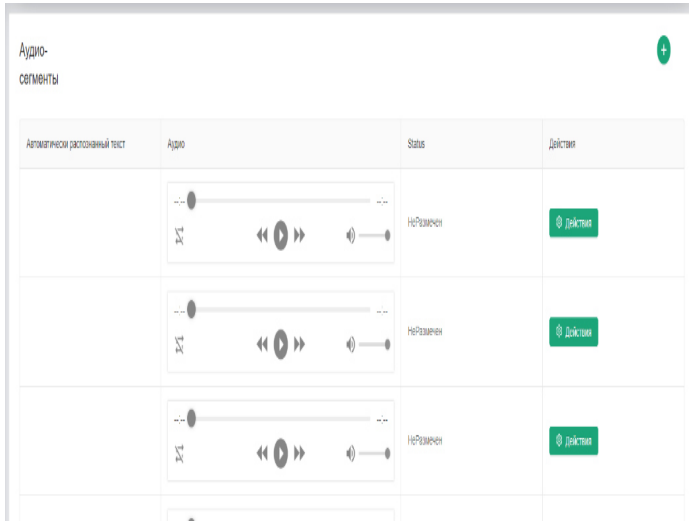


Рис. 3. Режим аннотирования записей

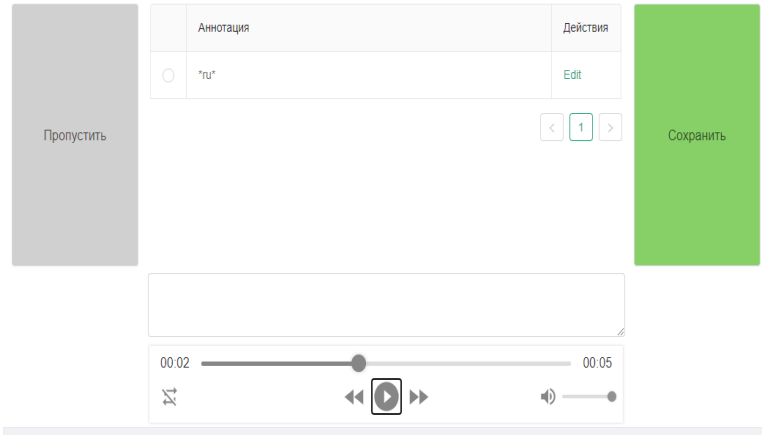


Рис. 4. Режим проверки аннотаций

4. Перспективы создания корпуса эфирной татарской речи

На первом этапе было решено использовать созданную систему для создания аннотированного корпуса эфирной татарской речи. Источником для разметки послужили переданные Институту прикладной семиотики АН РТ видеозаписи телепередач канала «ТНВ Планета» за декабрь 2019 года. Видеозаписи имеют формат AVI, стерео аудиоканал MP3 с частотой дискретизации 48 кГц, скорость потока 96 кб/с. Для разметки и включения в корпус аудиодорожки были выделены из видеофайлов и переконвертированы в форматWAVPCM 16 бит/сек 16 кГц. Общий объем записей составил 732 часа 59 минут и 29 секунд.

Были отобраны передачи, содержащие наиболее подходящий для первоочередного включения в корпус тип аудиозаписей: новостные выпуски, интервью, ток-шоу (передачи "Манзара", "Ком сәгате", "Адымнар", "Канун, парламент, җәмгыять", "Татарлар", "Халкым минем", "Таянуноктасы"). Время начала и окончания указанных передач

было определено вручную, так как имеющаяся на сайте телекомпании информация оказалась не точной. Итогом данного этапа работ стали 40 выделенных аудиофрагментов общей продолжительностью 23 часа 21 минуту и 47 секунд.

Сформированный список аудиозаписей был добавлен в систему, произведена их автоматическая фрагментация на отрывки, разделенные паузами (фрагментами, не содержащими речь) и длительностью не более 15 секунд. Таким образом, для разметки на сайте доступны 22 432 сегмента.

По мере накопления размеченного корпуса планируется обучить систему автоматического распознавания речи и добавить результаты её работы в качестве предварительного результата для упрощения и ускорения работы аннотаторов.

5. Заключение

В данной статье мы представили результаты нашей работы по созданию веб-инструмента для аннотирования речевого корпуса. Реализованный подход предметно-ориентированного программирования позволил обеспечить гибкую архитектуру программных средств. В качестве базовых возможностей были реализованы функции загрузки аудиофайлов, их сегментирования по паузам, краудсорсинговой разметки фрагментов, а также экспретной валидации результатов разметки.

В качестве первого проекта, реализация которого была начата на базе разработанного сайта, был выбран проект по созданию первого корпуса эфирной татарской речи.

Публикация осуществлена в рамках мероприятия 4.1.3. Государственной программы «Сохранение, изучение и развитие государственных языков Республики Татарстан и других языков в Республике Татарстан на 2014-2020 годы».

ЛИТЕРАТУРА

1. Система автоматического распознавания речи на татарском языке / А. Ф. Хусаинов, Д. Ш. Сулейманов // Программные продукты и системы. 2013. № 4. С. 301–304.

2. Программный комплекс для анализа речи (на примере распознавания фонем татарского языка) / А. Ф. Хусаинов // Доклады Томского государственного университета систем управления и радиоэлектроники. 2013. № 3 (29). С. 129–133.

3. Khusainov, A. F., Suleymanov, Dz. Sh. Towards Automatic Speech Recognition for the Tatar Language / A. F. Khusainov, Dz. Sh. Suleymanov // Proc. of the 16th International Workshop on Computer Science and Information Technologies (CSIT'2014). (Sheffield, September 6-22, 2014). – Ufa: USATU, 2014. – P. 97–100.

4. Документация по ASP.NET. URL: https://docs.microsoft.com/ru-ru/aspnet/core/?view=aspnetcore-3.1 [дата посещения: 1.05.2020].

5. React. JavaScript-библиотека для создания пользовательских интерфейсов. URL: https://ru.reactjs.org/ [дата посещения: 1.05.2020].

СЕКЦИЯ 2
СИСТЕМЫ И ТЕХНОЛОГИИ МАШИННОГО ПЕРЕВОДА

Zhetkenbay L., Sharipbay A., Bekmanova G., Yergesh B.
Eurasian National University named after L.N. Gumilyov, Kazakhstan, Nur-Sultan

DEVELOPMENT OF THE KAZAKH-TURKISH STATISTICAL MACHINE TRANSLATION SYSTEM

Abstarct. The creation of a machine translation system for texts between any pair ofn atural anguages is a complex task. Its solution requires a lot of time and research due to significant historical and grammatical (morphological and syntactic) differences between languages. Although machine translation systems have appeared on the Internet between them ention edpairsof Turkic languages, the qualityisin significant, but for Romance languages such systems were created with the advent of the first computers, and now they are very effective and of high quality, often their translation accuracy reaches 100%. As a result of research work, more than 212 thousand parallel proposals were received. We used them for a statistical machine translation from Turkish into Kazakh language based on phrases.

Keywords: *computer processing of natural language, machine translation, statistical machine translation, Kazakh and Turkish.*

Жеткенбай Л., Шарипбай А., Бекманова Г., Ергеш Б.Ж.
Евразийский национальный университет им.Л.Н. Гумилева, Казахстан, Нур-Султан

РАЗРАБОТКА СИСТЕМЫ СТАТИСТИЧЕСКОГО МАШИННОГО ПЕРЕВОДА ДЛЯ КАЗАХСКО-ТУРЕЦКОГО ЯЗЫКА

Аннотация. Создание системы машинного перевода текстов между любыми парами естественных языков является сложной проблемой, решение которой требует больших

временных и трудовых затрат из-за наличия между ними существенных исторических и грамматических (морфологических и синтаксических) различий. Однако для родственных языков, принадлежащих к одной языковой семье, эта проблема значительно упрощается. Но несмотря на это, для тюркских языков попытки решения проблемы машинного перевода появились недавно. До настоящего времени в интернете не появились эффективные системы машинного перевода между парами тюркских языков, хотя для романских языков такие системы были созданы с появлением первых компьютеров и сейчас они очень эффективны, зачастую правильность их перевода достигает до 100%. В результате научно- исследовательской работы было автоматизировано более 212 тысяч параллельных предложений. Они были использованы в статистическом машинном переводе на основе фраз с турецкого на казахский язык.

Ключевые слова: *компьютерная обработка естественного языка, машинный перевод, статистический машинный перевод, параллельный корпус, казахский язык, турецкий язык.*

1.Introduction

Today there are a lot of texts on the Internet in different languages of the world. This also applies to Turkic-language texts. Consequently, one of the important problems at the moment is not only the ability to quickly and efficiently search for information, but also the ability to create an effective machine translator among the Turkic languages. There fore, machine translation is the most pressing problem in natural language processing. There api development of information technology has led to the fact that, thanks to the digitalization of society in the daily life of a person, the volume of world communications is increasing everyday. Intercultural communication is becoming an integral part of the lives of many people, soothe use of machine translation on the Internet is a daily reality for a regular web user. A number of subtleties, inter language complex it yand uncertainty that occur in all works aimed at natural language processing. In this regard, the demand for work aimed at eliminating language barriers is higher than usual.

Over the past 25 years, numerous studies have been carried out in the field of computer processing of Turkic languages, systems and technologies for the active use of Turkic languages as a language for collecting, processing and transmitting information in cyber space are being developed [1-7]. Most of the work is devoted to research in the field of morphological analysis of texts, automation of morphological annotations [8-19].

There are now advanced systems in the field of machine translation. One of them, available in the menu in more than 100 languages, Google Translate from 2016 to the present translates the language of more than 59 countries using an artificial neural network to improve the accuracy of translation [20]. Similarly, but with less capabilities Yandex, PROMT, Translate.ru translation companies, such as, also use neural machine translation. There is an open-source neuro-machine translation system opennmt from Harvard University. Almost all of these systems cannot produce high-quality machine translation into these languages and vice versa.

Translation from one language to another requires learning the alphabet of the language, vocabulary, grammar rules and semantics of the language. Formulation of grammatical rules for machine translation, development and implementation of algorithms for morphological and syntactic analyzers and synthesizers of words and sentences of natural languages, taking into account their semantics.

2. Kazakh-Turkish parallel corpus

The creation of a parallel corpus in statistical and neuromachine translation is an important and rather difficult task. Currently, there are many electronic corpuses for the Turkic languages. Parallel Corpora is an electronic analogue of parallel translated texts, usually consisting of a set of blocks “original text and one / several of its translations” [21].

Among the parallel corpora, the most popular is the parallel corpus Europarl [22] for statistical machine translation, consisting of the set of works of the European Parliament, the revised corpus of the 36th Parliament of Canada, proposed by natural language processing specialists from the Institute of Information Sciences 2001-1A [23], Chinese-French parallel text news set [24], Anglo-

Percy, Anglo-Vietnamese parallel corpus [25]. The parallel corpus for the Kazakh language and the Russian language [26, 27], created by the research group of Nazarbayev University using various electronic resources, is also considered a great achievement in the field of machine translation of the Kazakh language.

Parallel corpora are used in research in the field of applied linguistics, namely: testing automated translation systems, filling the translation memory system, creating systems for automatic search for translation equivalents, etc. The importance of parallel corpora is associated with the fact that they allow to objectively determine how translators overcome difficulties in practice, and use this data to create models that match reality for aspiring translators. They also play an important role in the study of translation norms in specific sociocultural and historical contexts.

Without a doubt, the main problem in assembling a parallel corpus is the high labor intensity, that is, in most cases, aligning text manually looking at the finished translation takes a lot of time. Manual text alignment may be optimal for only a few thousand sentences. However, if the number of parallel sentences exceeds 100 thousand, then turning to automatic alignment methods is certainly rational and efficient from the point of view of time. Errors are detected when aligning automatic texts, especially in cases where rare words and phrases are encountered in the text. For this reason, with a significant reduction in costs, it becomes necessary to build approaches and algorithms, the leveling quality of which is close to the quality of manual leveling. Below are some of the issues of creating and using the corpus of Kazakh-language texts and parallel translations in Turkish.

Despite the fact that there are no parallel corpora available between the Kazakh and Turkish languages, we tried to solve this problem on our own. We searched for books of any genre (scientific, religious, literary books or articles, studies, interviews, etc.) translated from the Kazakh-Turkish or Turkish-Kazakh languages, scanning them and automatically translating the texts into electronic versions. Many mistakes occurred during the recognition (decoding) of texts in the Kazakh language.

Due to the lack of a program for converting error-free recognition of texts in the Kazakh language, this problem had to be solved manually. Although there was no particular difficulty in recognizing texts in Turkish, there were some errors. In particular, when the letters r and n are adjacent, it follows the letter m (r+n= rn=m), the letter y with the letter R, recognize the word that came in a row as rı = n (r+ı=rı=n), the letter l with the letter t, recognize the word that came in the row as d (tl=d), the letter y with the letter n that came in the row (nı=m), and Fig. 1 is an example from the Kazakh-Turkish parallel corpus.

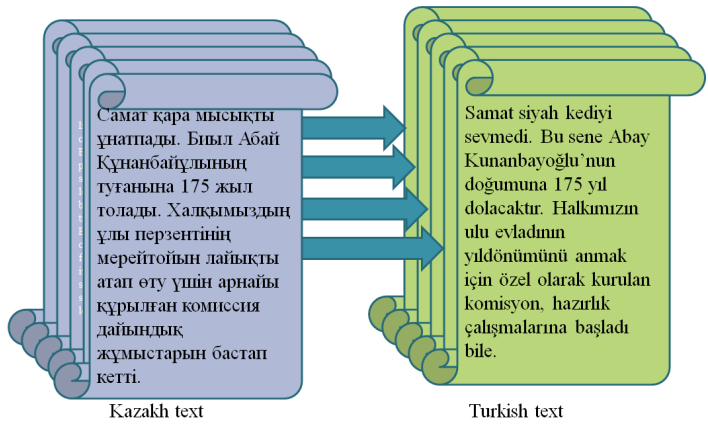


Fig.1. An xamplefrom the Kazakh-Turkish parallel corpus

3. Alignment of texts at the sentence level

Parallel text alignment is the definition of matching sentences in both halves of the parallel text. The alignment of texts at the level of sentences of the parallel corpus is one of the priority tasks in terms of importance and the order of their solution. It is a prerequisite for various aspects of linguistic research. During the translation process, sentences can be split, combined, deleted, inserted or changed in sequence. This makes alignment in most cases challenging. Text-parallel corpuses of samples (in the form of databases) are especially useful for a translator when working with strictly normalized (conditional) texts, the genre stylistic and stylistic design of such

texts practically does not allow variations, deviations from certain sociocultural norms. These texts include official documents, recipes, texts related to weather forecasts, contracts, etc. Just as texts of different styles differ in the vocabulary of lexical units used in certain texts, they also differ in the grammatical and syntactic structure of those proposals included in them.

In most cases, the identification between the sentences of the source and target text is not one-to-one, that is, several sentences of the translation may coincide in one sentence of the source text, or, conversely, some sentences or one paragraph of the source text may be completely omitted, the boundaries of sentences may not coincide, etc. In a couple of texts, it is especially characteristic of works of art that there is no identification between unambiguous sentences and phrases.

At the sentence level, alignment uses word methods, sentences in terms of length, number and static (in terms of the frequency of constituent words), these approaches do not require a developed vocabulary and continue to be used for languages whose resource reserve is not large. Length alignment methods are very sensitive to lagging or addition of sentences, since lagging or addition results in erroneous next alignment from the point of retention / addition to the end of the text.

We used bilingual dictionaries (Kazakh-Turkish dictionary) for automatic text alignment. The dictionary consists of over 76,000 words.

In the course of aligning parallel corpus sentences by rank, the number of all parallel sentences was 289 thousand. In each file, a different quality of parallel sentences was obtained. Later we received parallel sentences that were above the score of 0.2. Because below 0 it is considered a wrong sentence. Then the number of parallel proposals was about 217 thousand. We later removed duplicate sentences. Then the number of generalizing parallel sentences was more than 212 thousand. Below is the statistics of Kazakh-Turkish parallel datasets in Table 1.

Table. 1

Statistics of Kazakh-Turkish parallel datasets

Corpus	Corpus size (Number of sentences)	Number of words in a sentence	Average sentence size
Literary books	82516	909318	11,01
Religious books	56390	620290	11
Scientific books, articles	68500	756200	11,03
Web corpus	4936	59232	11,15
Total	212342	2345040	11,04

4. System of statistical machine translation in Kazakh-Turkish language

The system of statistical machine translation began to develop actively only in the early 90s of the twentieth century, in corpus linguistics, machine learning and information retrieval with the accumulation of data, during the technological and information boom. Therefore, large-scale search engines began to implement the statistical translation service into their interfaces. For example, the search engine Google switched to this technology in 2007 and introduced the Google Translate service. At the beginning of 2011, Yandex, like Google, introduced a system similar to machine translation.

In statistical translation, the main teaching idea is to establish correspondence between tokens in parallel corpuses. At the first stage, the corpus is aligned according to words, after which the algorithm calculates the probability of coincidence between phrases in two texts composed of tokens. In addition, the probabilities of replacing phrases with other phrases are also calculated. At the tuning stage, the algorithm determines the weights of various translation parameters. To make the text look as natural as possible, a language model is used, consisting of a text corpus in the target

language, which allows you to find out the probability of a phrase in this language.

Analytical and automatic metrics are used to test the results of model building. Typically, peer review compares translations of the two models. Ideally, the assessment should be done by a person who speaks both the source language and the target language (it is important to know the target language well). [Papineni et al. 2002] automatic assessments, such as the BLEUS core metrics mentioned in it, imply an assessment of the correspondence of unigrams, bigrams, trigrams, and quadrograms, and also use a list of synonyms in some methods [Banerjee, Lavie 2005]. Also, some methods are based on the assessment that the resulting translations are paraphrases of reference sentences [Russo-Lassner, Lin, Resnik 2005].

The construction of a phrasal table for machine translation begins with establishing a correspondence between words in two parallel corpuses. The GIZA ++ tool is the most popular tool used in Moses and other systems. This tool literally grinds in two directions to determine the final matches. For this reason, if the intersection of the two sets of matches and highly accurate smoothing can be obtained. On the other hand, you can maximize completeness by combining the results.

Accordingly, the next step in training a machine translation model is to build a phrase table. This part of the work is based on the match of previously received words. The Moses system use sthe heuristic method described in [28]. Phrases are smoothing as follows: it is necessary to find the intersection of the first two literal smoothing (text of the target language according to the words of the source language and vice versa). Subsequently, we add to them those that are in contact vertically, horizontally or diagonally with the matching matrix of the union. Starting from the upper left corner (first word), we repeat after the second word and then for the whole sentence. Finally, we add contactless merges to others. The prerequisite here is that the merge points join the literal blend merge. Also, adding each merge should be treated as adding a new word that has not previously been flattened.

Then a pair of phrases is output: phrases in which all the words of one phrase correspond only to the words of another phrase and which do not correspond to words outside the limits of this phrase, we call them corresponding to each other. After receiving all the phrases, the probability of their translation can be calculated using the relative frequency:

$$\varphi(t|k) = count(t,k) / \Sigma t\ count(t,k)$$

It should be noted that a similar model based on expanding intersections within a smoothing set of mergers was used in [29].

Also of interest are the methods [30], which make it possible to deduce unnoticed phrase pairs from the legs encountered in word smoothing. This effect is achieved by taking into account the likelihood of lexical translation, the length of phrases, and other parameters.

You can also note the method that allows you to get a couple of phrases without preliminary smoothing [31]. According to the principle, here you can search for a correspondence between groups of words that go right next to each other.

Moreover, when developing statistical machine translation systems, it can be useful to combine phrasal tables obtained in different ways [30]. And some techniques allow you to expand the phrasal table by adding specially composed paraphrases [32].

Probabilistic model. According to Bayes' rule, the best ebest translation for the phrase f can be calculated as follows:

$$k_{best} = \text{argmax}_k\ p(k|t) = \text{argmax}_k\ p(t|k)\ p_{lm}(k)$$

Where $p(t|k)$ is the translation model and $p_{lm}(k)$ – is the language model.

In addition, the translation model can be represented as the following formula:

$$p(\bar{t}_1^I | \bar{k}_1^I) = \prod_{i=1}^{I} \phi(\bar{t}_i^I | \bar{k}_i^I)\ d(start_i - end_{i-1} - 1)$$

where φ – is the probability of transfer, and d – is the probability of bias.

Words-sentences leveling alignment in Kazakh, Turkish language spresented in the figure 2.

Fig. 2 Words-sentences leveling alignment in Kazakh, Turkish languages

The conditional probability of each word must be calculated. If the Kazakh word P(K) and the Turkish word P(T) are probabilities, then

$$P(K|T) = P(T|K)\, P(T)$$

First step

The system knows that ty and yi are connected: P (ty / yi) = 0.80

The next step

The system knows that the cat and kedi are related: P (cat | kedi) = 0.85

Third

The system knows that black and siyah are related: P (black / siyah) = 0.50

Conclusion

The main purpose of “Natural Language Processing (NLP)”, which belongs to the field of artificial intelligence in computer

science, is to automatically analyze, understand, interpret and create natural language texts used by people in communication.

For the first time in the course of research work, a Kazakh-Turkish parallel corpus was developed. As a result of this corpus, a phrase-based machine translation was made. The more parallel texts accumulate, the better the results of statistical machine translation will be achieved. In the future, it is planned to increase the number of these parallel corpuses and carry out work on the creation of a neuro-machine translation system.

REFERENCES

1. Zhetkenbay L., Bekmanova G., Yergesh B., Sharipbay A. (2020). Method of Sentiment Preservation in the Kazakh-Turkish Machine Translation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (pp.538-549). Vol. 12250.

2. Yelibayeva G., Sharipbay A., Mukanova A., Razakhova B. (2020). Applied ontology for the automatic classification of simple sentences of the Kazakh language. 5th International Conference on Computer Science and Engineering, UBMK 2020 (pp. 13-18). # 9219461

3. Yergesh B., Bekmanova G., Sharipbay A., Yergesh M.(2017). Ontology-based sentiment analysis of Kazakh sentences. O.Gersavi et al. (Eds.): ICCSA 2017 (pp. 669-677). Part III, LNCS. Vol.10406.

4. Dönmez İ., Adalı E. (2018). Turkish document classification with coarse-grained semantic matrix. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (pp. 472-484). Vol. 9624.

5. Çolakoğlu T., Sulubacak U., Tantug A.C. (2019). Normalizing Non-canonical Turkish texts using machine translation approaches. ACL 2019 - 57th Annual Meeting of the Association for

Computational Linguistics, Proceed. of the Student Research Workshop(pp. 267-272).

6. Yildirim E., Tantug A.C. (2013). The feasibility analysis of re-ranking for N-best lists on English-Turkish machine translation. 2013 IEEE International Symposium on Innovations in Intelligent Systems and Applications, IEEE INISTA 2013. #6577652.

7. Eryiğit G., Nivre J.,Oflazer K. (2008). Dependency parsing of Turkish. Computational Linguistics (pp.357–389).

8. Oflazer K.(1995). Two-level Description of Turkish Morphology. Literary and Linguistic Computing (pp. 137-148). Vol. 9.

9. GulilaAdongbieke, MijitAblimit. (2004). Research on Uyghur word segmentation. Journal of Chinese information processing (pp.61-65).Vol.18(6) (in Chinese).

10. Altıntaş K., Çiçekliİ. (2001). A Morphological Analyser for Crimean Tatar. Proceed. of the 10th Turkish Symposium on Artificial Intelligence and Neural Networks, TAINN (pp. 180-189). North Cyprus.

11. Tantuğ A.C., Adali E., Oflazer K. (2006). Computer analysis of the Turkmen Language morphology. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (pp. 186-193). Vol. 4139 LNAI.

12. Orhun, M., Tantuğ, A.C., Adali, E., Sönmez, A.C. (2009). Computational comparison of the Uyghur and Turkish grammar // Proceed. - 2009 2nd IEEE International Conference on Computer Science and Information Technology, ICCSIT 2009(pp. 343-347). #5234696.

13. Sharipbayev A., Bekmanova G., Mukanova A., Buribayeva A., Yergesh B., Kaliyev A. (2012). Semantic neural network model of morphological rules of the agglutinative languages. The 6th International Conference on Soft Computing and Intelligent Systems The 13th International Symposium on Advanced Intelligent Systems (pp.1094-1099). Kobe, Japan.

14. Yergesh B., Mukanova A., Bekmanova G., Sharipbay A., Razakhova B. (2014) Semantic hyper-graph based representation of Verbs in the Kazakh language. Computacion y Sistemas (pp.627-635). Vol. 18(3).

15. Zetkenbay L., Sharipbay A., Bekmanova G., Kamanur U.(2016). Ontological modeling of morphological rules for the adjectives in Kazakh and Turkish languages. Journal of Theoretical and Applied Information Technology (pp. 257-263). Vol. 91. #2. ISSN: 1992-8645, E-ISSN: 1817-3195.

16. Bekmanova G., Sharipbay A., Altnbek G., Adalı E., ZhetkenbayL., Kamanur U., Zulkhazhav A. (2017). A uniform morphological analyzer for the Kazakh and Turkish languages. Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST 2017) (pp. 20-30). Moscow, Russia.

17. Kessikbayeva G., Cicekli I. (2016). Rule Based Morphological Analyzer of Kazakh Language. Linguistics and Literature Studies (pp.96-104). Vol. 4(1).

18. Makhambetov O., MakazhanovA., Sabyrgaliyev I., Yessenbayev Z. (2015). Data-driven morphological analysis and disambiguation for Kazakh. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (pp. 151-163). Vol. 9041.

19. Toleu A.,Tolegen G., Makazhanov A. (2017). Character-Aware neural morphological disambiguation. ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (pp. 666-671).# 2.

20. https://translate.google.com/intl/en/about/languages/

21. http://ling.ulstu.ru/linguistics/resourses/literature/articles/corpus_education_translation/

22. Philipp Koehn. (2005)Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit.

23. https://www.isi.edu/natural-language/download/hansard/

24. https://catalog.ldc.upenn.edu/LDC2018T17

25. http://catalog.elra.info/en-us/repository/browse/ELRA-W0126/

25. MyrzakhmetovB.,MakazhanovA. (2017). Initial Experiments on Russian to Kazakh SMT. Research in Computing Science (pp. 153-160). Vol. 117.

26. Makazhanov A., Myrzakhmetov B., Assylbekov Z. (2019). Manual vs automatic bitext extraction. LREC 2018 - 11th International Conference on Language Resources and Evaluation (pp. 3834-3838).

27. Och F., Ney H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Journal Computational Linguistics (pp 19-51). Vol. 29 (1). MIT Press Cambridge, MA, USA.

28. Tillmann C. (2003). A Projection Extension Algorithm for Statistical Machine Translation. Proceed. of the 2003 conference on Empirical methods in natural language processing (pp.1-8).

29. Venugopal A., Vogel S., Waibel A. (2003). Effective Phrase Translation Extraction from Alignment Models. ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (pp. 319-326). Vol. 1.

30. Vogel S, Zhang Y., Huang F., Tribble A., Venugopal A., Zhao B., Waibel A. (2003). The CMU Statistical Machine Translation System. Proceed. of the MT Summit IX. New Orleans, LA.

31. Callison-Burch C., Koehn P., Osborne M. (2006). Improved Statistical Machine Translation Using Paraphrases. Proceed. of the Human Language Technology Conference of the North American Chapter of the ACL (pp.17–24). New York.

Khusainov A.F, Gatiatullin A.R.,
Suleimanov D.Sh, Gilmullin R.A.
Institute of Applied Semiotics of the AS of the RT,
Russia, Tatarstan, Kazan

TO THE CREATION OF A COMPLEX OF MACHINE TRANSLATION SYSTEMS BETWEEN RUSSIAN AND TURKISH LANGUAGES "TURKLANG-7"

Abstract. The idea of the "TurkLang-7" project is to create datasets and neural machine translation systems for a set of Russian-Turkic low-resource language pairs. It's planned to achieve this goals through an hybrid approach to creating a multilingual parallel corpus between Russian and Turkic languages, studying the applicability and effectiveness of neural network learning methods (transfer learning, multi-task learning, back-translation, dual learning) in the context of the selected language pairs, as well as the development of specialized methods for unification of parallel data in different languages, based on the agglutinative nature of the selected Turkic languages (structural and functional model of the Turkic morpheme). In this paper we describe the main stages of work on this project.

Keywords*: neural machine translation, multilingual datasets, Turkic languages.*

Хусаинов А.Ф., Гатиатуллин А.Р.,
Сулейманов Д.Ш., Гильмуллин Р.А.
Институт прикладной семиотики АН РТ,
Россия, Татарстан, Казань

К СОЗДАНИЮ КОМПЛЕКСА СИСТЕМ МАШИННОГО ПЕРЕВОДА МЕЖДУ РУССКИМ И ТЮРКСКИМИ ЯЗЫКАМИ «TURKLANG-7»

Аннотация. Идея проекта «TurkLang-7» заключается в создании наборов данных разработки системы нейронного машинного перевода для русско-тюркских языковых пар с

ограниченными ресурсами. Достичь этой цели планируется за счет гибридного подхода к созданию многоязычного параллельного корпуса между русским и тюркским языками, изучения применимости и эффективности методов обучения нейронных сетей (трансферное обучение, многозадачное обучение, обратный перевод, двойное обучение) в контексте выбранных языковых пар, а также разработки специализированных методов унификации параллельных данных на разных языках, основанная на агглютинативной природе выбранных тюркских языков (структурно-функциональная модель тюркской морфемы). В этой статье мы опишем основные этапы работы над этим проектом.

Ключевые слова: *нейронный машинный перевод, многоязычные наборы данных, тюркские языки.*

1. Введение

Область построения систем автоматического машинного перевода получила стремительное развитие в последние годы, во многом благодаря успешному использованию современных методов машинного обучения. Однако методы нейросетевого машинного перевода, позволяющие достичь наилучших результатов для крупнейших пар мировых языков (англо-немецкой, англо-китайской и других), невозможно напрямую использовать в случае недостатка обучающих данных.

Особую актуальность имеет ряд подзадач, связанных с адаптацией и доработкой существующих подходов для случаев малоресурсных языков. В этой области достигнуты определенные успехи, в том числе в рамках технологий переноса знаний (transferlearning, zero-shotlearning) и искусственного увеличения объема обучающих данных (например, back-translation, duallearning). Однако для языков тюркской группы, относящихся к классу малоресурсных языков, исследование возможности построения многоязычной системы машинного перевода с русского и на русский язык, не проводилось.

Данный проект направлен на разработку методов и программных средств для целого ряда языковых пар, в которых

один язык является русским, а второй принадлежит к тюркской языковой группе. Благодаря решению поставленных в рамках проекта задач: накоплению параллельных обучающих корпусов данных, разработке метода унификации собранных параллельных корпусов на основе структурно-функциональной модели тюркских морфем, а также программных средств обучения многоязычного машинного переводчика на основе подходов переноса знаний и искусственного увеличения объема данных — планируется преодолеть проблему недостатка обучающих данных. Это должно позволить впервые создать систему перевода для крымско-татарско-русской языковой пары, кроме того, итоговая система машинного перевода также будет работать с ещё 6 языковыми парами (татарско-русской, башкирско-русской, чувашско-русской, казахско-русской, киргизско-русской и узбекско-русской). Общее количество носителей языков, для которых предполагается проведение данного исследования, составляет 57,93 млн человек [1], проживающих преимущественно на территории Российской Федерации и стран СНГ.

Результаты исследования позволят представить информацию о степени влияния множества параметров (объёма использованных корпусов родственных языков, искусственно сгенерированных параллельных данных, применение различных методик обучения и выбора архитектур нейросети) на качество работы итоговой системы машинного перевода.

2. Обзор

Подходы, применяемые к разработке систем машинного перевода, претерпевают серьезные изменения в последние годы. Значительные усилия направлены в равной степени на решение задач машинного перевода для случаев крупнейших мировых языков и малоресурсных языков. Количество и качество данных, доступных для выбранных языков, определяют набор алгоритмов и подходов к переводу.

Задача машинного перевода решается с помощью так называемых sequence-to-sequence моделей [2], построенных, например, на рекуррентных, сверточных нейросетях,

включающих элементы кодировщика и декодировщика (encoder/decoder архитектура). Модели, показывающие наилучшее качество работы, также включают механизм внимания (attention, self-attention).

Существуют различные архитектуры нейросетей, созданные с целью ускорить процесс обучения и улучшить качество работы переводчиков: рекуррентные нейросети [3], сверточные нейросети [4], модели Transformer [5] и EvolvedTransformer. Механизм внимания тоже был доработан: были предложены варианты multi-hopattention, self-attention и multi-headattention [6, 7, 8].

Выбор технологии построения системы машинного перевода очень сильно зависит от наличия и объема исходных обучающих данных. Наличие больших моно-корпусов для исходного и целевого языков позволяет использовать unsupervised-подходы к построению машинных переводчиков. Основная идея данного подхода заключается в построении единого векторного пространства слов/фраз для обоих языков. На текущий момент существуют варианты реализации данного подхода на основе статистического [9], нейросетевого [10] и гибридного подходов [11].

Были также предложены различные варианты использования моно-корпусов для улучшения качества перевода при обучении с частичным привлечением учителя (semi-supervised подход) [12].

Еще один способ использовать моноязычные данные – дополнить языковой моделью часть системы, отвечающую за декодировщик (decoder) [13]. Данный подход был использован ещё в самых ранних работах IBM [14]. Позже было показано, что дополнительная языковая модель для целевого языка перевода позволяет системам, построенным на статистическом подходе, улучшить естественность и корректность перевода [15]. Похожая стратегия в дальнейшем была также применена к нейросетевым системам машинного перевода [16]. Помимо использования во время декодирования, нейросетевые языковые и переводческие модели могут быть успешно интегрированы на внутреннем уровне за счет совмещения скрытых состояний

моделей [17]. Кроме того, нейросетевая архитектура позволяет использовать многозадачное обучение (multi-tasklearning) и совместное обучение параметров (parametersharing) [18].

И, наконец, в [19] предложили добавить вспомогательную задачу авто-кодирования для моноязычных данных, которая обеспечивает получение исходного приложения в результате последовательного перевода предложения в обе стороны.

В работе [20] авторы показали, что качество перевода в случае малоресурсных языковых пар может быть улучшено за счет синтетических данных, где предложения на исходном языке создаются простой копией предложений целевого языка.

Подход, предложенный в работе [21], предполагает очень эффективный способ автоматического увеличения данных для обучения (data-augmentation). Способ получил название back-translation (BT): сначала на доступных параллельных данных обучается вспомогательная система перевода с целевого языка на исходный, а затем эта система используется для перевода моноязычного корпуса целевого языка, тем самым увеличивая объем параллельного корпуса. Результирующий параллельный корпус используется в качестве обучающих данных для системы машинного перевода.

BT является простым для использования, так как не требует внесения изменения в обучающие алгоритмы машинного переводчика. Помимо основной задачи увеличения объема обучающих данных для малоресурсных пар языков, он также может быть использован для использования моноязыкового корпуса для задачи адаптации переводчика к конкретной предметной области [22]. В качестве последних идей по улучшению BT был предложено отказаться от генерации синтетических пар с помощью поиска по лучу (beamsearch) [23] или жадного поиска (greedysearch) [24]. Оба перечисленных алгоритма позволяют искать апостериорный максимум (MAP), то есть находить предложение с максимальной вероятностью согласно модели. Однако использование MAP может привести к менее разнообразному подкорпусу переводов, так как в случаях многозначности алгоритм всегда будет выбирать наиболее вероятный вариант [25]. В качестве альтернативы

рекомендуется использовать метод случайного выбора (randomsampling) [26]. Это позволяет сохранять лексическое разнообразие сгенерированных пар предложений. При этом можно внедрить дополнительные правила, позволяющие исключить очень редкие варианты переводов [27]. Кроме того, возможна генерация нескольких вариантов исходных предложений для каждого предложения. Важнейшее изменение к подходу было предложено в [28]: авторами был представлен итеративный процесс обучения/добавления синтетической части обучающего корпуса для улучшения качества финальных систем.

В работах [29] было показано, каким образом может быть улучшено качество переводчиков, если для обоих языков есть моноязычные корпуса. Использование одновременно двух корпусов позволяет перейти от BT к так называемому двойному обучению (duallearning): обучение происходит одновременно в обоих направлениях перевода, BT итеративно используется в обоих направлениях для постепенного увеличения размера обучающего корпуса и доли синтетических предложений.

Отдельного упоминания заслуживает подход Byte-pairencoding (BPE) [30], применяемый к задаче машинного перевода на основе базовых элементов, меньше целого слова (subword MT). Использование сегментации на основе BPE [31] позволяет, в том числе, решать задачу перевода с открытым словарем (система способна переводить любые слова, в том числе, те, которые не присутствуют в обучающих корпусах). BPE изначально был создан как алгоритм сжатия, но был адаптирован для сегментации слов следующим образом: каждое слово из обучающего словаря представляется последовательностью символов, завершающейся специальным символом конца слова; все символы добавляются в словарь элементов; определяются самые частотные пары символов — обнаруженные последовательности добавляются в словарь элементов и объединяются в корпус. Процедура повторяется до достижения заданного количество операций объединения.

3. План реализации проекта и промежуточные результаты

Задачу накопления параллельных обучающих данных для русского и группы тюркских языков планируется решать с помощью совокупности методов, включающих пополнение итогового корпуса из двуязычных Интернет-источников (новостные порталы, электронные библиотеки со свободной лицензией и другие), оцифровку печатных версий книг, имеющих перевод на один из выбранных в проекте языков, объединение уже созданных источников параллельных данных. Процесс решения данной задачи включает как экспертную работу по установлению источников данных, так и разработку необходимых методов для анализа, включающих реализацию алгоритмов выравнивания по документам, выравнивания по предложениям, фильтрация на основе набора эвристических правил, а также внедрение интеллектуальных моделей фильтрации пар параллельных предложений на основе имеющихся (для некоторых языков) подкорпусов хорошего качества.

Ключевая задача построения модели машинного перевода будет решаться на основе нейросетевого подхода. На начальном этапе планируется использование архитектуры нейросети Transformer, ключевой особенностью которой является использование механизма внимания (multi-headattention) и отсутствие сверточных и рекуррентных слоев.

Успешное использование заявленного подхода к обучению было невозможно для большинства из заявленных в проекте языковых пар (за исключением казахско-русской и татарско-русской пар). Комплексный подход к построению многоязычного переводчика должен позволить решить эту проблему за счет:

- использования различных подходов переноса знаний с одной языковой пары на другую (например, fine-tuning нейросети, предобученной на более обеспеченной ресурсами языковой паре/парах; внедрение токена для представления исходного языка на уровень embedding-слоя нейросети и обучение единой многоязычной нейросети);

- разработки общего представления частей слов для всех заявленных языков (например, общих byte-pairencoding-элеменов для всех языков);
- применения методов искусственного увеличения объема обучающих данных back-translation (использование промежуточных версий переводчика для наращивания объема параллельных данных на основе перевода моноязычных текстовых корпусов) и его модификаций, использующих метод randomsampling вместо beamsearch для получения переводов с более богатой лексикой;
- разработки методов унификации собранных параллельных корпусов для различных языковых пар, которая позволит более полно использовать корпуса большего размера, имеющиеся для одних языковых пар, при обучении модели переводчика в других парах. Предлагается использовать структурно-функциональную модель тюркской морфемы [32] с целью включения в обучающие данные информации о взаимосвязи грамматических ролей, выполняемых аффиксами в различных тюркских языках. Данная информация позволит на начальном этапе подготовки корпуса сформировать единые элементы для морфем в разных языках.

Таким образом, ключевые этапы работ, касающиеся подготовки обучающих данных и непосредственно обучения модели переводчика, опираются одновременно на методы машинного обучения и подход, основанный на правилах (базу данных для языков в структурно-функциональной модели тюркской морфемы).

На заключительном этапе будет проведено тестирование различных вариантов систем машинного перевода с целью выявления степени влияния использованных подходов, настроек, значений гиперпараметров нейросети и объемов использованных обучающих данных.

4. Заключение

В этой статье мы представили описание стартовавшего проекта “TurkLang-7”, направленного на создание комплекса систем машинного перевода для 7 пар языков. Ключевой

особенностью его реализации будет совмещение современных технологий машинного обучения с rule-based подходами, основанными на особенностях тюркских языков.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта№ 20-07-00823.

ЛИТЕРАТУРА

1. Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2020. Ethnologue: Languages of the World. Twenty-third edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

2. Sutskever, I. Sequence to sequence learning with neural networks [Text] / Ilya Sutskever, Oriol Vinyals, and Quoc V.Le. // Advances in Neural Information Processing Systems. – 2014. – P.3104–3112.

3. Bahdanau, D., Cho, K., Bengio, Y. Neural machine translation by jointly learning to align and translate [Text] / DzmitryBahdanau, Kyunghyun Cho, and YoshuaBengio // International Conference on Learning Representations (ICLR). – 2015.

4. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, N.Y. Convolutional sequence to sequence learning [Text] / Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin // International Conference of Machine Learning (ICML). – 2017.

5. Vaswani, A. Attention is all you need [Text] / Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and IlliaPolosukhin // Conference on Advances in Neural Information Processing Systems (NIPS). – 2017.

6. Gehring, J., Auli, M., Grangier, D., Yarats, D., Yann N Dauphin. Convolutional sequence to sequence learning [Text] / Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin // International Conference of Machine Learning (ICML). – 2017.

7. Paulus, R., Xiong, C., Socher, R. A deep reinforced model for abstractive summarization [Text] / Romain Paulus, CaimingXiong, and Richard Socher // International Conference on Learning Representations (ICLR). – 2018.

8. Johnson, M. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation [Text] / Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viegas, Martin Wattenberg, Gregory S. Corrado, Macduff Hughes, Jeffrey Dean // Transactions of the Association for Computational Linguistics, Vol. 5, – P.339-351. –2017.

9. Artetxe, M. Unsupervised Statistical Machine Translation [Text] / Mikel Artetxe, GorkaLabaka, EnekoAgirre // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. – P.3632–3642. – 2018.

10. Artetxe, M. Unsupervised Neural Machine Translation [Text] / Mikel Artetxe, GorkaLabaka, EnekoAgirre, Kyunghyun Cho // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. – P.3632–3640. – 2018.

11. lample, G. Phrase-Based & Neural Unsupervised Machine Translation [Text] / Guillaume Lample, Myle Ott, Alexis Conneau, LudovicDenoyer, Marc'AurelioRanzato // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. – 2018.

12. Munteanu, D.S. Improved machine translation performance via parallel sentence extraction from comparable corpora [Text] / D.S. Munteanu, A. Fraser, and D. Marcu // ACL, – 2004.

13. Caglar, G. On using monolingual corpora in neural machine translation [Text] / CaglarGulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, LoicBarrault, Huei-Chi Lin, FethiBougares, Holger Schwenk, and YoshuaBengio // arXiv preprint arXiv:1503.03535. – 2015.

14. Brown, P.F. A statistical approach to machine translation [Text] / Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin // Computational Linguistics, 16:79–85. – 1990.

15. Koehn, Ph. Statistical phrase-based translation [Text] / Philipp Koehn, Franz Josef Och, and Daniel Marcu // Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). – 2003.

16. He, W. Improved neural machine translation with smt features [Text] / Wei He, Zhongjun He, Hua Wu, and Haifeng Wang

// Conference of the Association for the Advancement of Artificial Intelligence (AAAI), – P.151–157. – 2016.

17. Caglar, G. On integrating a language model into neural machine translation [Text] / CaglarGulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and YoshuaBengio // Computer Speech & Language, 45. – P.137–148. – 2017.

18. Domhan, T. Using targetside monolingual data for neural machine translation through multi-task learning [Text] / Tobias Domhan, Felix Hieber // Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2017.

19. Cheng, Y. Semi-supervised learning for neural machine translation [Text] / Y. Cheng, W. Xu, Z. He, W. He, H. Wu, M. Sun, and Y. Liu // arXiv:1606.04596. – 2016.

20. Currey, A. Copied Monolingual Data Improves Low-Resource Neural Machine Translation [Text] / Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield // Proc. of WMT. – 2017.

21. Sennrich, R. Improving neural machine translation models with monolingual data [Text] / Rico Sennrich, Barry Haddow, and Alexandra Birch // arXiv preprint arXiv:1511.06709, – 2015.

22. Bertoldi, N., Federico, M. Domain adaptation for statistical machine translation with monolingual resources [Text] / Nicola Bertoldi and Marcello Federico // Workshop on Statistical Machine Translation (WMT). – 2009.

23. Sennrich, R. Improving neural machine translation models with monolingual data [Text] / Rico Sennrich, Barry Haddow, and Alexandra Birch // Conference of the Association for Computational Linguistics (ACL). – 2016.

24. Lample, G.. Unsupervised machine translation using monolingual corpora only [Text] /Guillaume Lample, Alexis Conneau, LudovicDenoyer, and Marc'AurelioRanzato // International Conference on Learning Representations (ICLR). – 2018.

25. Ott, M. Analyzing uncertainty in neural machine translation [Text] / Myle Ott, Michael Auli, David Grangier, and Marc'AurelioRanzato. // Proceedings of the 35th International Conference on Machine Learning, vol. 80, – P.3956–3965. – 2018.

26. Imamura, K. Enhancement of encoder and attention using target monolingual corpora in neural machine translation [Text] /

Kenji Imamura, Atsushi Fujita, and EiichiroSumita // Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, – P.55– 63. – 2018.

27. Graves, A. Generating sequences with recurrent neural networks [Text] / Alex Graves // arXiv, 1308.0850. – 2013.

28. Vu Cong Duy Hoang, Philipp Koehn, GholamrezaHaffari, Trevor Cohn. Iterative backtranslation for neural machine translation [Text] / Vu Cong Duy Hoang, Philipp Koehn, GholamrezaHaffari, and Trevor Cohn // Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, – P.18–24. – 2018.

29. Cheng, Y. Semisupervised learning for neural machine translation [Text] / Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu // Conference of the Association for Computational Linguistics (ACL). – 2016.

30. Gade, F. A New Algorithm for Data Compression [Text] / Philip Gage // C Users J., 12(2):23–38, February. – 1994.

31. Sennrich, R. Neural Machine Translation of Rare Words with Subword Units [Text] / Rico Sennrich, Barry Haddow, and Alexandra Birch // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). – Berlin, Germany. – 2016.]

32. Гатиатулин, А. Р. Многофункциональный Интернет сервис как инструмент для формирования и использования лексикографической базы тюркских языков / А. Р. Гатиатуллин // Сохранение языков народов мира и развитие языкового разнообразия в киберпространстве: контекст, политика, практика: матер. Междунар. конф. (Якутск, 1-5 июля 2019 г.). – Якутск, 2019.

СЕКЦИЯ 3
СИСТЕМЫ МОРФОЛОГИЧЕСКОЙ И СИНТАКСИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

Abjalova M. A.
Tashkent State University of Uzbek Language and Literature named after Alisher Navoi, Uzbekistan, Tashkent

AUTOMATIC ANALYSIS OF PHRASEOLOGICAL UNITS IN LINGUISTIC PROGRAMS

Abstract. Phraseological units are a linguistic phenomenon that requires separate study and plays an important role in the linguistic support of text-based programs. Despite the fact that the presence of a syntactic connection between the components of a phraseological unit does not affect its semantic integrity, this does not mean that the syntactic origin is always appropriate, although in some cases, the syntactic connection with a phrase is unambiguously free. Therefore, when creating linguistic software for programs for working with texts, it is advisable to consider word combinations as a single form, and the place of syntactic formation, as an indicator of the relationship form, should be strictly defined.

This article examines the place of phraseological units and their functions in the database of linguistic programs.

Keywords: *linguistic analysis, linguistic support, syntactic analysis, phraseological unity, phraseme, somatic phraseology, syntactic derivation.*

Абжалова М. А.
Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои, Узбекистан, Ташкент

АВТОМАТИЧЕСКИЙ АНАЛИЗ ФРАЗЕОЛОГИЧЕСКИХ ЕДИНИЦ В ЛИНГВИСТИЧЕСКИХ ПРОГРАММАХ

Аннотация. Фразеологические единицы – лингвистическое явление, которое требует отдельного изучения и играет важную

роль в лингвистическом обеспечении программ, основанных на текстах. Несмотря на то, что наличие синтаксической связи между компонентами фразеологизма не влияет на ее семантическую целостность, это не означает, что синтаксическое происхождение всегда уместно, хотя в некоторых случаях, синтаксическая связь с фразой однозначно свободна. Поэтому при создании лингвистического обеспечения программ для работы с текстами словосочетания целесообразно рассматривать как единую форму, а место синтаксического образования, как показатель формы отношения, должно быть строго определено.

В рамках данной статьи рассматривается место фразеологизмов и их функции в базе лингвистических программ.

Ключевые слова: *лингвистический анализ, лингвистическое обеспечение, синтаксический анализ, фразеологическое единство, фразема, соматическая фразеология, синтаксическая деривация.*

Автоматическое редактирование и анализ текстов на любом языке требует определенной лингвистической поддержки в памяти компьютера. Основа предложения – лингвистические правила и нормы. Словарный запас также включает богатство лингвистической и филологической лексики конкретного языка. Следовательно, при создании лингвистического процессора необходимо иметь достаточные знания об обрабатываемом естественном языке. Это, в свою очередь, гарантирует безупречный лингвистический процессор.

При создании лингвистического процессора для программы автоматического редактирования и анализа узбекских текстов лингвисту потребуется набор лингвистических словарей и грамматических правил узбекского литературного языка.

Фразеологизмы, по сути, являются плодом речи и художественного дискурса. Все уровни языка участвуют в создании и формировании языка. Однако, следует отметить, что язык, особенно его разнообразие лексической составной, т.е. лексический (фразеологический) уровень играет особую роль в создании и отражении языкового явления [1:22]. Например,

появление синонимичной серии из около сотни фразеологизмов (эвфемизмов), обозначающих значение одной лексемы *ўлмоқ* (смерт), создает возможность их функционального ограничения. Например:*“оламдан ўтмоқ”, “дунёдан ўтмоқ” (уйти из жизни), “дунёни тарк этмоқ” (покинуть мир), “омонатини топширмоқ” (исчезнуть с лица земли), “қулоғи остида қолмоқ” (проститься с жизнью), “жон бермоқ”(уйти в небытие), “у дунёга кетмоқ” (отойти на тот свет /отойти в мир иной)* и др. употребляются в разговорном стиле, а фраземы как *“вафот этмоқ”, “ҳаётдан кўз юммоқ”, “дунёдан кўз юммоқ”, ҳаёт билан видолашмоқ* могут встречаться публицистическом и формальном стилях. *“Аллоҳ раҳматига йўл тутмоқ”* (*Следовать пути по милости Аллаха), “шаҳодат шаробини ичмоқ” (пить вино мученичества) и “дорилфанодан дорилбақога рихлат қилмоқ”*(*переходить с “дорилфано к дорилбако”)* – относятся к литературному тексту. Очевидно, что учебный словарь фразеологизмов с учетом определенных методических особенностей является одним из источников, обеспечивающих совершенствование программы [2, 3].

Фразеологические единицы (ФЕ) комбинируются по смыслу, образуя семантическую целостность. Часто значение эквивалентно слову. Особый раздел лингвистики, изучающий фразелогические единицы, называется фразеологией.

В узбекском языке более тысячи ФЕ. В *“*Толковом словаре узбекского языка*”* встречается более восьмидесяти однокомпонентных фразеологизмов, а в *“*Фразеологическом словаре узбекского языка*”* Ш. Рахматуллаева – 158. В исследовании А. Исаева *“*Соматическая фразеология в узбекском языке*”* сообщается, что в нашем языке встречаются 127 фразелогических единиц с глазным компонентом [4:52].

Важно, чтобы фразеологизмы служили для выражения бесконечных и разнообразных эмоций, физиологических процессов человека в привлекательной художественной форме. Значительная часть фразеологизмов узбекского языка структурно представляются в виде: существительное + глагол, существительное + прилагательное, существительное + числительное + глагол, существительное + прилагательное,

существительное + числительное. Например: *кўз югуртирмоқ, кўзига иссиқ кўринмоқ, кўзига яқин кўринмоқ, кўзи ола-кула бўлмоқ, кўзи тўрт бўлмоқ, кўзи очиқ кетмоқ, кўзи очиқ, кўзи бежо, кўзи така-пука (закатывать глаза, выглядеть знакомо в его глазах, глаза искрятся, двоится в глазах, широко раскрытыми глазами, безжизненный взгляд, глаза от радости блестят).*

Все эти ФЕ принадлежат одному из трех индивидуумов, то есть говорящему, слушателю или другому человеку, поскольку они служат для выражения определенного внутреннего состояния индивидуума. Следовательно, слово, принадлежащее к категории существительных в ФЕ, принадлежащее к именному словосочетанию, или слово, принадлежащее к категории глаголов, приобретает грамматические особенности. Например: *"Қисқаси, энди унга хат ёзишга* ***қўлимбормади****"*(*Проще говоря, отныне у меня не поднимается рука писать ему письма)* (А. Мухтор). *"Қўлим бормади"* (У меня ***не поднимается рука /***У него ***не поднимается рука)***: существительное + глагол форма ФЕ, которая формирует отношение: моя "*қўли****м"*** (*рук****а)***(-а аффикс первого лица единственного числа категории собственности *-т* =>$N_{e.a.I.b}$ + V×). "*Бу саволдан унинг кайфи тарқалгандек бўлиб,* ***кўзлари***" (Его настроение, казалось, испортилось от этого вопроса, и его глаза) ("-лар" (*-а*) аффикс множественного числа, "*-и*" (*-а*) притяжательный аффикс III лица=>$N_{e.a.III.л.}$) "***мошдек очилди***" (широко раскрылись) (*-ыл* отношение идентичности, -ись 3 лицо множественное число =>Avd-$V_{pass.v.}$ + III.b.) (Ойбек. Детство), «О, братец Мухиддин, твое сердце тоже сожжено, добро пожаловать в ад», – сказал Мухиддин (Р. Файзи. В пустыню пришла весна). Грамматическое образование ФЕ "сердце сожжено": твое сердце *(-е* аффикс второго лица множественного числа притяжательной формы =>$N_{e.a.II.b.}$) тоже – частица, сожжено – лемма + прошедшее время (*-но*).

Фразеологизм – это лексическая единица, подобная слову. Он составляет словарный запас языка. Фразеологизм не образуется в процессе речи, как фраза или предложение, он воспроизводится как готовая речевая единица. Следовательно,

ФЕ – это языковой феномен, а не речевой. – Однако парадигматические и синтагматические отношения в ФЕ показывают, что ФЕ образуют отдельную область в лингвистической поддержке обработки текста, автоматического редактирования и анализа текста, программного обеспечения машинного перевода.

Вариация структурного выражения речи			
Единственное число	**модель**	**Множественное число**	**модель**
Я на седьмом небе от счастья	$N_{e.a.Ib.}+N_{k.a}+V_{Ib}$.“Бошим осмонга етди”	Мы на седьмом небе от счастья	$N_{e.a.Ik.}+N_{k.a}+V_{Ib.}$ “Бошимиз осмонга етди”
Ты на седьмом небе от счастья?	$N_{e.a.IIb.}+N_{k.a}+V_{Ib.}$ “Бошинг осмонга етдими?”	Вы на седьмом небе от счастья?	$N_{e.a.IIk.}+N_{k.a}+V_{Ib.}$ “Бошингиз осмонга етдими?”
Он на седьмом небе от счастья	$N_{e.a.IIIb.}+N_{k.a}+V_{Ib.}$ “Боши осмонга етди”	Они на седьмом небе от счастья	$N_{e.a.IIIk.}+N_{k.a}+V_{Ib.}$ “Боши осмонга етди”

Поскольку часть фразеологизма имеет жесткий шаблон, ее нельзя дословно перевести из одного языка на другой. Дословный перевод приводит к путанице. Поэтому при создании программного обеспечения автоматического перевода, мобильных приложений и лингвистической поддержки платформ словосочетания и структурные фразы предложений вводятся в базу данных в виде лексических единиц, а их перевод предоставляется на другие языки.

Несмотря на то, что фразеологизмы являются устойчивыми выражениями, синтаксические производные между их

компонентами динамичны. Эта ситуация затрагивает только формальную сторону фразеологизмов, а ее семантика сохраняет целостность и образность значения. Синтаксические отношения компонентов составного словосочетания, в свою очередь, являются первичными при создании фразеологического соединения в конкретное синтаксическое отношение. Например: "*сабр косаси тўлмоқ*" (чаша терпения уже наполнилась) – фразеологизм с предикативным знаком **N (N_o+$N_{э.а.}$=>именное словосочетание) + V:**"*Менинг сабр косам тўлди"* (Моя чаша терпения уже полна)*; "Сабр косанг тўлгандир"* (Наверное, твоя чаша терпения уже полна)*; "Унинг сабр косаси ҳам тўлди" (*Его чаша терпения также была уже полна*); "Сабр косамиз тўлиб бормоқда"* (Наша чаша терпения уже наполняется).

В целом, парадигматическое и синтагматическое формирование ФЕ представляют собой сложный структурный процесс социальной значимости, особенности которого более ярко выражены в тексте и в речевой ситуации. Однако, целесообразно воспринимать словосочетания как структурную единицу речи, как готовую лексическую единицу и определять место ее синтаксического соотношения.

ЛИТЕРАТУРА

1. Маҳмудов Н. Тилнинг сўз хазинаси ва оламнинг лисоний манзараси // Сўз санъати. International Journal of Word Art. 2018, vol. 1, issue 1, – PP. 22
2. Менглиев Б. ва бошқ. Ўзбек тили ибораларининг ўқув изоҳли луғати. – Тошкент: Янги аср авлоди, 2007.
3. Раҳматуллаев Ш. Ўзбек тилининг изоҳли фразеологик луғати. –Тошкент: Ўқитувчи, 2001.
4. Рашидова У. Семантико-прагматический анализ соматических выражений узбекского языка (на примере выражений с глазным, ручным и сердечным компонентами): Автореф. дисс. ф.ф.н. (PhD). – Самарканд. – 52 с.

Ayupov M.
Tatarstan Academy of Sciences, Kazan Federal University,
Russia, Tatarstan, Kazan

AUTOMATIC FILLING IN THE DATABASE OF THE PORTAL OF THE TURKIC MORPHEME BY MEANS OF PROCESSING BILINGUAL DICTIONARIES

Annotation. The article describes the automatic creation of a multilingual dictionary of root words of the Turkic languages using data of bilingual dictionaries. The resulting multilingual dictionary further will be imported into the database of the Turkic Morpheme Portal and will serve as its main table.

Keywords: *data base, dictionaries, software tools, Turkic language.*

Аюпов М.
Академия наук РТ, Казанский федеральный университет,
Россия, Татарстан, Казань

АВТОМАТИЧЕСКОЕ ЗАПОЛНЕНИЕ БД ПОРТАЛА ТЮРКСКОЙ МОРФЕМЫ С ПОМОЩЬЮ ПРОГРАММНОЙ ОБРАБОТКИ ДВУЯЗЫЧНЫХ СЛОВАРЕЙ

Аннотация. В статье описывается автоматическое создание многоязычного словаря основ тюркских языков с помощью обработки двуязычных словарей. Полученный многоязычный словарь в дальнейшем импортируется в базу данных портала тюркской морфемы и будет служить ее основной таблицей.

Ключевые слова: *база данных, словари, программные инструменты, тюркские языки.*

1. Введение

При создании больших многоязычных баз данных (БД) важная роль принадлежит заполнению этих баз данных, так как готовых решений не существует и подготовка нужных данных

занимает много времени. В рамках реализации базы данных [1] портала тюркской морфемы [2] применяются разные методы сбора данных. Одним из них является автоматическая обработка двуязычных словарей.

Для дальнейшей работы в БД портала тюркской морфемы были выбраны 7 тюркских языков: татарский, киргизский, узбекский, крымско-татарский, казахский, чувашский, башкирский. Решение задачи автоматического заполнения данными БД для вышеперечисленных языков требует большого количества лингвистических ресурсов. Анализ имеющегося на сегодня материала, готового для использования в данном проекте, показал, что практически отсутствуют двуязычные словари тюркских языков. Поэтому для создания связей между данными разных тюркских языков, было принято решение об использовании одного промежуточного языка. Этим языком был выбран русский язык, так как наиболее распространенными являются двуязычные словари, одним из языков в которых является русский язык. Сбор данных для тюркской базы данных сводится к следующему:

– с помощью программной обработки словарей находим переводы русского слова на разные тюркские языки,

– пытаемся автоматически найти связи между этими переводами,

– вручную обрабатываем спорные моменты.

2. Программная обработка двуязычных словарей

В ходе реализации проекта портала тюркской морфемы в предыдущие годы был вручную создан небольшой экспериментальный словарь основ тюркских языков. Для увеличения размера словаря найденные электронные версии двуязычных словарей были автоматически обработаны и объединены с помощью промежуточного словаря русского языка (рис. 1).

Tatar	Kirgiz	Uzbek	Cirim tatar
абайлылык, ихтираз,	аярдык, кыраакылык,	tuyg'unlik	ачыкъкозьлик, ихтият,
абайлы, сагаюлы, сак,	абайлагыч, кыраакы		ачыкъкозь, мукъайт,
йөгер, чап	жүгүр	chop, yugur	чап, ювур
бегемот, гиппопотам			
качкын	качкын, качуучу	qochkin	къачакъ, къачкъын
җәһәтлек, йөгереклек,		ildamlik, tezlik	сурьат, тезлик
качкан	качкы		
алгасак, алгыр, алгысак,	амалдуу, букта,	chopkir, ildam,	сурьатлы, тез, чабик,
бегония			бал-къаймакъ
йөгерешче, йөгерүче		yugurdak,	
афәт, бәла, бәла-каза,	алат, апат,	balo, baxtsizlik,	бахытсызлыкъ, беля,
болама, буталчыклык,	бүдүүчаң, иретсиздик,	sistemasizlik	къарышыкълыкъ,
банкротлан, бөл,	жакырдан, жардылан,		сюмеле, факъырлаш,
йолкышлык, мохрүмлек,	бакырдык, бакырлык,	faqirlik, muhtojlik	ёкъсуллыкъ, фукъарелик
факыйрь-фөкарә, ярлы-			
барлыксыз, малсыз,	жакыр, жарды, кедей,	faqir	джарлы, ёкъсул, зюгюрт,
бахетсез, кайгы-	бактысыз, бакытсыз,		бедбахт, гъарип,
бахыр, бичара,	бечара		байгъуш, бичаре
гидай	бакыр, кембагал		
бот, сан	бут, сан	son	

Рис. 1. Часть базы данных

Несмотря на то, что эти электронные версии словарей имели малый объем и не являлись полными, удалось добиться некоторого прироста. В итоге обновленный словарь основ тюркских языков имел следующие характеристики (табл. 1):

Таблица 1.

Статистика БД

Язык	Количество слов
татарский	33066
казахский	18757
башкирский	2977
крымско-татарский	7065
киргизский	9686
узбекский	5433

Наиболее полные двуязычные словари в основном были созданы и изданы в прошлом веке. У этих словарей отсутствует электронная версия, а если она и существует, то в большинстве случаев, как говорилось ранее, электронная версия является неполной и содержит малое количество слов. Поэтому для увеличения объема словаря основ тюркских языков возникла необходимость работы с бумажными двуязычными словарями. Далее рассмотрим этапы этой работы.

На первом шаге необходимо было найти отсканированную версию нужного словаря или, если нет версии с хорошим качеством, отсканировать словарь.

На втором шаге двуязычный словарь распознается с помощью специальных программ распознавания (рис. 2).

АККУРА`ТНО нареч. ұқыпты,мұқият,жинақы; ~ отвечать на письма хатқа мұқият жауап беру; ~ переписатьмұқият көшіру; 2. разг. абайлап,байқап,жайлап,білдірмей,ақырын; ~неситеабайлап апарыңдар; ~узнайжайлап біл; 3. разг. жүйелі,үнемі,үздіксіз; ~ навещать больногосырқатқа үнемі барып тұру.

Рис. 2. Пример словарной статьи

На третьем шаге с помощью программной обработки удаляется ненужная в дальнейшем информация, например, переводы примеров. То есть после третьего шага правая часть словаря превращается в словник, при этом информация о разных значениях сохраняется (рис. 3).

АККУРАТНО ұқыпты,мұқият,жинақы 2. абайлап,байқап,жайлап,білдірмей,ақырын 3. жүйелі,үнемі,үздіксіз.

Рис. 3. Словарная статья после удаления ненужной в дальнейшем информации

И, наконец, на последнем шаге полученный словник готовится для загрузки в базу данных:

– данные преобразовываются в табличный вид,

– информация о разных значениях удаляется и выводится в отдельный столбец в виде чисел,

– синонимичные переводы разбиваются на разные строки для получения окончательного вида таблицы, где строка имеет вид: одно слово – один перевод (рис. 4).

АККУРАТНО	ұқыпты	1
АККУРАТНО	мұқият	1
АККУРАТНО	жинақы	1
АККУРАТНО	абайлап	2
АККУРАТНО	байқап	2
АККУРАТНО	жайлап	2
АККУРАТНО	білдірмей	2
АККУРАТНО	ақырын	2
АККУРАТНО	жүйелі	3
АККУРАТНО	үнемі	3
АККУРАТНО	үздіксіз	3

Рис. 4. Данные для загрузки в БД

Для примера возьмем казахский язык. В предварительной базе данных для казахского языка имелось 18757 заполненных строк. Отсканированный pdf файл русско-казахского словаря содержал около 56000 словарных статей. Готовый к загрузке в БД словник содержит более 117000 строк.

3. Проблемы, возникающие во время обработки двуязычных словарей

Во время обработки двуязычных словарей возникает множество проблем. Рассмотрим некоторые из них.

При распознавании отсканированных двуязычных словарей с помощью специальных программ распознавания возникают орфографические ошибки, если встречаются неуверенно распознанные слова и слова, отсутствующие в словаре программы распознавания. Эти ошибки автоматически исправлять не получается, их выявлять и исправлять можно только во время ручного просмотра распознанного словаря.

Отсутствие пробелов – распространенный случай при работе с распознанным текстом (рис. 5). Например, в приведенном примере нет пробела между словом из левой части словаря (заглавное слово, с которого начинается словарная статья) и первым словом из правой части (часть, в которой объясняется заголовочная единица).

АБСОРБИ`РОВАТЬСЯсов.,несов. сорылу,сіңу,жұтылу.

АБСО`РБЦИЯж. физ., хим. абсорбция; сінірілу,жұтылу, сорылу.

АБСТРАГИ`РОВАНИЕ с.абстракциялаушылық.

АБСТРАГИ`РОВАТЬсов., несов. что абстракциялау,дерексіздендіру,жалпылау.

АБСТРАГИ`РОВАТЬСЯ сов.,несов. абстракциялану,дерексіздену.

АБСТРА`КТНОСТЬж. дерексіздік,абстрактылық.

АБСТРА`КТН||ый,-ая,-ое абстрактылы,дерексіз; ~ые понятиядерексіз ұғымдар; ~ое тождествоабстрактылы тепе-теңдік.

Рис. 5. Пример отсутствия пробелов

Во время обработки словарей знак переноса строки обычно означает начало новой статьи. Но так же встречаются лишние знаки переноса строки (рис. 6). Поэтому каждый случай необходимо анализировать.

ПЕРЕОБУЧА`ТЬ несов. см. переобучить.¶

ПЕРЕОБУЧА`ТЬСЯ несов. 1. см. переобучиться;¶

2 страд. отпереобучать.¶

ПЕРЕОБУЧЕ`НИЕ с. см. переобучить, переобучиться.-¶

ПЕРЕОБУЧИ`ТЬсов. кого-что қайтаоқыту, жаңаданүйретіпшығару.¶

Рис. 6. Пример лишнего знака переноса строки

В словарях часто встречаются толкования-отсылки. Например, во втором значении слова «переработаться» (рис. 7)

необходимо обратиться к слову «переработать» в 5 значении, где и будет дано толкование данного значения. Хорошо бы было, если толкование повторялось, а не использовалась отсылка, но в печатных словарях за счет отсылки экономилась бумага. Поэтому толкования-отсылки требуют дополнительной обработки.

ПЕРЕРАБО`ТА||ТЬСЯсов.·1. бойғасіну,·денегетаралу;·пища~ласьтамақбойғасінді;·2.·см.переработать5.¶

ПЕРЕРАБО`ТК||Аж.·1.·см.·переработать1–4;·2.·разг.·артықістелгенуақыт;·уплатитьза~уартықістелгенуақытүшінтөлем;·3.·(то,·чтопереработано)істепшығарылғанөнім,·ұқсатылғанбұйым.¶

ПЕРЕРАБО`ТОЧН||ЫЙ·~ая,·~оеұқсататын,·өңдейтін,·қайтажасайтын;·~ыйпунктөңдеупункті.¶

ПЕРЕРАСПРЕДЕЛЕ`НИЕс.см.перераспределить.¶

Рис. 7. Примеры толкования-отсылки

4. Заключение

Программная обработка двуязычных словарей позволила значительно сэкономить время для подготовки данных для загрузки в базу данных портала тюркской морфемы. Из-за особенностей строения различных словарей, для обработки каждого из них в программное приложение приходилось вносить изменения. Поэтому в дальнейшем планируется разработать универсальное приложение для обработки двуязычных словарей с возможностью настройки работы через интерфейс программы.

ЛИТЕРАТУРА

1. Сулейманов Д.Ш., Гатиатуллин А.Р., Альменова А.Б., Баширов А.М. Многофункциональная модель тюркской морфемы как база данных для лингвопроцессоров // Филология и культура. 2016. № 2 (44). С. 143-151.

2. Сулейманов Д.Ш., Гатиатуллин А.Р., Альменова А.Б., Баширов А.М. Многофункциональная модель тюркской морфемы: отдельные аспекты // В сборнике: TEL – 2016. Труды международной конференции по компьютерной и когнитивной лингвистике. Сер. "Интеллект. Язык. Компьютер" 2016. С. 168-171.

Dubrovina M.A.
Sankt-Petersburg State University,
Russia, Sankt-Petersburg

A FEW WORDS ABOUT THE IMPORTANCE OF THE SYNCHRONIC APPROACH IN THE ANALYSIS OF TURKIC MORPHOLOGICAL FORMS

Abstract. In any linguistic research, it is necessary to separate those meanings and functions of grammatical forms that existed in some previous periods of the development, and those that are productive and active at the moment that the researcher has chosen as the starting point of analysis. In Turkish, there are language tools that consist of two forms. From the historical point of view, each of the forms has its own grammatical meaning and has the right to be perceived as an independent unit. From the synchronic point of view, at the modern stage of language, these forms are a single whole, each form has its own meaning and cannot be analyzed separately from the entire form. Such analytical forms include complex verbal and adverbial constructions of modern Turkic languages.

Keywords: *Turkic languages, Turkic grammar, analytical forms, periphrastic forms, adverbial participle, synchrony*.

Дубровина М.Э.
Санкт-Петербургский государственный университет,
Россия, Санкт-Петербург

О ВАЖНОСТИ СИНХРОНИЧЕСКОГО ПОДХОДА ПРИ АНАЛИЗЕ ТЮРКСКИХ МОРФОЛОГИЧЕСКИХ ФОРМ

Аннотация. В любом лингвистическом исследовании необходимо отделять те значения и функции грамматических форм, которые существовали в некие предшествующие периоды развития данного языка, и те, которые являются продуктивными и действующими на тот момент, который выбран исследователем в качестве отправной точки анализа. В турецком языке выделяются языковые средства, которые состоят из двух

форм. С исторической точки зрении каждая из форм имеет свое собственное грамматическое значение и имеет право на то, чтобы восприниматься как самостоятельная единица. С синхронической же точки зрения, на современной этапе языка эти формы представляют собой единое целое, каждая форма уже своего собственного значения не имеет и не может анализироваться отдельно от всей формы. К таким сложным аналитическим формам можно отнести сложновербальные и обстоятельственные конструкции современных тюркских языков.

Ключевые слова: *тюркские языки, тюркская грамматика, аналитические формы, перифрастические формы, деепричастие, синхрония.*

1. Введение

Как известно, исследователь имеет возможность анализировать язык, исходя из двух точек наблюдения, во-первых, в исторической перспективе, т.е. рассматривать языковые факты в их диахроническом развитии, во-вторых, с позиции современности, т.е. наблюдать функционирование форм в настоящий момент, т.е. в синхронном срезе языка, совпадающим с каким-то определенным временным периодом. Такой синхронный срез будет представлять собой своего рода застывший отпечаток состояния языка и его системы, в котором все формы и модели имеют значения, актуальные для этого момента во времени.

Так, если исследователь утверждает, что он занимается турецким языком начала 21 века, то все обнаруживаемые им морфологические формы турецкого языка должны рассматриваться исходя исключительно из тех значений, которые они имеют на этот момент. Другими словами, необходимо отделять те значения и функции грамматических форм, которые существовали в некие предшествующие периоды развития данного языка, и те, которые являются продуктивными и действующими на тот момент, который выбран исследователем в качестве отправной точки анализа. Последнее утверждение, казалось бы, принимается всеми лингвистами, как

само собой разумеющееся. Однако практика показывает иное. В большинстве случаев, ученые продолжают описывать язык и его составляющие, полагаясь на те значения, которые пришли из прошлого, т.е. уже переставшими быть актуальными на момент наблюдения.

2. Статус бивербальных (аналитических и перифрастических) конструкций

В своем сообщении хотелось бы остановиться только на нескольких примерах, достаточно хорошо иллюстрирующих то, о чем было сказано выше.

Во-первых, это вопрос о статусе в современном турецком языке деепричастия –yıpи того глагола, который с ним сочетается, в конструкциях именуемых сложновербальными (бивербальных) типа, -yıp dur-. Этот вопрос представляет важность еще и потому, что практически во всех тюркских языках присутствуют аналогичные сочетания, в таких вариантах как: -п утыр (отыр, одур, ултыр), -п ят (чат), -п тур (тор-) [5].

Очень часто деепричастие в этих словосочетаниях рассматривается как некая самостоятельная форма, точнее, когда происходит анализ деепричастия, то одним из его значений нередко указывается возможность сочетания данного деепричастия с вспомогательными глаголами. Тем не менее, в указанных выше и аналогичных сочетаниях самостоятельное деепричастное значение формы уже утрачено, форма -yıp в таких случаях явно приобрела статус неотделяемой части единой аналитической или перифрастической конструкции, обладающей одним грамматическим значением (вида или способа действия) [4; 3]. То же можно сказать и о тех глаголах, которые употребляются в подобных сложновербальных образованиях. Они уже едва ли представляют собой самостоятельные лексические единицы, обладающие лексическим значением. Так, в турецком языке выделяются следующие аналитические показатели способов действия – -yıp kal-,-yıp dur-,-yıp git-,-yıp gel- [2]. С позиции синхронии можно говорить о том, что все эти сочетания представляют собой

отдельные морфологические аналитические показатели, каждый из них имеет свое значение, указывающее на тот или иной способ протекания исходного действия, в каждой форме – глаголы kal-, dur-, git-, gel- уже не воспринимаются в качестве лексических единиц обладающих своим значением - kal- оставаться, dur- стоять, git- идти, gel- приходить. Эти глаголы на настоящий момент выполняют исключительно техническую функцию, приобрели служебный статус и не должны отождествляться с имеющими то же самое звучание глаголами, от которых исторически возникли данные аналитические формы. На наш взгляд, на морфологический статус -yıp kal-,-yıp dur-,-yıp git-,-yıp gel- и подобных в других тюркских языках указывает повторяемость их одного и того же грамматического значения всякий раз, когда они употребляются после основы глагола. Раздельное написание этих элементов конструкции не может опровергать их уже ставшей грамматической сущность, т.к. на настоящий момент в системе турецкого языка обнаруживаются аффиксы, которые генетически восходят к сочетанию отдельных слов и форм. Например, показатель настоящего актуального времени имеет аффикс -ıyor, что представляет собой стяжение некогда деепричастной формы -yı и древнетюркского глагола движения jory-. Однако в синхронии языка едва ли носители осознают, что в этом аффиксе присутствует глагол. Поэтому бесспорно, что процесс перехода некогда обычных свободных словосочетаний деепричастной формы с самостоятельным глаголом на современном этапе развития языка прошел стадию морфологизации, в результате которого и -yıp, и глагол потеряли свои самостоятельные значения и слились в единое, но новое значение.

3. Конструкции, образованные с помощью глагольных имен

Второй вопрос, также порой связанный с неразличением понятий диахрония и синхрония является вопрос о статусе таких конструкций, как: -yacağı/ -dığızaman, -dığıiçin, -dığıgibiи многих

других подобных образований современного турецкого языка. Структура у них в большинстве случаев идентичная:

Основа глагола + показатель глагольно-именной формы + (падеж) + (аффикс принадлежности) послелог/служебное слово

1. Yirmi yaşında bir genç kız ol **masınarağmen** eski kafalı lıkta babasınataş çıkartıyordu. *'Несмотря на то, что она была молодой девушкой двадцати лет, она намного превосходила своего отсталого отца'*.

2. İçerigir **diğinizvakit** odadakimseyoktu. *'Когда вошли внутрь, в комнате никого не было'*.

3. Bugün buluş **tuktan başka** güzel bir müzeyi de ziyaret ettik. *'Мы не только сегодня встретились, но и посетили красивый музей'*.

Анализируя эти языковые средства лингвисты также зачастую продолжают в глагольных именах видеть самостоятельные формы, с которыми сочетаются послелоги или некие самостоятельные или служебные слова. Тем не менее, и это положение едва ли соответствует синхронному состоянию турецкого языка. Вероятно, в некие предшествующие периоды турецкого (османского) языка используемые в этих конструкциях формы с показателями –yacak / -dık воспринимались носителями языка в качестве продуктивных и самостоятельных компонентов словосочетаний, однако на сегодняшний момент это уже явно не так. Рассматриваемые конструкции представляют собой не свободные синтаксические словосочетания, а застывшие морфолого-синтаксические образования с единым (неделимым) значением. Возможно и в других тюркских языках можно обнаружить подобные же конструкции. Однако в академических грамматиках не всегда легко найти информацию об этом. Например, не всегда удается найти ответ на вопрос, присутствуют ли такие образования в татарском, башкирском, узбекском языках. По всей видимости, это происходит по той же причине, т.е. когда такие конструкции не всегда воспринимаются исследователями с синхронической точки зрения, в качестве перешедших в морфологию аналитических единиц.

Заключение

Таким образом, не видеть в рассмотренных выше сочетаниях морфологических форм является зачастую ошибочным, полагаясь исключительно на раздельно написание этих форм. В соответствии с известным утверждением В.В. Виноградова: «Морфологические формы — это отстоявшиеся синтаксические формы. Нет ничего в морфологии, чего нет или прежде не было в синтаксисе или лексике» [1]. Ведь на современном этапе присутствующая в разных тюркских языках временная форма -ыпты (ıptı) не понимается как сочетание деепричастия и аффикса прошедшего времени, а интерпретируется как единая форма давнопрошедшего времени. Точно также как в парадигме современной категории времени турецкого языка вычленяется форма с показателем -makta, которую необходимо интерпретировать как аффикс с единым темпоральным значением настоящего длительного времени, а не сочетанием масдара -mak и аффикса местного падежа -ta.

Список тем, при анализе которых необходимо отделять историческое от насущного можно продолжить. Актуальна она и для сферы словообразования, когда в качестве словообразовательных аффиксов перечисляют те показатели, которые таковыми были в предшествующие этапы и уже не работают в качестве словообразовательных на современном этапе (-ga/ka, например, в словах dalga, bölge, bilge, yufka).

ЛИТЕРАТУРА

1. Виноградов В.В. Русский язык. Грамматическое учение о слове. М. Л.: Учпедгиз, 1947. 784 с.
2. Гузев В.Г. Теоретическая грамматика турецкого языка /под ред. А.С. Аврутиной, Н.Н. Телицина. СПб.: Изд-во СПбГУ, 2015. 320 с.
3. Дубровина М.Э. Категория аспектуальности языка древнетюркских рунических памятников // Очерки по теоретической грамматике восточных языков / под ред. В.Г. Гузева. СПб: Издательский дом СПбГУ, 2011. С. 141-158.
4. Насилов Д.М. О грамматическом статусе некоторых сложновербальных конструкций // Востоковедение 9. Л.: Изд-во ЛГУ, 1984. С. 76-82.
5. Языки народов СССР Т. II. Тюркские языки. М.: Издательство «Наука», 1966. 531 с.

СЕКЦИЯ 4
ФОРМАЛЬНЫЕ МОДЕЛИ ДЛЯ ТЮРКСКИХ ЯЗЫКОВ

Abdurakhmonova N.
Tashkent state university of Uzbek language and literature named after Alisher Navoi,Uzbekistan, Tashkent
Aripov M.
National university of Uzbekistan,Uzbekistan, Tashkent
Norov A.
Karshi state university of Uzbekistan, Uzbekistan, Karshi

SYNTACTIC STRUCTURES FOR ONTOLOGICAL MODELS (AS EXAMPLE OF UZBEK LANGUAGE)

Abstract. The article deals with ontological models of syntactic structures of Uzbek language comparing example word combinations. According to syntactic models of the word combination divides two main types: nominal and verbal. Connecting of each words of models has three syntactic ways: agreement, government, adjoinment. In this paper analyzed syntactic relations and models in order to create meta language for NLP and other linguistic technology in the frame of Turkic languages under the project of «AP05132249 Processing of electron thesaurus of Turkic languages for creation multilingual information retrieval system and extracting knowledge» on agreement №132 on «12 » March 2018.

Keywords: ontological model; syntactic structures; relationship; grammatical classification; word combination; nominal and verbal adjournment; government; agreement

Абдурахмонова Н.
*ТГУ узбекского языка и литературы им. Алишера Навои, Узбекистан, Ташкен*т
Арипов М.
НУ, Узбекистан, Ташкент
Норов А.
КГУ, Узбекистан, Карши

СИНТАКТИЧЕСКИЕ СТРУКТУРЫ ДЛЯ ОНТОЛОГИЧЕСКИХ МОДЕЛЕЙ (НА ПРИМЕРЕ УЗБЕКСКОГО ЯЗЫКА)

Аннотация. В статье рассматриваются онтологические модели синтаксических структур узбекского языка на примере

сравнения словосочетаний. По синтаксическим моделям словосочетания делятся на два основных типа: именное и глагольное. Связь осуществляется тремя синтаксическими способами: согласование, управление, присоединение. В данной статье проанализированы синтаксические отношения и модели для создания метаязыка для НЛП и других лингвистических технологий тюркских языков в рамках проекта «AP05132249 Обработка электронного тезауруса тюркских языков для создания многоязычной системы поиска информации и извлечения знаний» (договор №132 от «12» марта 2018 года).

Ключевые слова: *онтологическая модель; синтаксические конструкции; отношение; грамматическая классификация; словосочетание; именная и устная отсрочка; правительство; соглашение*

I. Introduction

One of the linguistic properties of languages for natural language processing is syntax. There are two crucial components as constituencies of syntax: word combination and sentence. Syntactic parsing is crucial technology for each application of natural language processing: machine translation, question-answering system, information retrieval system and sentiment analysis, corpus linguistics. Consequently, building of the structure of text and word combinations plays essential role in order to identify the place of parts of speech. Each language has own linguistic peculiarities as according to typological system of languages. For example inflectional and agglutinative, having own ontological classification of parts of speech.

Word combination represents the combinations of words. Words belong to things and substance, quality, attribute, and action. Things and substance, quality, substance, attribute, and action interconnect each other in word combination, but they cannot apart from independently each other. Syntax of word combination is capability of adjoining of words that estimated as connection ways and schemata (forms) as well as components and forms of word combination associates closely with morphology. It is studied word

combination as a part of sentence and postfixes considered as morphological-syntactical category that joining each other's.

Word combination comprises semantically and grammatical attitudes of at least two words. One is component of word combination comes as head (governor) and other dependent word. Components interact each other's based on semantically and syntactical rules. Word combination plays role as nominative means of language by headword naming things, substance, quality, substance, attribute, and action interconnect each other's.

However, grammatical features can identify the functions of words in sentence there is not rigid word order in the sentence of agglutinative languages because of free place of parts of speech. For example in Uzbek it can be seen changeable position each component what focused part comes in the front of predicate:

Men bugun avtobusda ***universitetga*** *boraman.*

Men bugun universitetga ***avtobusda*** *boraman.*

Men universitetga avtobusda ***bugun*** *boraman.*

Predictable matching the role of parts of speech for Turkic languages seem difficult due to free placed in the text. However each language has own formal model can be used for parsing including several stages linguistically.

II. Related work

Ontological modeling is suitable for domain of concept to analyze in order specifies hierarchic attitudes of any entries.

If we compare ontological modeling syntactic structures of Uzbek and Kazakh languages, there are three types of syntactic relations between members of syntactic groups: agreement, government, adjoinment.

Three areas of work are essential for metadata to perform its functions: semantics to define the meaning of data, syntax to specify the data binding structure, and vocabulary to control the language [3].

Agreement is based on formal correspondence between members of a syntactic group in person and number of governor and subordinate. The definition in Uzbek there is some distinction that the model [NOUN+NOUN+Case] [Noun+POSS]: *talabalarning hammasi* in Uzbek, *мен келемін, оқушы оқиды, балалар жүгірді* in

Kazakh. Moreover, the predicate agrees with the subject in person and number: *Men o'quvchiman (Uzbek) - Мен оқушы (Kazakh).*

Government occurs by the case and particles of both languages according to what is nominal or verbal head word: *ukam uchun kitob, estalikka sovg'a, bog'dagi gullar (Uzbek); комиссияға мүше, өзіңе үлгі, елге жақсы, өзіңе өзің (Kazakh)*

Adjoinment can be expressed the words joined by the common grammatical function and meaning without any change in morphological forms between them: *qalin o'rmon, oq paxta, buyuk tarix; бұтақ күрек, отыз кітап, қызыл орамал, бұл үй, ол кім, қанша үлкен.*

Syntactic relation in word combination in Uzbek and Kazakh divides two types: nominal and verbal. In nominal word combination, nominal (adjective, noun, numeral, and pronoun) is considered governor.

III. **Ontological models of word combinations**

Ontology is used for formal and specialized concept and relations that belong to exact domain. Having advantage of ontology in NPL to create metalanguage in the sphere of machine translation (mainly, rule-based machine translation) or other purposes (information retrieval system, text analysis, annotation of text). Thanks to ontology, creating structure of information based on systematical and hierarchical data it aids to ease computational processing of the natural language. Effective way to create of ontology is representing OWL. "There are several types of ontologies. The word "ontology" can designate different computer science objects depending on the context. For example, an ontology can be:

— a thesaurus in the field of information retrieval or

— a model represented in OWL in the field of linked-data or

— a XML schema in the context of databases

— etc " [2].

Once analyzed on concrete artifact, a model can support relation and conclusion about important aspects of the underlying sphere. For the reason that any kind of model is a construct of understandings

according to a certain conceptualization. Furthermore, ontological model structure defines the set of grammatically construction in terms of the order parts of speech.

As similarities and distinctions of syntactic features between Uzbek and Kazakh languages compared by ontological models, it can ease to classify subclasses of relations of the words by different syntagmatic properties. Although Uzbek and Kazakh languages considered as agglutinative language, both of them have differences due to being allomorphs in terms of harmonies of phonemes in Kazakh.

Comparison of morphological peculiarieties, on the example of essential uzbek and kazakh languages

Tag	Name_English	UZB
WC	Word combination	ukam uchun sotib olmoq
COLC	Collocation	o‘z yog‘iga qovurilmoq
FP\ FCOLC	Free phrase\ Free collocation	xat yozmoq, kuchli iroda, ukamning kitobi
NP	Noun Phrase	bolalarning hammasi, intizomda birinchi, xushbo‘y hid
NA	Noun Adjoinment	ona vatan, bebaho sovg‘a
NG	Noun Government	ukam uchun sovg‘a, senga mukofot
NCS	Noun Collateral subordination	ukamning xati
VP	Verb Phrase	baland uchmoq, kulib gapirmoq
AGRM	Agreement	u o‘quvchi, mel keldim
SLP	Singular personal pronouns	Men talabaman
PPL	Plural personal pronouns	Ular talabalar

Table 1.

Predicative agreement relations between Uzbek and Kazak

Dependent	**Head**			
	N, Adj, Pro, NUM, ADV + personal endings		**V+Present / Future simple**	
	UZB	**KZ**	**UZB**	**KZ**
			-a / -y /-moqchi	**-a, -e, -й**
1 personal singular	-man	-мын, -мін, -бын, -бін, -пын, -пін	-man	-мын, -мін, -бын, -бін, -пын, -пін
2 personal singular	san	-сың,-сің	-san	-сың,-сің
2 personal singular formal	siz	-сыз, -сіз	-siz	-сыз, -сіз
3 personal singular		0	-di	-ды, -ді
1 personal plural	-miz	-мыз, -міз, -быз, -біз, -пыз, -піз	-miz	-мыз, -міз, -быз, -біз, -пыз, -піз
2 personal plural	-sanlar	-сыңдар, -сіңдер	-sanlar	-сыңдар, -сіңдер
2 personal plural formal	-sizlar	-сыздар, -сіздер	-sizlar	-сыздар, -сіздер
3 personal plural	0	0	0	-di /-dilar

Table 2.

Predicative agreement relations between Uzbek and Kazak

Dependent	**Head**					
	V+Past simple				**Condition**	
	UZB	**KZ**	**UZB**		**UZB**	**KZ**
	-b / -ib	**-п, -ып, -іп**	**--di**	**-ti, -ты**		
1 personal singular	-man	-пын, -пін	-m	-м	-y, -yin, -ay, -ayin	-йын, -йін, -айын, -ейін
2 personal singular	-san	-сың, -сің	-ng, -ing	-ң		0
2 personal singular formal	-siz	-сыз, -сіз	-ngiz, -ingiz	-ңыз, --ңіз	-ngiz, ingiz	-ңыз, -ңіз, -ыңыз, -іңіз
3 personal singular	-ti	-ты, -ті	0		-sin	-сын, -сін
1 personal plural	-miz	-пыз, -піз	-k	-қ, -к	-aylik	-йық, -йік, -айық, -ейік
2 personal plural	-sanlar	-сыңдар, -сіңдер	-nglar, -inglar	-ңдар, -ңдер	-ishgin	-ңдар, -ңдер, -ыңдар, -іңдер
2 personal plural formal	-sizlar	-сыздар -сіздер	-ingiz /-ingi-zlar	-ңыздар -ңіздер	ing-lar, in-giz	-ңыз-дар, -ңіз-дер, -ыңыз-дар, -іңіздер
3 personal plural	-lar	-ты, -ті	-lar		-sin, -sinlar	-сын, -сін

As we have seen above in many sides, especially personal endings and relativeness are common for both languages; however, there are some variations of affixes. There is instrumental case like *сегіз***бен** *кетті, алпыс***пен** *келді, әлдекім***мен** *қайтты* in the verbal government of Kazakh, but Uzbek this relativeness done by particles like *bilan, orqali, vositasida.*

To study probability of word combination modeling of parts of speech is useful for distribute to some group of syntactic relations. If we consider one example as classes and subclasses of words then we can see there are several kinds of types of models of word combinations in Uzbek:

Nominal adjointment:

1. Noun+Noun=> temir uskuna <=> Noun+Noun=> iron equipment
2. Adj.+Noun=> qulay imkoniyat <=>Adj.+Noun=> suitable opportunity
3. PNoun+Noun=>hamma ishtirokchilar <=>PNoun +Noun=> all participants
4. Num.Noun=> birinchi kun <=> Num.+Noun=>the first day
5. Gerund+Noun=> o‘qiyotgan qiz <=> Gerund+Noun=> reading girl
6. Infinitive+ Noun=>nishonlash kuni <=> Gerund +Noun=> celebrating day
7. Adv.+Noun=> sekin harakat <=>Adj.+Noun=> slow movement
8. (Noun+dagi)+ Noun=> devordagi rasm<=>Noun+be +Prep.+ Noun=> the picture is on the wall
9. (Infinitive+dagi)+ Noun=> ishlashdagi g‘ayrat <=> Noun+Prep.+Gerund=> enthusiasm in working
10. (Adv.+dagi)+ Noun=> yuqoridagi qavat <=> Adv+Noun=>upper floor
11. ↓PNoun+↓Adv. +Gerund+Noun=> (kimgadir) (sekin) o‘qib berayotgan qiz <=>Noun+Question word+be+Ving(Adv.) (to smb.) girl who is reading (slowly) to smb.
12. Noun|PNoun{ni, ga, da, dan}+Gerund+Noun=> maktabga ketayotgan qiz <=> Noun+Question word+be+Ving(Ad) +Prep.+Noun=> the girl who is going to school

13.Adj.+Gerund|Past participle+Noun=> yaxshi o‘qigan bola <=> Adj. | Adv+Gerund | Participle+Noun => well educated boy

14.Adv.+Gerund+Noun=>tez kelgan lahza <=> Adv.+Gerund+Noun =>fast coming time

15.(Noun+day|dek)+Adj.=> oyday oppoq<=>Adj.+like+Noun=> white like the moon

16.(Noun+dagi)+Adj.=> sinfdagi a’lochi<=> Adj.+Prep.+Noun=>the smart in the classroom

17.Adj.+Num.=> mo‘jizaviy yetti<=> Adj.+Num.=> marvelous seven

18.(Noun+dagi)+Num.=> rasmdagi bir<=>Num+Prep+Noun=>one in the picture

19.Noun+Infinitive=>kitob o‘qish (lekin *ism qo‘yish, nonushta qilish* bu so‘z birikmasi emas, ko‘makchi fe’lli so‘z qo‘shilmasi) <=> reading a book

20.Adj.+Infinitive=> qulay joylashish<=> Adj.+Gerund=>convenient placing

21.Adv.+Infinitive=> tez yeyish<=> Infinitive+ Adv. => to eat fast

Verbal adjoinment:

1. sifat+fe’l=> yaxshi o‘qimoq–reading well
2. ravish+fe’l=> astoydil o‘qimoq–studying hard
3. ravishdosh+fe’l=> kulib gapirmoq–speaking with smiling
4. Adj.+Verb=> yaxshi o‘qimoq<=>V+Adv=>read well
5. Adv.+Verb=> astoydil o‘qimoq<=>V+Adv=>study hard
6. Gerund+Verb=> kulib gapirmoq<=>V+Prep.+Gerund=>speak with smiling

Nominal government:

1. Noun+ dan+Noun=> Andijondan kelish <=>Noun+Prep.+Noun=> the letter from Andijan.
2. Noun+↓dan ham |↓dan ko‘ra+Adj.+↓roq=>onadan mehribon <=> Adj. + than + Noun=> kinder than mother .
3. Noun+dan+Infinitive=>ustozdan so‘rash <=> Gerund+Prep.+Adj.+Noun =>asking from the master.
4. Gerund+ {Noun} dan +Infinitive => bilgandan so‘rash <=> Gerund + Prep. + Adj. +Noun=> asking from educated person.

5. Gerund + dan + Adj.=>ko‘rgandan gumon<=> Adj. + {Prep.} + Gerund=>doubtful of seeing.

6. Noun | PNoun + dan + Num.=>hammadan birinchi <=> Num.+Prep.+PNoun|Noun=> the first all of them.

7. Noun | PNoun + dan +Adj.=>hammadan ustun <=> Adj. + Prep. + Noun | PNoun=> the best of all of them.

8. Adj. + ↓lar+dan+Num.=> a’lochilardan ikkitasi <=> Num. + Prep. + Adj.=> two of the smarts.

9. Adv. + dan+Adv.=> kechagidan erta<=>Adv.+than+Adv. => earlier than yesterday.

10.Adv. + dan+Infinitive=>ko‘pdan bilish<=> Infinitive + Prep. + Adj.=>to know from many (people).

11.Num.+dan+Num.=>yuztadan bittasi<=>Num.+Prep.+one out of hundred.

12.Noun+ga+Noun=> vatanga muhabbat <=>Noun+to + Noun=>love to homeland.

13.PNoun+ga+Noun=> hammaga do‘st <=> Noun + to + PNoun =>friend to everybody.

14.Gerund+ga+Noun=>o‘qiyotganga omad<=>Noun+to+Gerund+Noun=> luck to studying man.

15.Infinitive+ga+Noun=>o‘qishga mehr <=>Noun + to + Infinitive =>love to study.

16.Infinitive|Noun+ga+Infinitive=>o‘qishga intilish<=> Gerund|Noun +to+ Infinitive=> trying to study.

17.Noun+ga+Adv.=>bayramga yaqin<=>Adv. + to + Noun => close to holiday.

18.Noun+da+Noun=>yozuvdv xato<=>Noun+Prep.+Noun=> mistake in writing.

19.Noun+da+Num.=>tartibda birinchi <=>Num.+Prep. + Noun => the first of order.

20.Noun|PNoun+da+Adj.=>menda ko‘p<=>PNoun|Noun+have+Adj.=> I have many | much.

21.Adj.+ni+Infinitive=> qahramonni eslash <=>Infinitive+ Noun =>to remember hero.

22.Noun+ni+Infinitive=>farzandni sog‘inish<=>Gerund+Noun=> missing the child.

Verbal government

1. Noun+ga+Verb=>maktabga bormoq <=>Infinitive+Prep.+Noun=>to go to school.

2. Noun+ga+Infinitive=>daftarga yozmoq<=>Infinitive+Noun=>to write notebook.

3. Noun|Pronoun+dan+Verb=>universitetdan qaytmoq<=> Verb +Prep. +Noun=>return from the university.

4. Noun|Pronoun +ni+Verb=>hikoyani o‘qimoq <=>Infinitive +Noun=> to read story.

5. Noun +ni+ravishdosh=>ishni bajarib<=>Gerund+Noun=>doing work.

6. Noun+da+Verb=>maktabda o‘qimoq <=>Infinitive+Prep. +Noun=>to study at school.

7. Noun+da+Gerund=>osmonda uchib kelayotgan=>Gerund +Prep.+Noun=> flying in the sky.

Taking into consideration syntactic structures of word combination analyzed sentences via parts of speech. Usage protégé program input all linguistic data of structures, and then it becomes easy to split up components of sentence according to input ontological hierarchy.

Conclusion

This work has been implemented by project as mentioned above. Creation of Meta language for Turkic languages is crucial in order to NLP and other special purpose text analysis. All classes and subclasses with attributes input Protégé program to work further work. In spite of diversity of languages, there is commonness of grammatical rules among the Turkic languages. Entities inputted in Protégé as classes including object properties, data properties, individuals, annotation etc. Ontology grammatical rules of Turkic languages (Uzbek, Kazakh, Tatar, Turkish, Kyrgyz) could be used for computational language processing in perspectives.

REFERENCES

1. Abdurakhmonova N., Aripov M. Uzbek ontology of Uzbek language as example of adjective // Turklang – 2018. 6-international conference, Tashkent, 2018. – P. 234-237.

2. Catherine Roussey, Francois Pinet, Myoung Ah Kang, and Oscar Corcho. An Introduction to Ontologies and Ontology Engineering (G. Falquet et al., Ontologies in Urban Development Projects, Advanced Information 9 and Knowledge Processing 1, DOI 10.1007/978-0-85729-724-2_2, © Springer-Verlag London Limited 2011).

3. Duval, E., W. Hodgins, S. Sutton, and S. L. (2002). Weibel. Metadata principles and practicalities. 8(4): http:// www.dlib.org /dlib/april02/weibel/04weibel. Html.

Mongush Ch.M.
Tuva State University,
Russia, Tuva, Kyzyl

RECOGNITION OF THE AUTHOR'S STYLE OF STORYTELLERS OF THE TUVINIAN HEROIC EPOS USING THE METHODS OF ANALYSIS OF FORMAL CONCEPTS

Abstract. One of the important components of the ethnocultural heritage of the Republic of Tuva is the analysis of the texts of Tuvan heroic epic. The scientific base of the Tuvan Institute for Humanitarian and Applied Socio-Economic Research of the Republic of Tuva contains tape and handwritten records of all genres of Tuvan folklore, including about 300 Tuvan heroic epics in old dilapidated editions. Currently, teachers and students of Tuvan State University have created an electronic collection "Tuvan Heroic Epics" and introduced into the electronic corpus of the Tuvan language. This collection contains digitized texts of Tuvan heroic epics, their meta-descriptions and information about storytellers. Using this electronic collection, it is possible to solve linguistic and philological problems, which are reduced to the problem of conceptual modeling of the collection "Tuvan Heroic Epics". This article deals with the problem of establishing the author's style of storytellers when describing horse equipment in the works of the Tuvan heroic epic. To solve this problem, an algebraic approach is used, which in the literature is called the of the formal concept analysis. Within the framework of the formal concept analysis, the collection "Tuvan Heroic Epics" is presented in the formal context. Then the solution to the problem is aimed at identifying the set of all formal concepts of the formal context and linking them into a lattice. The resulting lattice serves as a conceptual model and a basis for solving the problem.

Keywords: *electronic corpus of the Tuvan language; Tuvan heroic epic; the author's style of the storyteller; pattern recognition; analysis of formal concepts.*

Монгуш Ч.М.
Тувинский государственный университет,
Россия, Тува, Кызыл

РАСПОЗНАВАНИЕ АВТОРСКОГО СТИЛЯ СКАЗИТЕЛЕЙ ТУВИНСКОГО ГЕРОИЧЕСКОГО ЭПОСА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ АНАЛИЗА ФОРМАЛЬНЫХ ПОНЯТИЙ

Аннотация. Одной из важных составляющих этнокультурного наследия Республики Тыва является тувинский героический эпос. Научная база Тувинского института гуманитарных и прикладных социально-экономических исследований Республики Тыва содержит магнитофонные и рукописные записи всех жанров тувинского фольклора, в том числе около 300 тувинских героических эпосов в старых ветхих изданиях. В настоящее время преподавателями и студентами Тувинского государственного университета создан электронный сборник «Тувинские героические сказания» и введен в электронный корпус тувинский язык. Этот сборник содержит оцифрованные тексты тувинских героических эпосов, их метаописания и сведения о сказочниках. С помощью этого электронного сборника можно решать лингвистические и филологические задачи, которые сводятся к проблеме концептуального моделирования сборника «Тувинские героические сказания». В статье рассматривается проблема установления авторского стиля рассказчиков при описании конского снаряжения в произведениях тувинского героического эпоса. Для решения этой проблемы используется алгебраический подход, который в литературе называется анализом формальных понятий. В рамках формального концептуального анализа сборник «Тувинские героические сказания» представлен в формальном контексте. Затем решение проблемы направлено на выявление множества всех формальных понятий формального контекста и связывание их в решетку. Полученная решетка служит концептуальной моделью и основой для решения задачи.

Ключевые слова: *электронный корпус тувинского языка; тувинский героический эпос; авторский стиль сказителя; распознавание образов; анализ формальных понятий.*

1. Введение

Электронная коллекция «Тувинские героические сказания» – информационная составляющая электронного корпуса тувинского языка [1]. В этой коллекции представлены паспорта тувинских героических сказаний в виде объектно-признаковой таблицы [11]. В таблице каждый столбец отвечает некоторому признаку, а строка устанавливает признаковое описание того или иного эпоса. В паспорт произведения входят сведения о сказителе, библиографические сведения, жанровые, стилевые и другие особенности текста. Таким образом, объектно-признаковая таблица является моделью представления электронной коллекции «Тувинские героические сказания», что допускает применение математических методов при решении прикладных задач таких, как установление авторских особенностей сказителей, использование языковых стандартов и клише, применение диалектных вариантов и т.д. [12; 8]. Эти задачи сводятся к задаче концептуального моделирования электронной коллекции «Тувинские героические сказания» [15].

В данной статье рассматривается задача установления авторского стиля сказителей при описании доспехов коня в произведениях тувинского героического эпоса. Для решения этой задачи применяется математический аппарат – анализ формальных понятий. Он возник как прикладное направление теории решеток [18; 19]. Анализ формальных понятий на основе алгебраической теории решеток Г. Биркгофа позволяет построить смысловую структуру предметной области [4]. В рамках этого подхода электронная коллекция «Тувинские героические сказания» представляется формальным контекстом. Тогда решение рассматриваемой задачи направлено на выявление множества всех формальных понятий заданного формального контекста и связывание их в решетку формальных понятий. В свою очередь, построенная решетка является

концептуальной моделью предметной области и основой для решения прикладных задач [10].

2. Концептуальное моделирование коллекции «Тувинские героические сказания»

Рассмотрим основные понятия математического аппарата, применяемые в настоящей статье [4; 5; 7].

Пусть для рассматриваемой предметной области определены непустые конечные множества:

G – множество объектов,

M – множество признаков.

Непустое бинарное отношение инцидентности I между этими множествами трактуется следующим образом: для $g \in G$ и $m \in M$ пара (g, m) означает, что объект g обладает признаком m и наоборот, признак m свойственен объекту g [17]. Тогда в анализе формальных понятий тройка $K = (G, M, I)$ называется формальным контекстом предметной области. В таблице 1 представлен пример матрицы инцидентности I формального контекста $K = (G, M, I)$, где $G = \{1, 2, 3, 4\}$, $M = \{a, b, c, d\}$. А символ «+» означает отношение инцидентности, т.е. объект g обладает признаком m.

Таблица 1.

Формальный контекст $K = (G, M, I)$

g / m	a	b	c	d
1	+		+	
2			+	+
3	+			+
4	+	+	+	
5		+	+	

Если $A \subseteq G$ и $B \subseteq M$, то пара операторов

$$A' = \cap_{g \in A} g' = \{m \in M \mid \forall g \in A\ (g, m) \in I\}, \quad (1)$$

$$B' = \cap_{m \in B} m' = \{g \in G \mid \forall m \in B\ (g, m) \in I\}. \quad (2)$$

называется отображениями Галуа и задает соответствие между частично упорядоченными множествами $(2^G, \subseteq)$ и $(2^M, \subseteq)$. В этом

случае A'– множество признаков, характерных всем объектам из A, а B'– совокупность объектов, обладающих всеми признаками из B. Двойное применение отображения Галуа определяет оператор замыкания на совокупность всех подмножеств множеств объектов, признаков, и обозначается A'' и B''.

Формальным понятием называется пара множеств (A, B), где $A \subseteq G$ и $B \subseteq M$, если выполняется $A' = B$ и $B' = A$ [Mongush, 2019]. Другими словами, множества A, B образуют формальное понятие (A, B) в формальном контексте K= (G, M, I) тогда и только тогда, когда

$A = A''$ и $B = B''$. (3)

Рассмотрим подмножества A = {4, 5} и B = {b, c} множеств G, M из примера. Для подмножеств A и B найдем A'' и B'', используя формулы (1) и (2):

$A'' = (A')' = (\{b, c\})' = \{4, 5\} = A$.

$B'' = (B')' = (\{4, 5\})' = \{b, c\} = B$.

Тогда согласно формуле (3) пара (A, B) образует формальное понятие формального контекста K= (G, M, I), т.е. объекты 4 и 5 обладают признаками b, c. В таблице 1 формальное понятие (A, B) выделено двойными линиями. Определение формального понятия полностью соответствует традиционной трактовке термина «понятие» в гуманитарных науках. Объемом формального понятия является множество A, а содержанием – множество B. А также если уменьшить объем формального понятия, то увеличивается его содержание и наоборот, если уменьшить содержание, то объем формального понятия станет больше.

Обозначим через FC – множество всех формальных понятий формального контекста K= (G, M, I). Тогда введем на FC отношение частичного порядка $\sqsubseteq$:

(A_1, B_1) $\sqsubseteq$ (A_2, B_2), если $A_1 \subseteq A_2$ (или $B_2 \subseteq B_1$),

где A_1, $A_2 \subseteq G$ и B_1, $B_2 \subseteq M$.

На множество всех формальных понятий FC определим операции пересечения $\sqcap$ и объединения $\sqcup$ при помощи $\cap$ и $\cup$:

(A_1, B_1) $\sqcap$ (A_2, B_2) = ($A_1 \cap A_2$, $(A_1 \cap A_2)'$),

(A_1, B_1) $\sqcup$ (A_2, B_2) = ($(B_1 \cap B_2)'$, $B_1 \cap B_2$).

В этом случае, частично упорядоченное множество (FC, $\sqsubseteq$) образует решетку L = (FC, $\sqcap$, $\sqcup$) [21], которая называется

решеткой формальных понятий контекста K= (G, M, I) [9]. Решетка формальных понятий формального контекста является концептуальной моделью, которая определяет смысловую структуру исследуемой предметной области, и служит основой для решения прикладных задач [6]. В качестве примера рассмотрим задачу распознавания авторских особенностей при описании снаряжений коня в текстах тувинского героического эпоса.

3. Установление авторского стиля сказителей тувинского героического эпоса

Для решения задачи распознавания авторского стиля сказителей с помощью математического аппарата анализа формальных понятий необходимо:

1) формировать формальный контекст K= (G, M, I) тувинских героических сказаний;

2) найти множество всех формальных понятий FC;

3) построить концептуальную модель L рассматриваемой предметной области;

4) производить запросы на установление авторского стиля сказителей.

Формирование формального контекста K= (G, M, I) осуществляется на основе баз данных «Тувинские героические сказания», «Клише и стандарты в текстах тувинских героических сказаниях» и разработанной программы «Программа формирования контекста для электронной коллекции Тувинские героические сказания» [2; 3; 13; 14]. Формальный контекст K = (G, M, I) содержит 12 тувинских героических сказаний:

G = {Демир-Шилги аттыг Тевене-Мөге; Мөге Шагаан-Тоолай; Танаа-Херел; Каңгывай-Мерген; Алдын-Чаагай; Хан-Шилги аттыг Хан-Хүлүк; Арзылаң-Кара аттыг Чечен-Кара Мөге, Арзылаң-кара аттыг Хунан-Кара, Сарыг-Хемниң иштин чурттаан Тавын-Хаан, Хан-Шилги аттыг Хан-Күчү-Маадыр, Алдын-сарыг аттыг Аңчы-Кара, Элестей ашак}, а в качестве признаков выбираются языковые стандарты, описывающие снаряжения коня

M = {стандарты, характеризующие седло; стандарты, характеризующие хлыст; стандарты, характеризующие потник; стандарты, характеризующие аркан, стандарты, характеризующие узду; стандарты, характеризующие лассо; стандарты, характеризующие стремена}.

Нахождение множества всех формальных понятий FC и построение решетки формальных понятий L реализуется с

помощью разработанной программы «Программа FCACorpus концептуального моделирования тувинских текстов методами анализа формальных понятий» [16]. Применение данного комплекса программ к построенному формальному контексту $K = (G, M, I)$ дало множество FC, содержащее тридцать формальных понятий. С помощью найденных формальных понятий программа FCACorpus построила концептуальную модель, которая определяет смысловую структуру исследуемой предметной области. Данная решетка формальных понятий является основой для решения лингвистических и филологических задач, в том числе распознавания авторских особенностей сказителей тувинских героических сказаний.

Для установления авторских особенностей необходимо выполнять запросы на выявление знаний из решетки формальных понятий. Например, задан запрос: необходимо найти все языковые стандарты, которые характеризуют авторский стиль сказителя Ооржак Ч-Х.Ч., т.е. найти все формальные понятия, включающие в содержании Ооржак Ч-Х.Ч.

При реализации данного запроса программа FCACorpus отбирает из 30 формальных понятий 8 и строит решетку, которая описывает авторские особенности сказителя Ооржак Ч-Х.Ч. (рис. 1).

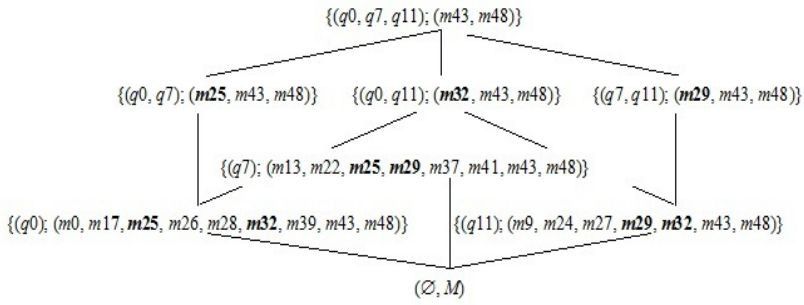


Рис. 1. Решетка формальных понятий, характеризующая авторский стиль сказителя Ооржак Ч-Х.Ч.

На рис. 1 формальное понятие

{(*q*0, *q*7, *q*11); (*m*43, *m*48)} = {(Демир-Шилги аттыг Тевене-Мөге, Арзылаң-кара аттыг Хунан-Кара, Элестей ашак); (Ооржак Чанчы-Хөө Чапаажыкович, Барун-Хемчикский р-н)}

означает, что авторский стиль сказителя Ооржак Чанчы-Хөө Чапаажыковича из Барун-Хемчикского района выявляются в эпосах Демир-Шилги аттыг Тевене-Мөге, Арзылаң-кара аттыг Хунан-Кара, Элестей ашак.

В решетке формальные понятия означают:

- {(*q*0, *q*7); (*m*25, *m*43, *m*48)} = {(Демир-Шилги аттыг Тевене-Мөге, Арзылаң-кара аттыг Хунан-Кара); (Алдын шалба, Ооржак Чанчы-Хөө Чапаажыкович, Барун-Хемчикский р-н)},
- {(*q*0, *q*11); (*m*32, *m*43, *m*48)} = {(Демир-Шилги аттыг Тевене-Мөге, Элестей ашак); (Алдан кулаш сыдым, Ооржак Чанчы-Хөө Чапаажыкович, Барун-Хемчикский р-н)},
- {(*q*7, *q*11); (*m*29, *m*43, *m*48)} = {(Арзылаң-кара аттыг Хунан-Кара, Элестей ашак); (Хүмүш чүген, Ооржак Чанчы-Хөө Чапаажыкович, Барун-Хемчикский р-н)}.

Таким образом, при описании аркана сказитель Ооржак Чанчы-Хөө Чапаажыкович применяет клише «Алдын шалба», а лассо характеризует с помощью языкового стандарта «Алдан кулаш сыдым», сказитель изображает узду, применяя стандарт «Хүмүш чүген».

В решетке последние формальные понятия описывают отдельные произведения сказителя Ооржак Чанчы-Хөө Чапаажыкович со своими характерными языковыми стандартами.

Аналогичным образом выполняется запрос: необходимо найти все языковые стандарты, которые характеризуют авторский стиль сказителя Ондар Тевек-Кежеге. В этом случае программа FCACorpus выдала решетку формальных понятий, которая приведена на рис. 2.

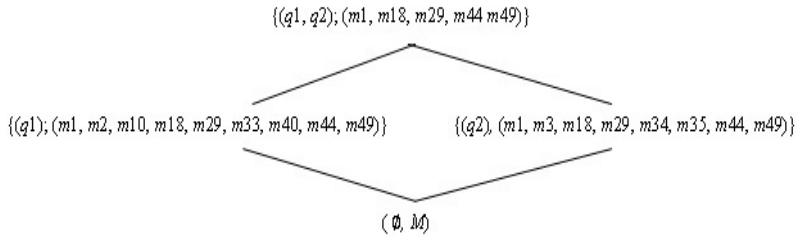


Рис. 2. Решетка формальных понятий формальных понятий, характеризующая авторский стиль сказителя Ондар Тевек-Кежеге

В решетке согласно формальному понятию{(*q*1, *q*2);(*m*1, *m*18, *m*29, *m*44 *m*49)} = {(Мөге Шагаан-Тоолай, Танаа-Херел); (Арт болган алчайган-калчайган эзер, Хөл болган хөлбең кара чонак, Хүмүш чүген, Ондар Тевек-Кежеге, Сут-Хольский р-н)} сказителю Ондар Тевек-Кежеге из Сут-Хольского района присущи языковые стандарты «Арт болган алчайган-калчайган эзер», «Хөл болган хөлбең кара чонак», «Хүмүш чүген» при описании седла, потника и узды.

Следует заметить, что сказители Ооржак Чанчы-Хөө Чапаажыкович и Ондар Тевек-Кежеге используют общий языковой стандарт««Хүмүш чүген» при описании узды.

Заключение

Экспертами было установлено, что полученные результаты в статье являются лингвистически правильными, то есть соответствуют действительности. Для полного анализа авторских особенностей сказителей тувинского эпоса с помощью анализа формальных понятий необходимо расширить формальный контекст. Также из полученных результатов следует, что созданные средства могут быть использованы для решения аналогичных задач анализа текстов в рамках электронного корпуса тувинского языка.

ЛИТЕРАТУРА

1. Бавуу-Сюрюн М. В. Вопросы создания электронных ресурсов тувинского языка: некоторые итоги, неотложные задачи и перспективы [Электронный ресурс] // Новые исследования Тувы. 2016, № 4. URL: https://nit.tuva.asia/nit/article/view/610 (дата обращения: 28.09.2020).

2. Бавуу-Сюрюн М.В., Далаа С.М., Монгуш Ч.М., Ондар, М.В. Тувинские героические сказания. Свидетельство о государственной регистрации базы данных № 2017620090 / Зарегистрировано в Реестре баз данных 19 января 2017 г.

3. Бавуу-Сюрюн М.В., Далаа С.М., Монгуш Ч.М., Ондар М.В. Клише и стандарты в текстах тувинских героических сказаниях. Свидетельство о государственной регистрации базы данных № 2017620024 / Зарегистрировано в Реестре баз данных 10 января 2017 г.

4. Биркгоф Г. Теория решеток. М.,1984: Наука, 568 с.

5. Быкова В.В., Монгуш Ч.М. Алгоритмы концептуального моделирования и классификации текстов в корпусе тувинского языка // Программные продукты и системы, 2017. Т. 30. № 3. С. 487–495. DOI: 10.15827/0236-235X.119.487-495

6. Быкова В.В., Монгуш Ч.М. Декомпозиционный подход к исследованию формальных контекстов // Прикладная дискретная математика. 2019. № 44. С. 113–126. DOI: https://doi.org/10.17223/20710410/44/9 (дата обращения: 28.09.2020).

7. Гретцер Г. Общая теория решеток М.: Мир, 1982. 456 с.

8. Карелова О.В. К вопросу изучения индивидуального стиля автора // Известия Российского государственного педагогического университета им. А.И. Герцена. 2006. Т. 20. № 3. С. 24–29.

9. Кузнецов С.О. Автоматическое обучение на основе анализа формальных понятий // Автоматика и телемеханика. 2001. № 10. С. 3–27.

10. Монгуш Ч. М. Распознавание индивидуального авторского стиля сказителей тувинского героического эпоса [Электронный ресурс] // Экономика и менеджмент систем управления. 2018. Т. 29. № 3.1. С. 184–194. URL: http://www.sbook.ru/emsu/ (дата обращения: 28.09.2020).

11. Монгуш Ч. М., Ондар М.В. База данных и средства создания контекстов для представления и анализа тувинского героического эпоса // Программные продукты, системы и алгоритмы. 2017. № 3. С. 1–6. DOI: 10.15827/2311-6749.24.261.

12. Монгуш Ч.М. Метатекстовая разметка в Национальном корпусе тувинского языка: структура и функциональные возможности [Электронный ресурс] // Новые исследования Тувы. 2016. № 4. С. 1–8. URL: https://nit.tuva.asia/nit/article/view/613 (дата обращения: 28.09.2020).

13. Монгуш Ч.М. Программа формирования контекста для электронной коллекции «Тувинские героические сказания» [Электронный ресурс] // Инженерный вестник Дона. 2018. № 2(49). С. 119–128. URL: http://www.ivdon.ru/ru/magazine/archve/N2y2018/5039 (датаобращения: 28.09.2020).

14. Монгуш Ч.М. Программа формирования контекстов в корпусе тувинского языка. Свидетельство о государственной регистрации программы для ЭВМ № 2018618908 / Зарегистрировано в Реестре программ для ЭВМ 23 июля 2018 г.

15. Монгуш Ч.М. Разработка метода и средств фрагментации и дефрагментации формальных контекстов : автореф. дисс. ... канд. физ.-мат. наук. Красноярск, 2019. 19 с.

16. Монгуш Ч.М., Быкова В.В. Программа FCACorpus концептуального моделирования тувинских текстов методами анализа формальных понятий. Свидетельство о государственной регистрации программы для ЭВМ № 2018618907. / Зарегистрировано в Реестре программ для ЭВМ 23 июля 2018 г.

17. Bykova V.V., Mongush Ch.M. On Algebraic Approach of R. Wille and B. Ganter in the Investigation of Texts // Journal of Siberian Federal University. Mathematics and Physics. 2017. Vol. 10. № 3. P. 372–384. DOI: 10.17516/1997-1397-2017-10-3-372-384

18. Ganter B., Wille R. Formal Concept Analyses: Mathematical Foundations. Springer Science and Business Media, 2012. 314 p.

19. Kuznetsov S.O., Ganter B., Eklund P.W., Sertkaya B. Machine Learning and Formal Concept Analysis // 2th International Conference on Formal Concept Analysis: proceedings. Berlin Heidelberg : Springer, 2004. Vol. 2961. P. 287–312.

20. Mongush Ch.M., Bykova V.V. On decomposition of a binary context without losing formal concepts // Journal of Siberian Federal University. Mathematics and Physics. 2019. Vol. 12. № 3. P. 323–330. DOI: 10.17516/1997-1397-2017-10-3-372-384

21. Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts // 7th International Conference on Formal Concept Analysis: proceedings. Darmstadt, Germany : Springer-Verlag, 2009. P. 314–339.

Pankov P.S., Bayachorova B.J., Karabaeva S.J.
Institute of Mathematics of NAS of KR, KNU named after J.Balasagyn, KSUCTA n.a. N. Isanov, Kyrgyzstan, Bishkek

MATHEMATICAL MODELS OF INTERACTIVE EDUCATIONAL SOFTWARE FOR HUMAN CONTROL

Abstract. In this paper we consider a general mathematical model for interactive software. We carry out a survey of existing and conceivable software, including one proposed and developed with author's participation. The paper highlights the objectives and specifics of various software including "independent presentation" of an object.

Keywords: *human control, mathematical model, computer model, classification, application, training, learning.*

Панков П.С., Баяшорова В.Ж., Карабаева С.Ж.
Институт математики НАН КР, КНУ им. Я. Баласагына, КГУСТА им. Н. Исанов, Кыргызстан, Бишкек

МАТЕМАТИЧЕСКИЕ МОДЕЛИ ИНТЕРАКТИВНОГО УЧЕБНОГО СОФТА ДЛЯ УПРАВЛЕНИЯ СО СТОРОНЫ ЧЕЛОВЕКА

Аннотация. Рассмотрена общая математическая модель для интерактивного программного обеспечения. Проведен обзор существующих и возможных видов программного обеспечения, в том числе предложенного и разработанного авторами. В статье подчеркиваются цели и особенности различных видов программного обеспечения, в том числе "независимое представление" объекта.

Ключевые слова: *человеческое управление, математическая модель, компьютерная модель, классификация, приложение, тренировка, обучение*.

1. Introduction

Training devices for hunting, horsemanship and war are known since ancient times. Mechanical flight simulators appeared together with the development of aviation. Computers gave the opportunity to create simulations with real-time feedback and elements of virtual reality. These ideas were also implemented in computer games. Educational computer software and educational games developed together with the development of personal computers.

The paper contains a general mathematical model of such software, remarks on its implementation as a computer model, a list of known and possible kinds of software (some of which has been implemented with authors' participation).

2. Mathematical model for human control

$N_0:=\{0,1,2,...\}$ contains values of discrete time t; $R_+:=[0,\infty)$;

Denote X as the space of states x (including virtual media and objects in it); $X_0 \subset X$ as the set of targets; $Q: X \to R_+$ as the target function to be minimized;

V as the set of observable (affectable by human interaction) elements of X; $W: X \to V$ is a given function;

P as the set of random elements p;

U as the set of possible actions u by the user (control).

We will consider discrete models. Continuous models are obtained from discrete ones by setting time divisions/steps to zero.

We propose the system

$x[0]$ $(x[0]=Z(p[0]))$ *is given (either* $x[0] \notin X_0$ *or* $Q(x[0]) > 0$*);* (1)

$v[t]=V(x[t])$; $x[t+1]=F(x[t], u[t], p[t])$, *or* $x[t+1]=x[t]+G(x[t], u[t], p[t])$, $t \in N_0$ (2)

where $u[t]$ is the action of the user influenced by information $v[t]$; $Z(p):P \to X$ is a function of random generation of initial data.

The goal is either to reach $x[t+1] \in X_0$ in minimal time or to minimize $Q(x[t+1])$ in a given time.

Two options of input *$x[0]$ by the user* and *random* give us two modes: learning mode and exam mode.

In advanced software *TaskLang* [6] the user can choose functions F (or G) too.

It may be $x=\{x_1,...,x_n\}$, $x_1,...,x_n$ are input independently; it is a necessary-collective task for n users.

The principle of duality [4]: (narrow V and wide U) and (wide V and narrow U) yield similar efficiency.

This principle extends for different kinds of human activity: Duality of available information and available goal achievement capacity.

3. Computer model specifics

3.1. Input of information *v[t]*

- common (by means of eyesight, hearing, vibration - vestibular apparatus);
- by means of special devices (earphones, binocular displays);
- to brain immediately.

3.2. Output of control *u[t]:*

- common (by means of hands, foots and voice; by top of head in diving suits);
- by reading nerve impulses in hands [5];
- from brain immediately.

General conclusion from [5]: using appropriate equipment for feedback, each physiological display (breath, pulse etc.) by human or animal with cognitive ability (ape, dolphin, dog) can be used for control.

3.3. There is Galileo-Einstein's principle of relativity: if we observe uniform movement of an object towards us then we cannot detect whether the object is moving, or we are.

For virtual motion the condition of uniformity (i.e. no acceleration) is not necessary. Hence, we receive a principle of relativity in virtual motion: if we observe movement in a kinematical space [2] then we can interpret it either movement of space toward us or our movement toward space.

The first interpretation prevails in software for scientific purposes (Mathematica, MathLab, MathCad) and the second one does in computer games.

4. Some cases of computer-human control

4.1. Simulation cases of real control where real training is too difficult, expensive or dangerous include: spaceship, aircraft, boat,

U-boat, artillery, launching big rockets, manufacturing processes, medical operations. They consist of random generation of media and random generation of emergences. Simulation made for a crew (for instance, pilot, co-pilot and navigator at aircraft) is an example of necessary-collective activity.

Remark. Some simulators are mixed computer-mechanical solutions that involve vibration and physical inclinations.

4.2. Computer games. Notes:

- some computer games arose from items listed in 4.1;
- computer games involving simulations of real objects (geographical map, concrete vehicles) may be considered educational;
- there are some hints in computer games useful to forthcoming proposals.

Remark. We do not consider games "person versus person" and "team versus team" by means of computers.

4.3. Imitation of physical-chemical experiments - "virtual laboratory".

4.4. Enhancing virtual reality. We [2] proposed to present abstract spaces in form close to presentation of the metric space.

Definition. A pair: a set X of points and a set K of **routes** is said to be a **kinematic space (**each route M, in turn, consists of the positive real number T_M (**time** of route) and the function $m_M : [0, T_M] \to X$ (**trajectory** of route)) if the following conditions are fulfilled: (K1) For $x_0 \neq x_1 \in X$ there exists such $M \in K$ that $m_M(0) = x_0$ and $m_M(T_M) = x_1$, and the set of values of such T_M is bounded with a positive number below (infinitely fast motion is impossible); (K2) If $M = \{T_M, m_M(t)\} \in K$ then the pair $\{T_M, m_M(T_M - t)\}$ is also a route of K (the reverse motion is possible); (K3) If $M = \{T_M, m_M(t)\} \in K$ and $T^* \in (0, T_M)$ then the pair: T^* and function $m^*(t) = m_M(t)$ $(0 \le t \le T^*)$ is also a route of K (one can stop at any desired moment); (K4) concatenation.

We implemented controlled motion in Riemann surfaces, Moebius band, projective plane and topological torus.

4.5. Experimental mathematics [7]. On one hand, it is using well-known software (Mathematica, MathLab, MathCad), on the other hand a search for new mathematical facts (hypothesis) - a separate direction of investigations.

4.6. Training in deciphering the simplest ciphers alongside with evaluating the knowledge of a language [1].

4.7. Complex examination (for example, [12]) including multimedia tasks, interactive tasks of optimization and solving equations, tasks with objects with-out. Primary versions of such software for Kyrgyz language, mathematics and informatics were implemented and are in use.

4.8. Measuring imagery [13]. **Definition**. The problem is said to be intellectual eye measurer (or measuring imagery) - its conditions are strict but the approximate answer is permissible; using any tool (computer, pen-and-paper, reference book) is forbidden; in sciences the time given to answer is about 20 - 30 seconds to avoid immediate mental counting. If the answer is a real number then $Q(x) = |x - x_0|$ or $Q(x)=|log(x/x_0)|$ (for $x_0 >0$) where x_0 is the exact answer.

We have introduced competitions on students' capacities in this subject matter.

4.9. Necessarily-collective tasks [14]. For example such task includes: transformation of sign systems: the first teammate is given a drawing (a set of similar drawings); s/he describes it in a prescribed language (during 15-20 minutes) and this text is sent to the second teammate by an intermediary; s/he restores the drawing (the consequence of drawings) (during 10-15 minutes).

4.10. Software to correct pronunciation.

4.11. Independent interactive presentation of objects. If a computer presentation does not depend on the user's knowledge and skills on similar objects then it is said to be independent.

4.11a. Interactive presentation of some mathematical objects [8].

4.11b. Interactive presentation of basic of language. Earlier, learning a living language was implemented with the assistance (including bilingual dictionaries and text-books) of persons who had a complete command of it; investigating of a dead language was done by means of remained bilingual texts and texts with additional implicit suggestions and conclusions. 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.

Before our publications, existed software to learn languages were based on languages native to the user.

We proposed [3; 9; 10; 11] **Definition**. Let any "notion" (word of a language) be given. If an algorithm acting at a computer: - performs (generating randomly) sufficiently large amount of situations 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".

Simple mathematical models consist of fixed (F_i) and movable (M_j) sets and temporal sequence of conditions of types $(M_j \subset F_i)$, $(M_j \cap F_i = \varnothing)$, $(M_j \cap F_i \neq \varnothing)$.

Remark. 4.10) can also be involved.

Sketches of such software were implemented for Kyrgyz, English and Turkish languages. A proposal for Chinese language was in.

5. Conclusion

We hope that developing this method would yield new types of educational software both interesting and useful for students. For instance, combination of 4.3) and 4.11a) can give independent presentation of some physical notions; adding of mathematical tasks with physical content can give a complex examination in physics.

REFERENCES

1. Борубаев А.А., Панков П.С. Дискретная математика (допущено МОН КР в качестве учебного пособия для преподавателей вузов). – Бишкек: изд. КРСУ, 2010. – 123 с.

2. Борубаев А.А., Панков П.С. Компьютерное представление кинематических топологических пространств. – Бишкек: КГНУ, 1999. – 131 с.

3. Карабаева С. Единый алгоритм словоизменения и представление про-странства в кыргызском языке. – Saarbrűcken, Deutschland: Lap Lambert Aca-demic Publishing, 2016. – 62 с.

4. Панков П.С. Двойственность параметров управления и наблюдения при получении гарантированных оценок //

Проблемы теоретической кибернетики: Тез. докладов IX всесоюзной конференции (сентябрь 1990 года). – Волгоград, 1990. – Часть 1(2), с. 57.

5. Панков П.С. Проект автоматизации ввода дискретной информации с помощью биотоков, считываемых с мышц человеческих рук // Проблемы автоматики и управления: Научно-технический журнал / НАН КР. – Бишкек: Илим, 2005. – С. 125-127.

6. Панков П., Баячорова Б., Жураев М. Кыргыз тилин компьютерде чагылдыруу. – Бишкек: Турар, 2010. – 172 б.

7. Grenander U. Mathematical experiments on the computer. - New York: Academic Press, 1982. – 525 p.

8. Pankova M. Mathematical models of notions of Chinese language // Abstracts. Issyk-Kul International Mathematical Forum, Bozteri, Kyrgyzstan, 2015. – P. 80.

9. Pankov P.S., Aidaraliyeva J. Sh., Lopatkin V. S. Active English on computer // Conference «Improving Content and Approach in the Teaching of English Language in the Context of Educational Reform», Bishkek, 1996. – pp. 25-27.

10. Pankov P.S., Alimbay E. 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, Poland, 2005. – Pp. 357-360.

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

12. Pankov P., Dolmatova P. 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, 2009, Poznan, Poland. – P. 502-506.

13. Pankov P.S. Independent learning for Open society // Collection of papers as results of seminars conducted within the frames of the program «High Education Support». Bishkek: Foundation «Soros-Kyrgyzstan», 1996. - Issue 3, pp. 27-38.

14. Pankov P.S. Necessarily-Collective Computer Competitions for Schoolchildren // Information Technologies at Schools: Proceedings of the Second International Conference on Informatics in Secondary Schools "Evolution and Perspectives", 2006, Vilnius, Lithuania. – Pp. 585-588.

Khamroeva Sh., Mengliev B.
Tashkent State University of Uzbek Language and Literature named after A.Navai, Uzbekistan, Tashkent

MODELING AFFIXES FOR THE MORPHOLOGICAL ANALYZER OF THE UZBEK LANGUAGE

Abstract. The article discusses the issues of modeling the position of morphemes in word forms for the morphological analyzer of the Uzbek language. Formative affixes of verbs have different meanings, which are associated with their positions in the word form. The order and sequence in the placement of grammemes are related to its meaning and grammatical feature: the derivative that forms a new lexical meaning is added first, the formative affix is after the derivative, and the inflectional affix, which does not affect the lexical meaning, but connects the word, has a postposition. The study of formative morphemes shows that they are added to the stem or root in a specific order and form a specific sector. Since morphemes are found in all parts of speech, it is necessary to take into account the positions of formative morphemes in each part of speech. This article discusses the issue of modeling the position of formative affixes in verbs. By modeling the positions of morphemes in word forms, one can eliminate the grammatical homonymy that is formed in speech.

Keywords: *modeling, morphological analyzer, Uzbek language, arrangement of affixes, position of affixes.*

Хамроева Ш., Менглиев Б.
Ташкентский государственный университет узбекского языка и литературы им. А.Навои, Узбекистан, Ташкент

МОДЕЛИРОВАНИЕ АФФИКСОВ ДЛЯ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА УЗБЕКСКОГО ЯЗЫКА

Аннотация. В статье рассматриваются вопросы моделирования позиции морфем в словоформах для морфологического анализатора узбекского языка. Формообразующие аффиксы глаголов имеют разные значения, которые связаны с их позициями в словоформе. Порядок и

последовательность в размещении граммем связаны с ее значением и грамматической особенностью: дериватема, которая формирует новое лексическое значение, добавляется первым, формообразующий аффикс стоит после дериватемы, а словоизменительный аффикс, который не влияет на лексическое значение, но связывает слово, имеет постпозицию. Изучения формообразовательных морфем показывает, что они добавляются к основе или корню в определенном порядке и образуют определенный сектор. Поскольку формообразования встречаются во всех частях речи, необходимо учитывать позиции формообразовательных морфем в каждой части речи. В данной статьи рассматривается вопрос моделирования позиции формообразовательных аффиксов в глаголах. Моделируя позиции морфем в словоформах, можно устранить грамматическую омонимию, которая образуется в речи.

Ключевые слова: *моделирование, морфологический анализатор, узбекский язык, аранжировка аффиксов, позиция аффиксов.*

Введение. Об окружении формообразующих морфем в словоформах. В обычном порядке аффиксов в слове формообразующая морфема стоит после словообразующей. Об этом подробно изложено в работах А.Ходжиева, С.Усманова [7: 46]. Объясняется это тем, что словообразовательная морфема связана с материальной стороной слова, а формообразующая морфема связана с формальной стороной слова. Формообразующие морфемы представляют грамматическое значение, также влияют на лексическое значение слова. Они формируют грамматическое значение слово связанное, с лексическим значением основы.

Основная часть. Порядок грамматических элементов в структуре слова имеет определенную закономерность. Граммемы размещаются в следующем порядке: основа (корень) + словообразовательный аффикс + формообразовательный аффикс + словоизменительный аффикс. Порядок и последовательность в размещении граммем связаны с его значением и грамматической особенностью: дериватема, которая формирует новое лексическое значение, добавляется

первой, формообразующий аффикс стоит после дериватемы, а словоизменительный аффикс, который не влияет на лексическое значение, но связывает слово, имеет постпозицию. Нормативное положение аффиксов иногда нарушается. Например: *опа-лар-им // опа-м-лар, айт-ди-нг-лар // айт-ди-лар-инг.* Но это редкое явление [3: 134].

Порядок присоединения формообразовательных и словоизменительных аффиксов к глагольным основам имеет множество особенностей. В этом статье мы рассматриваем особенности позиций формообразующих аффиксов в словоформах. Наши наблюдения за формообразовательными аффиксами показали, что они добавляются к основе или корню в определенном порядке и образуют определенный сектор [8: 104-105]. Поскольку формообразования встречаются во всех частях речи, необходимо учитывать позиции формообразовательных аффиксов в каждой части речи. В данной статье будем рассматривать формообразовательные аффиксы в только в глаголах.

Локализация формообразовательных аффиксов в глаголах. В ряд формообразовательных аффиксов глагола входят аффиксы залога, причастия, деепричастия, масдара, синтетического типа глаголов, аффиксы указывающие степени и продолжительности действия [5: 18-20]. Отдельные вопросы формообразовательных аффиксов в глаголах отражены в исследованиях Н.Баскакова, А.Щербака, А.Гулямова, А.Хожиева, также в работах европейских языковедов Ж.Вандриеса и Г.Глисона. В узбекском языкознании формообразующие морфемы изучались в монографическом плане в работах Т.У.Мирзакулова [2]. Поэтому не останавливаясь на обычных порядках формообразующих морфем, перейдем к выводам о позиционных правилах, основанных на наших наблюдениях. В глаголах (V) наблюдается следующий порядок локализации морфем: V + словообразовательный аффикс + формообразующий аффикс. Состав формообразовательных афиксов достаточно изучен [6: 115-124]. В некоторых случаях наблюдается добавление понудительного залога (каузатив) к основе глагола дважды:

том+из+дир, кий+дир+тир, кий+гиз+тир, ол+дир+тир, кел+тир+тир. Такие случаи должны быть описаны в базе правил дополнительно к существующей модели $V+f_1^v+p_2^v$, (глагол + формообразовательный аффикс + словоизменительный аффикс). В вышерассмотренном порядке морфема *-дир* может образовывать грамматическую омонимию, находясь в постпозиции. Такая омонимия отличается от модели $V+f_1^v+p_2^v$.

В узбекском языке формы страдательного и возвратного залога схожи, и их автоматическое распознавание не может быть проведена стандартной моделью, поскольку в обоих случаях их локализация и форма аффикса одно и то же. Отметим, что употребление страдательного или возвратного залогов зависит от самой основы. Сравните:

1) значение возвратного залога: *туғилди, кўринди, сўкинди, ютинди, танилди*и т.п.;

2) значение страдательного залога: *тешилди, қорилди, тикилди (тикмоқ, қарамоқ эмас)* и т.п.;

3) значение возвратного и страдательного залога (полисемантическая форма): *қайрилди, айрилди, бўғилди, отилди, йиғилди* и т.п.

В своих опубликованных работах мы описали регулярные и не регулярные морфемы, которые могут присоединяться ко всем глагольным основам или только к определенным. Морфемы возвратного и страдательного залога не могут присоединяться подряд ко всем основам. Выше мы показали на примерах, что эти морфемы привязываются к определенным основам. Для создания лингвистического обеспечения (базы данных) морфоанализатора узбекского языка требуется инвентаризация всех глаголов узбекского языка на предмет возможности прибавления этих форм. Определение типов залоговых морфемы осуществляется на основе инвентаризации, а не на основе модели автоматического анализа.

Формообразующая морфема глагола *-(и)ш* может выступать как морфема масдара и взаимно-совместного залога (в некоторых случаях форма возвратного залога). Чтобы различать такие грамматические омонимы и правильно анализировать их

при морфоанализе, в базу данных должны быть включены следующие позиционные модели:

1) в позиции $V+f_1^v$ или $V+f_1^v+f_2^v$ морфема *-(и)ш* означает морфему масдара: *чиқ+ар+иш; том+из+иш, кий+гиз+иш, кий+дир+иш, тасдиқла+т+иш, суя+н+иш, чўз+ил+иш*;

2) в позиции $V+f_1^v+p_1^v$или $V+f_1^v+f_2^{v+}p_1^v$ форма *-(и)ш* означает морфему взаимно-совместного залога: кел+иш+ди, суриштир+иш+ди, ўтир+иш+ди, кел+тир+иш+ди. Существует еще одна позиция грамматической омонимии, которая должна представляться другой моделью. Например: глагол + взаимно-совместный залог + взаимно-совместный залог: *сўзла+ш+иш.* В этой позиции форму *–ш* следует рассматривать как форму взаимно-совместного залога.

3) в словах *жойлашди, керишди, қоришди* форма *–ш* дает значение возвратного залога. Это связано с проблемами переходных/непереходных глаголов (конечно, эта проблема теоретически ожидает своего решения) [3: 142]. Поэтому эту ситуацию следует определять на основе правил исключений, а не на основе моделирования.

Формы масдара, причастия и деепричастия имеют положительные и отрицательных формы. Положительная форма представляется нулевой формой, отрицательная форма реализуется в деепричастиях аффиксами *-май/масдан (ўқиб – ўқимасдан*), в причастиях — аффиксом *-ма (ўқиган-ўқимаган)*, в масдарах — аффиксом *-маслик (ўқиш – ўқимаслик)*. Эти формы представляют собой сложные суффиксы, созданные с помощью *–ма* [3: 148]. Эти сложные формы образуют омонимы с другими формами, выражающими грамматическое значение, поэтому мы фильтруем их путем моделирования. Необходимо различать аффикс деепричастия *–масдан,* означающий отрицательную форму деепричастия от формы причастия. Например, в таких позициях как в предложениях *ишёқ**масдан** ҳамма безор, борсакел**масдан** ҳеч ким қайтмаган* аффиксы *–мас и –дан* − это отдельные морфемы, два аффикса стоящие рядом (т.е. причастие в форме исходного падежа). А в предложениях *Ишни тугат**масдан** кетмайди; дарсни бажар**масдан** келмайди – масдан — эт*о одна сложная морфема, не разлагаемая на

составляющие аффиксы. В таких случаях формальный автоматический морфоанализ не может дать верного результата, необходимо привлекать семантический или синтаксический анализ: словоформы могут различаться в зависимости от синтаксической функции в предложения. Вопросы разграничения позиций грамматической многозначности и омонимии с семантическо-синтаксическим фильтром в данной исследовательской задачи не рассматриваются. Проблема требует специального исследования в рамках проблемы создания автоматического семантического анализатора.

Отрицательная форма глагола в некоторых случаях представляется в виде: *кел+моқчи+мас+ман, бор+моқчи+мас+ман, суя+н+моқда+мас+сан*. В таких случаях форма *-мас* рассматривается как сокращенный вариант неполного глагола *эмас*. Сравните: *кел+моқчи+мас+ман = кел+моқчи+эмас+ман*. Этот случай можно представить в виде модели: $V+ p_1^{v}+f_1^{v+}p_2^{v}$.

Фонетическая форма (алломорф) глагола − формант *-вер,* является алломорф вспомогательного глагола *бер*. В таких случаях будет изменение в орфографии форманта. Например: *кел+а+вер+а+ди = кел+а бер+а+ди*. Такие изменения можно описать с помощью модели $V+f_1^{v}+V+f_2^{v}+p_1^{v}$.

Правила: если будет {глагол-основа+ f_1^{v}+вер+$V+f_2^{v}+p_1^{v}$}, то форма *-вер* = вспомогательный глагол. В зависимости от положения формы деепричастие образует омонимию с формой прошедшего времени *-иб*. Анализируем позиции:

1) в слоформах *ўқиб, ёзиб, бориб* морфема *-иб* представляет грамматическое значение деепричастия; в синтетических формах глагола *ўқиб кўрмоқ, айтиб бермоқ, гуллаб қўймоқ* морфема *–иб* реализует связь между основным и вспомогательным глаголом; в составных глаголах *бориб келмоқ, олиб бормоқ, кириб чиқмоқ* морфема *–иб* функционирует как относительная форма, которая служит для соединения основного и вспомогательного глаголов. Следовательно,

а) если реализется модель $V+f_1^{v}$(и)б, то есть{основа-глагол+(и)б+0}, тогда —*иб*= форма деепричастия;

б) если реализуется модель $V_1+f_1^v$(и)б+$V_2+f_1^v+f_1^v/p_1^v$, то есть {осново-глагол+(и)б+ осново-глагол+f_1^v/p_1^v}, тогда форма *-иб*= связка основного и вспомогательного глагола;

2) в словоформах *ўтирибман, бўлибди, борибсан, узоқлашибмиз –иб* действует как аффикс прошедшего времени. Данная ситуация представляется моделью $V+f_1^v$(и)б+p_1^v, то есть{глагол-основа+(и)б+словоизменительный аффикс}, тогда форма -иб = прошедшее время глагола.

Морфема второго лицо *-(и)нгиз* означает и формы повелительного наклонения [1: 30-34]: *чиқмангиз*. В устной речи эта морфема употребляется в формах *-(и)нглар, -(и)нгизлар*. Например: *кел+ингиз, кел+ингиз+лар*. Эти формы необходимо записывать в лингвистическое обеспечение как альтернативу (эквивалент), потому что художественная литература, материал устной речи (даже диалекта) может подвергаться морфологическому анализу для различных целей пользователя. В этом случае полезно в качестве альтернативы определить типичные случаи отклонения от нормы литературной речи, чтобы анализатор смог правильно различать грамматическое значение. Модель можно представить как: $V+f_1^v+p_1^v$*-(и)нгиз* = $V+f_1^v+p_1^v$ *-(и)нглар/ -(и)нгизлар.*

Морфема деепричастия *-гани* образует омонимичную позицию с формой причестия *–ган с* притяжательным суффиксом. Сравните: *Ишла**гани** келишди; айт**гани** борганди* (деепричастие цели) // *Ишла**гани** йўқ, ишла**гани**нг йўқ; айт**гани**ни қилди, айт**гани**нгизда билардик*(форма причастие с прятежательности). При моделировании этих ситуаций морфоанализатор выполняет безошибочный анализ:

1) если будеть $V+f_1^v$ (гани) + 0, то есть {основа-глагол+гани+0}, тогда f_1^v =форма деепричастие цели;

2) если будеть $V+f_1^v$(ган+и)+p_1^v+0, то есть {глагольная основа+(ган+и)}, тогда, f_1^v = (ган+и), то есть форма причастия с аффиксом прятежательности;

3) если будеть $V+f_1^v$(гани)+йўқ, то есть {глагольная основа+(ган+и)+йўқ}, тогда f_1^v = форма причастия с аффиксом прятежательности;

4) если $V+f_1^v$(гани)+ p_1^n(-га/ -да/ -дан), то есть {глагольная основа+(ган+и)+ (-га/ -да/ -дан}, тогда f_1^v = форма причастия с аффиксом прятежательности.

Частица -*дир* обычно ставится в самом конце цепочки глагольных словоформ: *боргандир, келмагандир.*

Аффикс сказуемости -*дир* тоже находится в том же положении. Следующая модель будет достаточной, чтобы различать эти двух омонимичных аффиксов:

1) если $V+f_1^v+f_2^v$(дир), то есть {глагольная основа +(формообразующий аффикс+дир}, тогда f_2^v =частица;

2) если N/Adj/P/Num+p_1^n(дир), то есть {основа (имя существительное /имя прилагательное/имя числительное/местоимение) + дир}, тогда p_1^n = аффикс сказуемости.

Следующие позиции можно твердо определить как общий нормальный порядок для морфем, образующих форму глаголов:

1) в первой позиции после основы стоит форма залога, порядок которого следующий: основа+возвратный залог+понудительный залог +взаимно-совместный залог: *безантиришди, ювинтиришди, безантирилди, ювинтирилди, то есть* $V+f_1^{v(voice)}+f_2^{v(voice)}+f_2^{v(voice)}$;

2) после формы залога идет отрицательная форма -*ма*: $V+f_1^v+f_2^{v(neg)}$;

3) после формы отрицания идут формы причастия, деепричастия и масдара (*бормаган, бормай, бормасдан, бормаслик*) и формы наклонения (*бормасин, бормайлик, бормай*): $V+f_1^v+f_2^{v(neg)}+ f_3^{v(fin)}$ ёки$+f_1^{v(mood)}$;

4) после формы отрицания также идут формы категории времени, категории числа (*бормасин, бормайлик, бормади*): $V+f_1^v+f_2^{v(neg)}+p_1^v+ p_2^v$.

Формообразующие морфемы, обозначающие интенсивность движения, должны быть включены в ряд квазиграммем. Они представлены в таблицах 1-2.

Таблица 1.

Морфемы, указывающее на слабый уровень движения

-мсира/-имсира:	йиғламсира, кулимсира
-қира/-инқира:	ишонқира, оқаринқира, ўчинқира
-ш/-иш:	тўлиш, қизиш, оқариш, тўхташ (юраги)
-қ/-иқ/-к/-ик:	тутақ, толиқ, шишиқ, жунжик, кўник

Таблица 2.

Морфемы, указывающее на сильный уровень движения

-ла/-ала:	қувла/қувала, ишқала, савала, сийпала, чўқила, чайқала, силтала, опичла
-қи/-ғи:	юлқи, сизғи, тўзғи, бижибижғи
-чи:	типирчила, тепчи, терчила
-а:	бур −бура, қувон −қувна, урин −урна
-ғила/-кила/-қила/ -ғила:	югургила, титкила, тепкила, чопқилла, турткила, тортқила, эзғила, чўзғила.
-га/-ка/-қа:	сурга, сурка, чайқа.

Такие морфемы относятся к квазиграммемам, потому что они не являются членами какой-либо парадигмы. Как видно из таблицы, большинство формообразующих морфем в глагольной словоформе, указывающих на интенсивность действия, по своей природе омонимичны. Чтобы различать их значение, мы моделируем их в соответствии с категорией, после которой они идут: $V+f_1$: *-ка* (*сурка*), *-а* (*бура*), *-ала* (*қайтала*), *-имсира* (*кулимсира*), *-инқира* (*босинқира*), *-кила* (*титкила*), *-чай* (*букчай*).

Если {V+f_1| *-ка, -а, -ала, -имсира, -инқира, -кила, -чай*}, тогда {f_1| *-ка, -а, -ала, -имсира, -инқира, -кила, -чай*} = формообразующие морфемы указывающие на интенсивность действия. С помощью этой модели становится ясно, что эти морфемы, добавленные после глагола, являются формообразующими аффиксами, указывающими на интенсивность действия.

Некоторые из этих аффиксов образуют омонимии с словообразующими аффиксами (*-ла, -чи, -а*), формообразующими аффиксами (*-ш,-иш*), словоизменительными морфемами (*-га, -ка, -қа*). Их можно отфильтровать по следующей модели: если {V+f_1 (*-ла, -чи, -а, -ш, -иш, -га, -ка, -қа*)}, то есть если эти формы добавлены после основы глагола, тогда f_1= формообразующая морфема, обозначающая интенсивность действия.

Известно, что в узбекском языке способ аффиксации при образовании глагола производится от неглагольной формы основы. Основываясь на этом правиле, мы моделируем позиции морфем следующим образом:

1) если {N/A /Num /Adv/ Sim +d_1 (*-ла, -чи, -а*)}, то есть, если эти морфемы добавлены после неглагольной основы, тогда d_1 = словообразовательный аффикс (поскольку выше было указано, что эти суффиксы являются формообразующими морфемами при добавлении после глагольной основы);

2) если {N/A /Num /Adv/ Sim /причастие, деепричастие, масдар+r_1*-га, -ка, -қа*}, то есть, если эти формы добавлены после неглагольной основы, тогда r_1 = словоизменительный аффикс (поскольку выше было указано, что эти суффиксы являются формообразующими морфемами при добавлении после глагольной основы).

Заключение. Констатируем, что анализ и предложенное моделирование сочетаний формообразующих морфем позволяет морфоанализатору различать их от словообразовательных морфем и морфем в синтаксических формообразованиях. На основе 5 инвариантных моделей нами предложены 20 специфичных моделей для анализа морфемы глаголов с целью

определения позиции формообразующих морфем для морфоанализатора.

ЛИТЕРАТУРА

1. Аламова М. О повелительно-желательное наклонение на тюркских языках / Узбекский язык и литература, 1966. № 6. - С.30-34.(Аъламова М. Туркий тилларда буйруқ-истак майли ҳақида / Ўзбек тили ва адабиёти, 1966. № 6. – Б.30-34).
2. Мирзакулов Т.У. Формаобразующие сложные аффиксы в узбекском языке: автореф. дисс. ... канд. филол. наук. – Тошкент, 1980.
3. Менглиев Б., Холиёров О., Абдурахманова Н. Универсальный учебник узбекского языка. (Перераб. 3-е издание). – Ташкент: Академнашр, 2014. - 389 с. (Mengliyev B., Xoliyorov O`., Abdurahmonova N. O`zbek tilidan universal qo`llanma. (Qayta ishlangan 3-nashri). – Toshkent: Akademnashr, 2014. – 389 b).
4. Нариманова М.Д. О некоторых особенностях глаголов в современном узбекском языке. Сборник научных трудов. Вопросы узбекского языкознания. ТошДУ, Ташкент, 1983. - Б. 52-59. (Нариманова М.Д. Ҳозирги ўзбек тилида феъллардаги даража ясовчиларнинг баъзи бир хусусиятлари ҳақида. Илмий ишлар тўплами. Ўзбек тилшунослиги масалалари. ТошДУ, Тошкент, 1983. – Б. 52-59).
5. Нематов Х., Гуломов А., Кадыров М., Абдураимова М. Родной язык: учебник для 6 класса. - Т .: Учитель, 2004. (Ne’matov H., G‘ulomov A., Qodirov M., Abduraimova M. Ona tili: 6- sinf uchun darslik. – T.: O‘qituvchi, 2004).
6. Тоджиев Ё. Синонимия аффиксов, образующих глагольные уровни. Научная работа ТашГУ. Вып. 443. Вопросы узбекского языкознания. – Тошкент, 1973. – Б. 115-124. (Тожиев Ё. Феъл даражаларини ясовчи аффикслар синонимияси. Научные труды ТашГУ. Вып. 443. Вопросы узбекского языкознания. – Тошкент, 1973. – Б. 115-124).

7. Усмонов С. Морфологическая структура слов современного узбекского языка // Научные труды ТашДПИ им. Низами, 42 тома, 2 кн. – Ташкент, 1963. (Усмонов С. Ҳозирги ўзбек тилида сўзнинг морфологик тузилиши // Низомий номидаги ТошДПИ илмий асарлари, 42 том, 2-китоб. – Тошкент, 1963).

8. Грамматика узбекского языка. Морфология. Iтом.– Ташкент, 1975. – 608 б. (Ўзбек тили грамматикаси. Морфология. I том. – Тошкент, 1975. – 608 б).

9. Гуломов Ю.Г. Аффиксы уровня глагола в узбекских диалектах. Сборник научных трудов. Вопросы узбекского языкознания. ТошДУ – Ташкент, 1983. – С.64-68. (Ғуломов Ё.Г. Ўзбек шеваларида феъл даражалари аффикслари. Илмий ишлар тўплами. Ўзбек тилшунослиги масалалари. ТошДУ – Тошкент, 1983. – Б.64-68).

10.Ходжиев А. Формообразующие аффиксы. – Узбекский язык и литература, 1977. (Ҳожиев А. Форма ясовчи аффикслар. – Ўзбек тили ва адабиёти, 1977).

11.Ходжиев А. Формообразование на современном узбекском языке. – Ташкент, 1979. (Ҳожиев А. Ҳозирги ўзбек тилида форма ясалиши. − Ташкент, 1979).

12.Современный узбекский литературный язык. – Ташкент, 1980. (Ҳозирги ўзбек адабий тили. – Тошкент, 1980).

Khakimov M. Kh.
National University of Uzbekistan named after Mirzo Ulugbek, Uzbekistan, Tashkent
Kadirov B.
Karshi State University, Uzbekistan, Karshi

ALGORITHMS FOR ANALYZING ENGLISH TEXTS GENERATED BY WITH THE USE OF AN EXTENSIBLE INPUT LANGUAGE

Abstract. Prominent feature at the computer analysis of offers stated in a natural language, is logic and linguistic communications between parts of offers. It is possible to show these communications in a general view at construction of formal logiko-linguistic models and it is much more accurate by working out of mathematical models of a natural language. In both variants construction of models the expanded source language is used. In the present work analysis algorithms of English scientific texts are stated.

Keywords: *English language, semantics, logic-linguistic communication, logic-linguistic model, the mathematical model, the expanded source language, algorithm.*

Хакимов М.Х.
Национальный университет Узбекистана им. Мирзо Улугбека, Узбекистан, Ташкент
Кадиров Б.
Каршинский государственный университет Узбекистан, Карши

АЛГОРИТМЫ АНАЛИЗА АНГЛИЙСКИХ ТЕКСТОВ, СФОРМИРОВАННЫХ С ПРИМЕНЕНИЕМ РАСШИРЯЕМОГО ВХОДНОГО ЯЗЫКА

Аннотация. Характерной особенностью при компьютерном анализе предложений, изложенных на естественном языке, являются логические и лингвистические связи между частями предложений. Эти связи можно показать в общем виде при

построении формальных логико-лингвистических моделей и намного четко при разработке математических моделей естественного языка. В обоих вариантах построения моделей используется расширяемый входной язык. В настоящей работе изложены алгоритмы анализа английских научных текстов.

Ключевые слова: *английский язык, семантика, логико-лингвистическая связь, логико-лингвистическая модель, математическая модель, расширяемый входной язык, алгоритм.*

1. Введение

При анализе предложений, составленных на естественном языке, основное внимание уделяется на синтаксические и семантические связи между его частями, для которых семантические базы [1] имеют важнейшую роль. В настоящей статье приведены разработанные укрупненные алгоритмы для анализа научных текстов английского языка (АЯ), в соответствии с их логико-лингвистическим [2] и математическим моделям [3]. Метод анализа предложений, почти идентичен, который был использован при анализе частей предложений [5, 6].

При составлении алгоритмов были использованы терминальные символы из расширяемого входного языка [4]. Следует отметить, что расширяемый входной язык предназначен для обработки естественных языков и с момента опубликования он был расширен несколько раз, а последняя версия расширения не была опубликована.

Терминальные символы и их назначения:

EVX – входной текст для процесса анализа (например, текст на английском языке);

L1 – база слов из предметных областей;

CC – база существительных;

GG – база глаголов;

NN – база наречий;

PP – база прилагательных;

MM – база местоимений;

FF – база числительных;

Y – база союзов;
L – база модальных слов;
L(GG) – база модальных глаголов;
EE1 – временная таблица.

2. Укрупненные алгоритмы анализа научных предложений

Укрупненные алгоритмы предложений английского языка с их семантическими базами [1], приведенными выше обозначениями логико-лингвистическими [2] и математическими моделями [3], состоят из 12 вариантов.

Алгоритм N1.

1. Выбрать существительному из EVX соответствующее ему существительное из базы CC.
2. Выбрать глаголу из EVX соответствующий ему глагол из базы GG.
3. Если имеется наречие в EVX, тогда выбрать соответствующее ему наречие из базы NN.
4. Записать выбранные слова в таблицу EE1.

Алгоритм N2.

1. Выбрать артиклю из EVX соответствующее ему слово из базы L1.
2. Если имеется прилагательное из EVX, тогда выбрать соответствующее ему прилагательное из базы PP.
3. Выбрать существительному из EVX соответствующее ему существительное из базы CC.
4. Если имеется глагол из EVX, тогда выбрать соответствующий ему глагол из базы GG.
5. Если имеется другое прилагательное из EVX, тогда выбрать соответствующее ему прилагательное из базы PP.
6. Записать выбранные слова в таблицу EE1.

Алгоритм N3.

1. Выбрать артиклю из EVX соответствующее ему слово из базы L1.
2. Если имеется прилагательное из EVX, выбрать соответствующее ему прилагательное из базы PP.

3. Выбрать существительному из EVX соответствующее ему существительное из базы CC.

4. Если имеется глагол из EVX, выбрать соответствующий ему глагол из базы GG.

5. Если имеется наречие в EVX, тогда выбрать соответствующее ему наречие из базы NN.

6. Записать выбранные слова в таблицу EE1.

Алгоритм N4.

1. Выбрать местоимению из EVX соответствующее ему местоимение из базы MM.

2. Выбрать глаголу из EVX соответствующий ему глагол из базы GG.

3. Если имеется существительное из EVX, тогда выбрать соответствующее ему существительное из базы CC.

4. Если имеется прилагательное из EVX, то выбрать соответствующее ему прилагательное из базы PP.

5. Записать выбранные слова в таблицу EE1.

Алгоритм N5.

1. Выбрать существительному из EVX соответствующее ему существительное из базы CC.

2. Выбрать глаголу из EVX соответствующий ему глагол из базы GG.

3. Если имеется местоимение из EVX, выбрать соответствующее ему местоимение из базы MM.

4. Если имеется существительное из EVX, тогда выбрать соответствующее ему существительное из базы CC.

5. Записать выбранные слова в таблицу EE1.

Алгоритм N6.

1. Выбрать существительному из EVX соответствующее ему существительное из базы CC.

2. Выбрать глаголу из EVX соответствующий ему глагол из базы GG.

3. Если имеется прилагательное из EVX, то выбрать соответствующее ему прилагательное из базы PP.

4. Если имеется существительное из EVX, тогда выбрать

соответствующее ему существительное из базы CC.

5. Если имеется наречие в EVX, тогда выбрать соответствующее ему наречие из базы NN.

6. Если имеется местоимение из EVX, тогда выбрать соответствующее ему местоимение из базы MM.

7. Если имеется другой глагол из EVX, тогда выбрать соответствующий ему глагол из базы GG.

8. Записать выбранные слова в таблицу EE1.

Алгоритм N7.

1. Выбрать местоимению из EVX соответствующее ему местоимение из базы MM.

2. Выбрать глаголу из EVX соответствующий ему глагол из базы GG.

3. Если имеется числительное из EVX, тогда выбрать соответствующее ему числительное из базы FF.

4. Если имеется существительное из EVX, тогда выбрать соответствующее ему существительное из базы CC.

5. Если имеется союз из EVX, тогда выбрать соответствующий ему союз из базы Y.

6. Если имеется прилагательное из EVX, то выбрать соответствующее ему прилагательное из базы PP.

7. Выбрать существительному из EVX соответствующее ему существительное из базы CC.

8. Записать выбранные слова в таблицу EE1.

Алгоритм N8.

1. Если имеется артикль из EVX, тогда выбрать соответствующее ему слово из базы данных L1.

2. Выбрать существительному из EVX соответствующее ему существительное из базы CC.

3. Выбрать глаголу из EVX соответствующий ему глагол из базы данных GG.

4. Если имеется прилагательное из EVX, то выбрать соответствующее ему прилагательное из базы PP.

5. Если имеется союз из EVX, тогда выбрать соответствующий ему союз из базы данных Y.

6. Если имеется другое прилагательное из EVX, то выбрать соответствующее ему прилагательное из базы PP.

7. Записать выбранные слова в таблицу EE1.

Алгоритм N9.

1. Если имеется местоимение из EVX, тогда выбрать соответствующее ему местоимение из базы MM.

2. Выбрать модальному глаголу из EVX соответствующий ему модальный глагол из базы L(GG).

3. Выбрать глаголу из EVX соответствующий ему глагол из базы GG.

4. Если имеется наречие из EVX, тогда выбрать соответствующее ему наречие из базы NN.

5. Записать выбранные слова в таблицу EE1.

Алгоритм N10.

1. Выбрать существительному из EVX соответствующее ему существительное из базы CC.

2. Выбрать модальному глаголу из EVX соответствующий ему модальный глагол из базы L(GG).

3. Выбрать глаголу из EVX соответствующий ему глагол из базы GG.

4. Выбрать наречию из EVX соответствующее ему наречие из базы NN.

5. Записать выбранные слова в таблицу EE1.

Алгоритм N11.

1. Выбрать модальному слову из EVX соответствующее ему модальное слово из базы данных L.

2. Выбрать местоимению из EVX соответствующее ему местоимение из базы MM.

3. Выбрать глаголу из EVX соответствующий ему глагол из базы GG.

4. Если имеется союз из EVX, тогда выбрать соответствующий ему союз из базы данных Y.

5. Если имеется местоимение из EVX, то выбрать соответствующее ему местоимение из базы MM.

6. Записать выбранные слова в таблицу EE1.

Алгоритм N12.

1. Выбрать предлогу из EVX соответствующий ему предлог из базы данных D(E).
2. Выбрать местоимению из EVX соответствующее ему местоимение из базы MM.
3. Выбрать существительному из EVX соответствующее ему существительное из базы CC.
4. Выбрать глаголу из EVX соответствующий ему глагол из базы GG.
5. Если имеется числительное из EVX, тогда выбрать соответствующее ему числительное из базы FF.
6. Выбрать существительному из EVX соответствующее ему существительное из базы CC.
7. Записать выбранные слова в таблицу EE1.

Заключение

Каждый из изложенных алгоритмов может быть использован при анализе входного текста изложенного на английском языке. Так как предложения английского языка могут иметь соответствующие разновидности логико-лингвистических и математических моделей, поэтому были предложены различные алгоритмы.

Следует подчеркнуть, что каждый естественный язык имеет динамический характер развития, то и весь процесс обработки также имеет динамический характер. Отсюда вытекает, что любая часть процесса обработки естественного языка может претерпеть изменения при обработке на компьютере.

ЛИТЕРАТУРА

1. Абдурахмонова Н., Хакимов М.Х. Семантические базы английского языка для многоязычной ситуации компьютерного перевода // Труды научной конференции "Проблемы современной математики" 22-23 апреля 2011 г, г. Карши, с.311-314.

2. Абдурахмонова Н., Хакимов М.Х. Логико-лингвистические модели слов и предложений английского языка для многоязыковых ситуаций компьютерного перевода // Компьютерная обработка тюркских языков. Первая международная конференция: Труды - Астана: ЕНУ им. Л.Н. Гумилева, 2013, с. 297-302.

3. Хакимов М.Х. Математические модели слов и предложений по типам английского языка для системы машинного перевода // Проблемы вычислительной и прикладной математики, 2017, № 5, с. 50-55.

4. Хакимов М.Х. Расширяемый входной язык математического моделирования естественного языка для многоязычной ситуации машинного перевода. ЎзМУ хабарлари, №1, 2009, с. 75-80.

5. Хакимов М.Х., Кадиров Б. Инглиз тилидаги от, феълва сон сўзларни келтириб чиқарувчи математик моделнинг алгоритмлари // Informatika, axborot texnologiyalari va boshqaruv tizimi: bugun va kelajakda / Respublika ilmiy-amaliy konferensiyasi materiallari TO'PLAMI, 2018, b. 36-40.

6. Хакимов М.Х., Кадиров Б. Инглиз тилидаги сифат, олмош ва равиш сўзларни келтириб чиқарувчи математик моделнинг алгоритмлари // Informatika, axborot texnologiyalari va boshqaruv tizimi: bugun va kelajakda / Respublika ilmiy-amaliy konferensiyasi materiallari TO'PLAMI, 2018, b. 40-46.

Khamroeva Sh.
Tashkent State University of Uzbek Language and Literature named after A.Navai, Uzbekistan, Tashkent

ARRANGING MORPHEMES FOR THE DATABASE OF THE MORPHOLOGICAL ANALYZER OF THE UZBEK LANGUAGE

Abstract. The article discusses the issues of modeling the position of morphemes in word forms for the morphological analyzer of the Uzbek language. Formative affixes of different parts of speech have different meanings associated with their positions. The order and sequence in the placement of grammeme are related to its meaning and grammatical peculiarity. The study of formative morphemes shows that they are added to the stem or root in a specific order and form a specific sector. Since shapers are found in all parts of speech, it is necessary to take into account the positions of shapers in each part of speech. This article discusses the issue of modeling the position of formative affixes for nouns, adjectives, numerals and pronouns. By modeling the positions of morphemes in word forms, one can eliminate the grammatical homonymy that is formed in speech.

Keywords: *modeling, morphological analyzer, Uzbek language, arrangement of affixes, position of affixes.*

Хамроева Ш.
Ташкентский государственный университет узбекского языка и литературы им. А.Навои, Узбекистан, Ташкент

АРАНЖИРОВКА МОРФЕМ ДЛЯ БАЗЫ ДАННЫХ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА УЗБЕКСКОГО ЯЗЫКА

Аннотация. В статье рассматриваются вопросы моделирования позиции морфем в словоформах для морфологического анализатора узбекского языка. Формообразующие аффиксы разных частей речи имеют разные

значения, связанные с их позициями. Порядок и последовательность в размещении граммамем связаны с его значением и грамматической особенностью. Изучение формообразовательных морфем показывает, что они добавляются к основе или корню в определенном порядке и образуют определенный сектор. Поскольку формообразователи встречаются во всех частях речи, необходимо учитывать позиции формообразователей в каждой части речи. В этой статье рассматривается вопрос моделирования позиции формообразовательных аффиксов именных частей речи: существительных, прилагательных, числительных и местоимений. Моделируя позиции морфем в словоформах, можно устранить грамматическую омонимию, которая образуется в речи.

Ключевые слова: *моделирование, морфологический анализатор, узбекский язык, аранжировка аффиксов, позиция аффиксов.*

Введение

Формообразующие аффиксы разных частей речи имеют разные значения, связанные с их позициями. Порядок и последовательность в размещении граммамем связаны с их значением и грамматической особенностью: дериватема, которая формирует новое лексическое значение, добавляется первой, формообразующий аффикс стоит после дериватемы, а словоизменительный аффикс, который не влияет на лексическое значение, но связывает слово, имеет постпозицию. Изучение формообразовательных морфем показывает, что они добавляются к основе или корню в определенном порядке и образуют определенный сектор. Поскольку формообразователи встречаются во всех частях речи, необходимо учитывать позиции формообразователей в каждой части речи.

Основная часть

1. Аранжировка формообразующих морфем имен существительных. Ряд формообразующих морфем имен существительных включает аффиксы оценочности, уменьшительно-ласкательные аффиксы, морфемы

обозначающие уважение. Форма, относящаяся к категории оценочности, — это средство, с помощью которого говорящий выражает в своей речи положительное или отрицательное отношение к объективному существу. По мнению лингвистов, в узбекском языке принято выражать значение оценочности (личного отношения), уменьшительно-ласкательности, уважения и дискриминации с помощью различных суффиксов или аффиксоидов. Например, суффиксы *-ча, -чак, -чоқ* в таких словах, как *қизча* (девочка), *йигитча* (мальчик), *келинчак* (невестка), *қўзичоқ* (ягненок) выражают грамматическое значение уменьшительности. В словах *болагина, укажон, холажон, акахон, Раҳимбой, Қўчқортой, Гулсинбиби, Раънохон, Мохирабону* аффиксоиды *-гина, -жон, -хон, -той, -бой, -биби, -бону* обычно выражают значение ласкательности, а иногда и значение дискриминации. Итак, необходимо разделять формы оценочности на две основные группы: формы ласкательности, уменьшительности (1) и формы уважение (2) [5: 241].

В узбекском языке существует третий тип формы оценочности. Он образуется путем размещения в словоформе суффикса *-лар* после суффикса притяжательности в терминах родства. Например: *опамлар, дадамлар, амакимлар, бобомлар.* В таких случаях значение “уважение” проявляется в морфеме *-лар*. Категория личного отношения не меняет лексического значения слова, за исключением его словообразовательной функции (например, *ўргимчак, қўғирчоқ, кўрпача, боғча*), но добавляет коннотативное значение к основному понимаемому значению, т.е. *стул* и *стулча* в основном одно и то же – *стул* [5: 245]. В морфоанализе узбекского языка формы оценочности определяются как квазиграммемы и помещаются в базу данных.

Обычный порядок формообразующих морфем имен существительных можно представить следующим образом [5: 134.]: морфема уменьшительности / морфема ласкательности + число. Менее распространенная форма уменьшительности и ласкательности имеет предпозицию от широко распространненных форм этой категории: *тойчоқча, тойлоқча*.

После основы в именах существительных прибавляются морфемы оценочности (уменьшительности, уважение,

ласкательности) [2] – f_1^n, потом прибавляется множественная форма [10] – f_2^n: *той+чоқ+лар*.

Часто морфемы оценочности (субъективные формы оценки) повторно используются в составе словоформ [4: 16]. В этом случае пассивная форма предшествует активной: $N+f_1^n+f_2^n$: *-чоқ+ча*: *тойчоқча*; *-чиқ+ча*: *қопчиқча*; *-а+ча: уяча*; *-ка+ча*: *йўлкача/йўлакча*; *-чак+ча*: *тугунчакча*. В структуре морфемы этих примеров размещение аффиксов в диахроническом плане наблюдается до синхронных аффиксов. Однако в слово *ўғилчавой* ($N+f_1^n+f_2^n$), используемое в устной речи, используется после активного суффикса *-вой* аффикс *-ча*. Этот обратный порядок объясняется тем, что форма *-вой/-бой* является аффиксоидом.

В связи с появлением нового лексического значения в словах *шолча, боғча (детский сад, не полисадник), кўрпача -ча* не считается формообразующей морфемой, то есть слово не делится на морфемы [5: 175].

Такие слова-исключения включаются в словарь исключений: *-ча* не участвует в анализе морфем. При анализе справа налево форма *-ча* может быть найдена как аффикс, но когда идет поиск корневой части, эти слова обнаруживаются в качестве основы в словаре исключений; анализ делает вывод, что эти слова равны одной лемме.

Морфемная структура слова имеет ряд особенностей: среди них самой значимой является аффиксальная омонимия. Морфемы имеющие одинаковую форму порождают некоторые закономерности, связанные с расположением суффиксов. Расположение формы *-гина* f_2 в слове *қўзичоқгиналаримиз* ($N+f_1^n+f_2^n+f_3^n+r_1$) отличается от порядка в слове *ёшларгина* ($N+f_1^n+f_2^n$). В первом примере *-гина* действует как формообразующая морфема (f), а во втором примере действует как частица: формообразующая морфема *–гина* занимает предпозицию от множественного числа и притяжательной формы, а форма частицы *(-гина*) занимает постпозицию. Сравните:

1) *дадагиналари* (*-гина* представляет формообразующую морфему ласкательности);

2) дадаларигина (*-гина* представляет форму частицы).

В словоформах *болагинага / болагагина; акагинаси / акасигина* также проявляется аффиксальная омонимия. В наборе правил морфоанализа эти позиции формы *-гина* представляются в виде следующей модели:

1) если {N + гина + множественное число+притяжателное}, тогда —*гина* = морфема ласкательности и уменьшительности;

2) если {N+ множественное число + притяжателное +гина}, тогда *-гина* = форма частицы.

Эти две рассматриваемые формы в речи различаются посредством интонации, но морфоанализатор не распознает эту функцию без правил. Ввод в базу данных вышеупомянутых моделей позволяет морфоанализатору различать значения этих форм.

Форма *-гина* (f_2^n) означает ласкательность при использовании после уменьшительной формы *-ча* в таких словах, как *укачагинам, қизчагинам,* потому что морфема, обозначающая ласкательность, добавляет дополнительный смысл к уменьшительной форме: смысл ласкательности прибавляется в уменьшительную форму [8: 65; 4: 18].

Среди формообразующих морфем имен существительных существуют морфемы, которые исторически состоят из двух аффиксов, таких как *-даги, -ники, -гача.* Сейчас они считаются сложными морфемами и образуют омонимы с формами *-да, -ни, -ги, -ча.* Однако морфоанализатор с анализом направления справа налево в целом находит и анализирует правильно эти формы: *-даги, -ники, -гача*, поскольку в базе данных есть комментарий, что эти формы квазиграммемы. Другая причина отсутствия ошибок в морфоанализе состоит в том, что формы *-ги, -ки, -ча* всегда предшествуют формам *-да, -ни, -га* и не приводят к позиционной омонимии.

Морфема *-лар* при выражении значения уважения занимает позицию после формы притяжательности. Сравните: *опамлар // опаларим.* Также в словоформе *опаларим –лар* означает обобщенность, в словоформе *опамлар* − специфичность. Мы предлагаем следующие модели, чтобы отличить это состояние в морфологическом анализе: заметим, что f = *-лар*:

1) если {R+f+r}, тогда f = множественное число;

2) если {R+r+f}, тогда f = значение “уважение”.

В значениях словоформ *борларинг // боринглар, ўзларинг // ўзинглар* есть стилистические различия: *борларинг, ўзларинг* означает дискриминацию [11]. Существует как функциональная, так и семантическая разница в положении морфем в словоформе *бизларники* и *бизникилар.* В первой словоформе морфема *–лар* представляет множественное число, и считается формообразующей морфемой; во второй словоформе морфема *−лар* считается как формообразующая, так и словоизменительная морфема. Словоформа *бизларники* − это местоимение, обозначающее принадлежность, оно представляет неопределенность, неопределенно указывает на человек или вещь; словоформа *бизникилар* – это местоимение, обозначающее принадлежность множества людей. Следовательно, это смысловое различие также должно быть морфо-стилистически отфильтровано.

Кроме того, в таких словах, как *болаларсиз // боласизлар* изменение позиции морфемы *-лар* влияет на значение и функцию. Очевидно наличие семантических различий в предложениях *болаларсиз яшаш зерикарли // боласизлар йиғилиши* [7, 9]. Функция этих морфем заключается в следующем:

1) означает отрицание ситуации (*болаларсиз яшамоқ* − қандай яшамоқ?) (жить без детей – как жить?): действует как словообразовательный аффикс;

1) отрицает существование человека: (*боласизлар − боласиз оила*) (бездетная − бездетная семья); действует как словообразующая морфема.

Отметим, что при таких сочетаниях аффиксов возможны и кардинальные различия в значениях: меняется даже части речи. Для таких случаев также следует разработать семантическую фильтрацию.

2. **Аранжировка формообразующих морфем имен прилагательных.** Существует типичный порядок формообразующих морфем имен прилагательных [5: 134]. Формообразующий аффикс имен прилагательных *–роқ* употребляется после аффиксов *-иш, -имтир*, он также

добавляется после суффикса, который образует причастие, деепричастие: *оқишроқ, кўкимтирроқ, совинқираганроқ, тортиниброқ.*

К основе имен прилагательных (группа А) сначала добавляется морфема субъективной оценки, либо аффикс обозначающий меньшинство $A + f_1^a$: *кўкимтир, оқчил, оқиш; яхшигина.* Затем добавляется форма сравнительной степени: $A+f_1^a+f_2^a$: *кўкимтирроқ, оқишроқ, сарғишроқ, кичкинагина.* Форма *-гина*, добавленная к основе имени прилагательного, означает усиление-ударение в отличие от существительного, поэтому анализатор должен различать это следующим образом: если {А+гина}, *-гина* = частица. Такая же процедура наблюдается и в наречиях: *астагина, кўпгина, тезгина.*

3. Аранжировка формообразующих морфем имен числительных. Существует обычный порядок формообразующих морфем имен числительных [5: 134]. Формообразующий аффикс *-ча*, который несет значение предположения, имеет постпозицию после элемента безударного *-та*, что означает приблизительность счета: например, *ўнтача (примерно десять)*.

Формообразующие аффиксы в основах числительных занимают следующие позиции:

1) безударная морфема *-та*, образующая числительное, находится в первой позиции: f_1^{num};

2) морфема *-дан*, обозначающая значение деления − $Num+ f_1^{num}+f_2^{num}$: *учтадан* (такое моделирование морфемы *-дан* связано с необходимостью отличать эту форму от формы исходного падежа);

3) форма *-ча*, образующая числительное со значением приблизительности − $Num + f_1^{num} + f_2^{num}$: *ўнтача, бештача* (в этой позиции форма *-ча*, которая представляет формообразующую морфему числительного приблизительного значения, отличается от формы словообразующего аффикса наречия и формы уменьшительности.

4. Аранжировка формообразующих морфем в местоимениях. Существует обычный порядок формообразующих морфем местоимений [5: 134]. Безударная морфема *-дир*, которая выражает значение подозрения, обычно добавляется как частица после всех видов суффиксов: *нима-лар-ни-дир (чего-то).* Аффикс *-дир*, который означает значение неточности, обычно стоит после основ-местоимений (Pr) и имеет функцию частицы, занимает постпозицию среди других суффиксов: $Pr+f_1+f_2+f_3^p$ – *кимларнидир, бизларникидир,*

улардандир. Лингвистическая модель формы *-дир* в такой позиции имеет следующий вид: если {Pr+f_1+f_2+дир}, то *-дир* = частица со значением неточности.

Вывод

Констатируем, что моделирование позиции формообразующих морфем для лингвистической базы данных морфоанализатора узбекского языка гарантирует распознавание многозначности и грамматической омонимии словообразовательных и словоизменительных морфем. Нами были разработаны 5 специфичных инвариантных моделей для имен существительных, 1 модель для имен прилагательных, 1 модель для числительных, 2 субмодели для местоимений.

ЛИТЕРАТУРА

1. Абузалова М.К. Субстанциал морфология, валентлик ва синтактик қурилма: Филол. фан. доктори (DSc) диссер. Автореферати. − Самарқанд, 2018. − 84 б.
2. Гулямов А.Г. О некоторых особенностях аффиксов – с уменьшительно-ласкательным значением в узбекском языке / Научные труды ТашГУ. Вып. 268. Языкознание и литературоведение. − Ташкент, 1964.
3. Журабаева М.К. К вопросу о категории уменьшительно-ласкательности в узбекском языке Труды аспирантов ТашГУ. Вып. 360. Литература и языкознание. − Ташкент, 1970.
4. Махкамов Н. Аффиксальный и лексико-аффиксальный плеоназм в узбекском языке: Автореф. дис. …канд. филол. наук. − Ташкент, 1983. – 24 с.
5. Mengliyev B., Xoliyorov O`., Abdurahmonova N. O`zbek tilidan universal qo`llanma. (Qayta ishlangan 3-nashri). – Toshkent: Akademnashr, 2014. – 389 b.
6. Ўзбек тили стилистикаси. − Ташкент, 1983. – 216 б.
7. Степанова М.Д. Теория валентности и валентный анализ. – М., 1973.
8. Қўнғуров Р. Субъектив баҳо формаларининг семантик ва стилистик хусусиятлари. − Ташкент, 1980. – 178 б.
9. Қўчкортоев И.К. Сўз маъноси ва унинг валентлиги. − Ташкент, 1977.
10. Ғуломов А. Ўзбек тилида кўплик категорияси. − Ташкент, 1943. −146 б.
11. Щербак А.М. Последовательность морфем в словоформе как предмет специального лингвистического исследования. − Вопросы языкознания, 1983. − № 3.

СЕКЦИЯ 5
ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В СОХРАНЕНИИ И ИЗУЧЕНИИ ТЮРКСКИХ ЯЗЫКОВ

Eşref Adali
Istanbul Technical University,
Turkey, Istanbul
Zhumadillaeva A.
L.N. Gumilyov Eurasian National University,
Kazakhstan, Nur-Sultan

COMPARISON OF LANGUAGES

Abstract. Although there are more than 4000 languages spoken in the world, the number of commonly used languagesis limited. Turkish ranks eighth in the list of common usage. Although there are no criteria developed for the scientific comparison of languages, we can reasonably use criteria such as intelligibility, sound harmony, energy saving, vocabulary, efficiency, clarity, regularity of grammatical structures. We would like to emphasize that the evaluation we will make in this paper is made not in terms of linguistics but from an engineer viewpoint.

Keywords: *Turkish language, means of communication, harmony of sounds, formants of sounds, vocabulary, syllabic structure of the Turkish language, linguistic statistics.*

Ешреф Адалы
Стамбульский технический университет,
Турция, Стамбул
Жумадиллаева А.
ЕНУ им. Л.Н. Гумилева, Казахстан, Нур-Султан

СРАВНЕНИЕ ЯЗЫКОВ

Аннотация. Хотя в мире говорят на более чем 4000 языках, количество часто используемых языков ограничено. Турецкий язык занимает восьмое место в списке употребления. Хотя не

существует критериев, разработанных для научного сравнения языков, мы можем разумно использовать такие критерии как понятность, гармония звуков, экономия энергии, словарный запас, эффективность, ясность, регулярность грамматических конструкций

Мы хотели бы подчеркнуть, что оценка, которую мы сделаем в этой статье, сделана не с точки зрения лингвистики, а с точки зрения инженера.

Ключевые слова: *турецкий язык, средства коммуникации, гармония звуков, форманты звуков, словарный запас, слоговая структура турецкого языка, лингвостатистика.*

1. UNDERSTANDABILITY

A method used to measure the communication fidelity of telephone lines can also be used to measure the u*nderstandability* of a language. The way of doing the experiment is as follows: Two people who do not see each other start talking over the phone. For example, Eren is on one end of the phone and Gözde is on the other. These people will be assumed to use proper and measured language. Eren reads a text given to him with a certain speed; Gözde on the other side writes what she heard. When Eren's reading is finished, Gözde begins to read a different text of the same size. This time Eren writes what he heard. When the experiment is over, the writings of both sides are compared with the proof texts. Naturally, there will be words that both listeners hear wrongly and therefore misspell them. The degree of u*nderstandability* can be calculated as follows [1]:

$$AD = 1 - \frac{YS_1 + YS_2}{M_1 + M_2}$$

In this equation,
AD: degree of u*nderstandability*
YS: Number of wrong
YS_1: The wrong number of the first listener
YS_2: The wrong number of the second listener
M: Number of words in the text read
M_1: Number of words in the text read by the first reader

M_2: the number of words in the text read by the second reader

This experiment is done between two different people and if the comprehensibility levels found at the end of each experiment are averaged, the u*nderstand ability* level of a language is measured. When the same experiment is performed similarly for different languages, the degree of u*nderstand ability* will be calculated for other languages. In order for the experiment to be scientific, it should be paid attention that the same telephone connection is used and that the education levels and speech smoothness of the selected people are approximately the same. U*nderstand ability* experiments are carried out in the communication media of the Turkish language. These experiments show that the u*nderstand ability* of Turkish is high.

2. EASY OF SAYING, NICE SOUNDING, ENERGY CONSERVATION

Various opinions are expressed by linguists about the ease of utterance of languages and the way they sound. Linguists make the eastern characterization of Italian for Turkish. It is a fact that these evaluations are subjective.

But

- Having Turkish vowel and consonant vocal harmony rules,
- Having regular stresses in words and sentences,
- The richness of vowel letters and the duration of which can be extended up to twice.

It can be said as a technical evaluation that its features make Turkish a pleasant-sounding language.

2.1. Vowel Harmony

The existence of phonetic rules in Turkish and related languages is known. Vowel harmony rules are actually a result of the nature of our vocal organs. Figure-1 and Table-1 are given in order to show the vowel harmony rules of Turkish. When Figure-1 and Table-1 are examined together, the following results can be easily drawn [1]:

The sound we make from the back of our mouth when our lips are straight and open is the "a" sound. After the "a" vowel, we can

say "a" without distorting the shape of our mouth and lips, and "ı" with a little change.

The sound we make from the middle of our mouth when our lips are straight and closed is the "ı" sound. After the "ı" vowel, we can say "ı" without distorting the shape of our mouth and lips, and "a" by changing it very little.

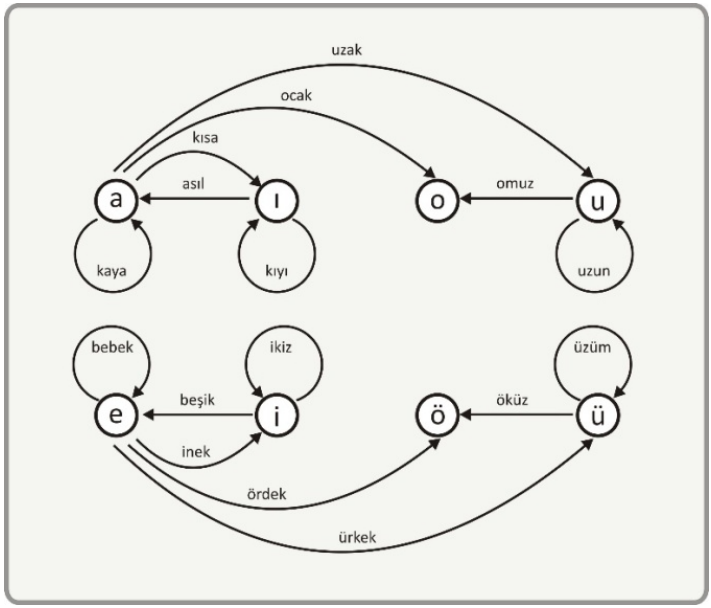


Figure-1: Vowelharmony of Turkish

Table 1.

Clustering of Vowel of Turkish According to Their Origination Places

	Front		Middle		Back	
	Flat	**Round**	**Flat**	**Round**	**Flat**	**Round**
Close	i	ü	ı			u
Open	e	ö			a	o

When our lips are round and open, we can make the "o" sound from the back of our mouths, after the "o" vowel we can make the "o" sound again. However, this situation is not encountered in Turkish words because it does not sound nice. The "u" and "a" sound can be easily produced with a small change in the lip structure.

- While our lips are round and closed, we can make the "u" sound from the back of our mouths, after the "u" vowel we can come back to the "u" vowel or we can make the "a" sound by changing our mouth shape a little.
- One of the sounds we make from the front of our mouth when our lips are straight is "e" and the other is "i". There is no need to change our mouth structure in order to remove the "e" after the "e". After the vowel "e", we can make a small change in the shape of our mouth and lips, making one of the sounds "i", "ö" and "ü". After the "i" vowel, the vowel "i" and "e" can be easily removed without making any changes. However, it is difficult to say the vowels "e" or "i" after the vowels "ö" and "ü".
- When our lips are round and open, we can make the "ö" sound from the front of our mouths, after the "ö" vowel, we can make the "ö" sound again. However, this situation is not encountered in Turkish words because it does not sound nice.
- When our lips are round and closed, we can make the "ü" sound from the front part of our mouth, followed by "ü" again.

By looking at the explanations above, we can generalize below:

- The consecutive vowel, which we can remove without any change in our mouth structure, is the same as the first vowel.
- Whatever the condition of our lips when singing the first vowel (open or closed), keeping the same when singing the consecutive vowel makes it easier to say the consecutive vowel.
- Whatever the condition of our lips (flat or round) when singing the first vowel, keeping the same when singing the consecutive vowel makes it easier to say the consecutive vowel.
- Switching from round vowels to straight vowels "o" and "u" can only be done if you stay in the back region. In this case, the transition can only be to the vowel "a".

- Transition from straight "e" to round vowels ("ü" and "ö"), provided that they stay in the same region (front). When we show the transition between vowels in the vowel's quadrant, the comments we made above are clearly visible, Figure-2 [1].

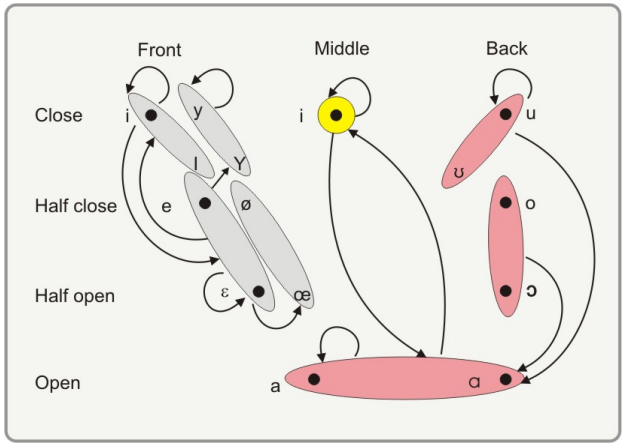


Figure-2: Transitions between vowels

Let's try to read these two artificial words to see the effect of celebrity harmony on utterance.

tenteredi, tintoridö

The first word is easier to say and less tiring on our jaws. Although this word is not a Turkish word, it is easy to say because the arrangement of the vowels is in accordance with Turkish sound rules.

In studies on vowels of Turkish, it has been tried to find the formants of vowel sounds. Naturally, a certain number of male and female subjects are used in these studies. The electrical signals of the subjects' speech are recorded, and then trying to find formants. There are studies by different researchers on this subject [2], [3], [4] and there are differences between the results obtained in these studies. According to the results of the study conducted by E. Malkoç, the

basic frequency (fo) and formant values of Turkish vowels are summarized in Table-2. If you pay attention to Table-2, the basic frequency and formant frequencies of male and female voices are different. In addition, the frequency values given in this table are the average values obtained from many subjects, not one person. Figure-3 shows the formant values of Turkish vowels together with the IPA's formant values. As seen in this comparative way, the vowels in the phonetic alphabet of IPA are close to the vowels of Turkish [1].

Table 2.

Basic Frequencies and Formants of Turkish Vowels

letter	gender	fo (Hz)	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)
a	female	236	771	1338	2998	4168
	male	130	642	1128	2714	3707
e	female	231	578	2205	2961	4128
	male	127	470	1866	2563	3715
ı	female	233	492	1629	2976	4232
	male	128	396	1500	2479	3782
İ	female	245	430	2591	3325	4308
	male	138	306	2111	2897	3751
o	female	243	564	959	2976	3794
	male	130	483	860	2733	3668
ö	female	212	543	1636	2764	3947
	male	124	469	1510	2439	3554
u	female	247	452	961	2940	3825
	male	141	379	980	2490	3558
ü	female	234	424	1938	2742	3694
	male	139	333	1769	2337	3342

2.2. Harmony of Consonants

In Turkish, consonants also follow certain rules. Consonants are divided into subsets as shown in Table-4.

• *Hard and soft consonants according to the vibration of sound beams*

• *Continuous and discontinuous consonants, depending on whether the vocal path is blocked or open after the vocal chords.*

• *Lip, palate, tooth and larynx consonants, depending on where the sound is produced.*

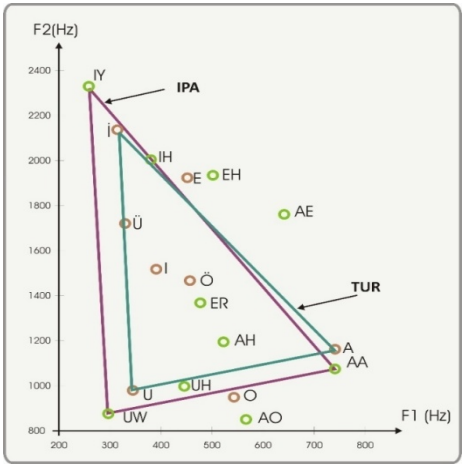


Figure-3: Comparison of Turkish Formants with IPA Formants

The general status of consonants is shown in Table-3.

Table 3.

Clustering of Turkish Consonants

	Voiced consonants			**Volatile consonants**	
	Continuous			Discontinuous	
	fluent	infiltrating	explosive	infiltrating	explosive
Double lip	m		b		p
Tooth-lip		v		f	
Tooth	n, l, r	z	d	s	t

Tooth-palate		j	c	ş	ç
Front palate	n		g		k
Back-palate	ñ	ğ	g		k
Glottal				h	
Half vowel	y				

The classification of consonants of Turkish is shown in Table-4 and their harmony is shown in Figure-4.

Table 4.

Classification of Turkish Consonants

Hard Consonants (HC)	ç, f, h, k, p, s, ş, t
Soft Consonants with No Hard Equivalent (NSC)	l, m, n, r, y
Soft with Hard Equivalent (ESC)	b, c, d, g, ğ, j, v, z

It is known that Turkish is rich in vowel sounds. Vowel letters are used more in Turkish than in other languages. The result of a study conducted by E. Çiçek and A. E. Yılmaz is shown in the order of the frequency of letters in Table-5 [5].

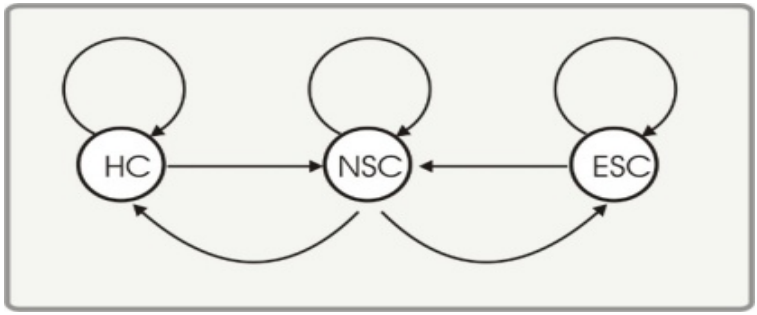


Figure-4: **Rules of harmony of consonants in Turkish**

Table 5.

Frequency of Use of Letters in Turkish

Letter	*Freq. %*	*Letter*	*Freq. %*	*Letter*	*Freq. %*	*Letter*	*Freq. %*
a	*11,46*	*ı*	*4,56*	*ü*	*1,92*	*c*	*0,92*
i	*9,32*	*t*	*3,60*	*ş*	*1,53*	*p*	*0,87*
e	*9,07*	*m*	*3,51*	*z*	*1,5*	*ö*	*0,77*
n	*7,42*	*y*	*3,32*	*g*	*1,15*	*f*	*0,49*
r	*7,04*	*s*	*3,15*	*h*	*1,11*	*j*	*0,05*
l	*6,4*	*u*	*3,14*	*ğ*	*1,047*		
K	*4,65*	*b*	*2,67*	*ç*	*1,046*		
d	*4,6*	*o*	*2,58*	*v*	*1,01*		

According to the results of this study, the ratio of vowel used in Turkish to all letters is 42.82% and consonants 57.18%. Results related to other languages are presented in Table-6, calculated from the data in reference [6]. As a result, it can be said that vowels are used extensively in Turkish.

Table 6.

Vowel and Consonant Usage Rates of Different Languages

Languages	*Vowel letter frequency (%)*	*Frequency of consonant use (%)*
Italian	*47,617*	*52,383*
Finnish	*45,65*	*54,355*
French	*44,811*	*55,189*
Spanish	*44,23*	*55,77*
Turkish	*42,82*	*57,18*
Czech	*41,52*	*58,47*
Polish	*39,20*	*60,803*
German	*38,238*	*61,762*
English	*38,1*	*61,9*
Swedish	*36,19*	*59*

Easy to Speak

The harmony of vowels and consonants in Turkish facilitates the production of these sounds in the sound production organ.

Nice Sound

We can compare the regular arrangement of vowels and consonants within the word to the arrangement of notes in a musical work. The sounds made when we press the piano keys indiscriminately are not pleasing to our ears. On the other hand, when we start playing a musical piece, a sound that everyone likes is created. Because in the musical work, the notes are arranged in a certain order. The fact that vowels and consonants are arranged in accordance with certain rules in Turkish makes Turkish sound nice.

According to a research conducted by the Department of Phonology of the University of Trier in Germany:

- Vowels are the melody sounds of a language.
- The aesthetics of a language is directly proportional to the number of its vowels.
- The longer the duration of the vowel sounds brings fluency to the language.

The results of this research show why Turkish sounds good. There are 13 vowel sounds in Turkish. In addition, the duration can be doubled when used with the thick ordinary vowel sounds "ğ".

Energy Conservation

The rules in the arrangement of vowels and consonants in Turkish are the natural result of the structure of our vocal organs. When we want to make a sound without tiring or forcing our vocal organs, we need to follow the vowel and consonant rules. When we evaluate it from this point of view, we can say that Turkish is an energy-saving language. The jaw of a Turkish-speaking person does not get tired.

3. VOCABULARY

The effectiveness of a language can also be measured by the richness of its vocabulary. The size of a language's vocabulary is directly proportional to the person's ability to think. When investigating the richness of a language's vocabulary, looking only at the number of words in the dictionary does not give correct results.

When we look at the dictionaries related to the English or French language, we can see the richness of words. On the other hand, vocabulary in Turkish dictionaries can be seen as little. In fusional languages, it is necessary to give a new name to every concrete and abstract object. This necessity causes a wide vocabulary in such languages. In additive and Hami-Sami languages, it is possible to derive new words with the additions added to the root word. Even if some of the words that can be derived are not included in dictionaries, their meaning is understood by those who use that language. We can explain the situation with well-known examples.

In Table-7, Turkish words derived from the word eye, which is the organ of vision, by adding only constructional suffixes and the corresponding English words are shown.

The word eye, which means mirror, seen in Table-7 is a word used in the past. The fact that we can easily interpret all four words derived from this word, which is not used today, after learning its meaning, is a proof of Turkish's ability to derive words. Another result of Table-7 is that there is a word in English for Turkish words. However, most of them are not related to the organ of vision. This situation may cause difficulties in human perception and interpretation of words.

Table 7.

Turkish Word Derivation Ability

Turkish	English	Turkish	English
Göz	Eye	**Göz**lemci	Observer
Gözlük	Eyeglasses	**Göz**lemcilik	Observation
Gözlükçü	Optician	**Göz**de	Favourite
Gözlükçülük	Opticians	**Göz**gü	Mirror
Gözcü	Watchman	**Göz**gülük	Mirror stand
Gözcülük	Opthamolgy	**Göz**gücü	Mirror maker
Gözlem	Observation	**Göz**gücülük	Mirror makers business
Gözleme	Observing		

Hami-Sami languages have the ability to derive new words from a root. For example, words such as *ketebe*, *kitabkatib, katibe, mekteb* (*clerks*, book, *clerk, lady clerk*, s*chool)* can be derived from a root consisting of "KTB". The derivation form also specifies the meaning of the new word. For example, the lady who reads on kettle, writes scribe, writes on scribe gives the meaning of the place where the school is read.

Although Turkish words (except reinforcement prefixes and foreign words) only have suffixes, English can also have prefixes. There are those who consider the lack of prefixes in Turkish as a deficiency. These people are the ones who used to use negative suffixes to words according to Arabic and Persian rules in the past. For example e*mevcut, namevcut* (*available, absent)*. In the same habit, they also advocate giving words a gender feature. For example, like *memur, memure* (*officer, lady officer*). It is clear that such ideas are irrelevant for Turkish. In Turkish, new words or additional words are used instead of prefixes. Examples on this issue are given in Table-8.

Table 8.

Prefixes are not used in Turkish

Turkish	English
Olanak**sız**	**Im**possible
Yenidencanlandırmak	**Re**activate
Tepkin	**Re**active
Düzeltmek	**Re**form
Mut**suz**	**Un**happy
Önlem	**Pre**caution
Önyargı**lı**	**Pre**conceived
Özür**lü**	**Dis**abled

While evaluating the richness of a language's vocabulary, it is not enough to look only at the noun noble words, it is also necessary to examine the verb noble words. Apart from the basic form of every action in Turkish, such as active, passive, reflexive, active, and

causative states provide a significant advantage over other languages. The verb structures of Turkish are shown in Table-9.

Languages develop according to the lifestyle, habits and interests of the nations that use that language. While the number of words in this field is increasing in societies that are interested in philosophical issues, the number of words describing kinship is increasing in societies with strong family ties. While some languages try to explain the events with concrete verb words, some languages prefer to explain with idioms. It may be necessary to describe an action or situation described with an idiom in a one-page article in another language.

Table 9.

Predicate Structures in Turkish

Predicate type	Turkish	English
Active	Görmek	To see
Passive	Görünmek	To be seen
Passive	Görülmek	To be seen
Reflexive	Giyinmek	To dress
Reciprocal	Görüşmek	To see each other or to discuss
Causative	Görüştürmek	To bring someone to see or to discuss each other
Causative in second degree	Görüştürtmek	To have somebodies to see or to discuss each other
Passive causative	Görüştürülmek	To be brought to see or to discuss somebody

One of the issues that is mentioned incompletely in Turkish is that words do not change according to gender. For example, the feature of giving different names to men and women who have undertaken the same profession or duty such as *melik-melike (kink-quin), rahip-rahibe (priest-nun), memur-memure (officer-lady officer)*, is not available in Turkish. When it is necessary to specify the gender of the person, this deficiency is eliminated by putting a

word before or after the word that determines the gender. For example*, öğretmen hanım (lady teacher) or hanım öğretmen(lady teacher)*.

In languages where words are given femininity and masculinity, for example, in French and Arabic, the prefix at the beginning of the word indicates the gender of the word. To give an example from French, *la mur* (wall), *le port* (door), the word wall is female and the word door is male. It is really funny and difficult to understand why such a distinction is necessary, perhaps it is meaningless.

4. EFFİCİENCY

Adding a suffix to words in agglutinative languages give new meanings to words. These features not only enrich the vocabulary of the language, but also make the language more efficient. Active and passive forms of actions are seen in all languages. However, forming reflexive, active, causative, secondary causative, causative passive forms of actions is not as easy and regular as in Turkish. It is not even possible in some languages.

An indicator of the development of a language is the meanings attached to words. For example, the most basic use of *kırmak* (*breaking)* is to (*smash hard things by hitting or crushing)*. It is possible to come across many meanings of the word to break in developed languages such as Turkish. For example:

- *Soğukhayvanlarıkırdı (öldürmek) – Coldkill animals*
- *Fiyatlarıkırmak (indirimyapmak) -Breaking prices (discounting)*
- *Kalbimikırdı (gücenmek) - It broke my heart (offended)*
- *Pulunukırmak (tavlaoyununda) - Breaking your checker (in backgammon game)*
- *Buğdayıkırdırmak (kabaöğütmek) - Crushing the wheat (coarse grinding)*
- *Direksiyonukırmak (Çevirmek, yönünüdeğiştirmek) - Breaking the steering wheel (turning, changing direction)*
- *Soğuğunbelinikırmak (Soğuğunetkisiniciddiolarakazaltmak) - Breaking the waist of the cold (seriously reducing the effect of the cold)*
- *Senet kırmak (senediparayaçevirmek) - Issuing bills (converting the bill into cash)*
- *Dersikırmak (derstenkaçmak) - Breaking the lesson (avoiding the lesson)*

5. CLARITY

A certain order is expected in the order of the words in the sentence. In every language, certain rules are followed in ordering the words in sentences. If this order is changed in some languages, the meaning of the sentence is distorted or lost. In Turkish sentences, the words are arranged in a certain order. However, there is no loss of meaning when the order is changed. However, the difference in emphasis occurs. For example, the meanings of these three sentences are the same, the emphasized word is different.

- *Elinisabunlayıka.* - Wash your hand with soap. (emphasize hand)
- *Sabunlaeliniyıka.* - Wash your hands with soap. (emphasize soap)
- *Yıkaelinisabunla.* - Wash your hand with soap (emphasize washing)

The sentence structure in Indo-European languages is rigid. Therefore, when the places of the words in the sentence are changed, the meaning of the sentence changes or deteriorates. The flexibility of sentence structure increases the meaning power of Turkish and provides skill and convenience to the speakers. However, this feature causes additional difficulties in natural language processing studies.

6. BEING RULES

It is thought that languages develop with the development of humanity, so it should not be expected to have certain rules. However, this opinion is not valid for Turkish. It is as if a language congress was held 5000 years ago and the rules of Turkish were determined. These rules have not changed until today. This feature of Turkish astonishes linguists. The unchanging rules of Turkish have never been broken in phonics and morphology. An exception can be given in the rules of morphology. As it is known, when possessive suffixes are added to a word ending with a vowel, the letter "n" is added in between. According to this rule, when we should say su+n+unbut say *suyun*. This situation, which we can consider as the rule breakdown, actually stems from the fact that the word *su* was used as *suy* in the past.

We can also examine languages in terms of case suffixes. The meaning of the case suffixes in Turkish is very precise. This certainty is not at the same level in Indo-European languages. For example:

- *Müziğidinliyorum.* - I'm listening to the music. (I am listening to music)
- *Partililerekonuşuyorum.* - I am speaking to the party members. (I am speaking to members of parties)
- *Senidüşünüyordum.* - I was thinking of you. (I think of you)

As it is known, there are six cases of nouns in Turkish: Lean, -i (indicating), -e (orientation), -de (presence), -den(leaving), -in (affiliation). These attachments can be thought of as vectors because they indicate direction. It cannot be wrong in the meanings given by the case suffixes. For example, someone is listening; is told to someone. Note that the same direction indicator (to) is used for these two sentences in English.

The roots of Turkish verbs never change. For example, if we write the action to go for all times(**git**+tim, **git**+mişim, **gid**+iyorum, **gid**+eceğim,**gid**+erim) we can clearly see this feature. In the example we have given, the letter "t" can turn into "d" due to the phonetic rule, and this is a rule. There is not even a letter similarity between the present and the past of the act of going in the Indo-European language family, Persian, French and English. For example,

- I go (present tense) I went (past tense) (English)
- Je vais (present tense) Je suisallé (past tense) (French)
- Hahemreft (present tense) reftem (past tense) (Persian)

Such irregularities are not encountered in the shooting of Turkish verbs.

Turkish is a regular language in terms of phonetic structure. A word can be easily divided into nouns based on rules. Turkish phonemes and possible numbers are given in Table-10.

Table 10.

Syllable Patterns in Turkish and Their Possible Numbers

Syllable structure	***S***	***SZ***	***SZZ***	***ZS***	***ZSZ***	***ZSZZ***
Possible number	*8*	*128*	*112*	*152*	*2312*	*2128*

(S: voweland Z: consonant)

Looking at the numbers given in Table-10, it can be said that there are 4840 syllables in Turkish. Accordingly, it can be said that when 4840 syllablesare voiced, Turkish text vocalization can be realized. How to emphasize the syllable in the word in Turkish is shown in Figure-5

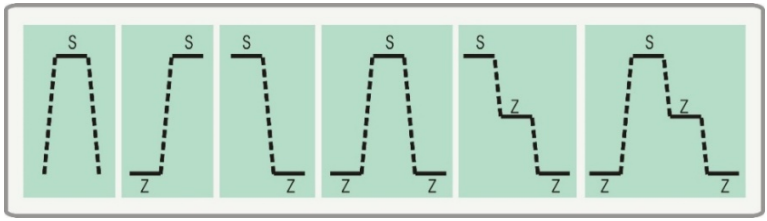


Figure-5: Peak ups and downs in Turkish syllable

There is no need to use a dictionary to separate a Turkish word into its syllables, it can be accomplished using a very easy algorithm.

Syllable ending with a vowel are called open vowels, and those ending with consonants are called closed vocalizations. Univocal syllable in Turkish must also conform to one of these six forms. Most of the monophonic root words are closed syllable. Al, vur, sev,

at, kuş, aş.Turkish syllable are as follows:

1 - The first of the two consonant sounds side by side in the word forms a syllable with the vowel before it, the second with the vowel after it: bir-lik

2 - The first two of the three side by side consonants in the word form the syllable with the vowel before it, and the third with the vowel after it: kork-maz, Türk-çe

3 - The first two of the four side by side consonants in the word form the syllable with the vowel before it, and the last two with the vowel after it.

Words are sets of sounds with one or more syllables that serve as a grammar or meaning in a sentence. Every word has a meaning. When Turkish is evaluated in terms of phonology, it is clear that words cannot be randomly derived. Phonetic rules and phonemes are determinants when deriving a word. A question may come to mind: "Can an artificial dictionary be created in accordance with the

phonetic rules of Turkish?" The results of a study conducted in order to find an answer to this question are presented below: As a result of the examinations made on Turkish texts, it was observed that the average length of the root words was approximately five. The lengths of the root words are given in Figure-6 [7].

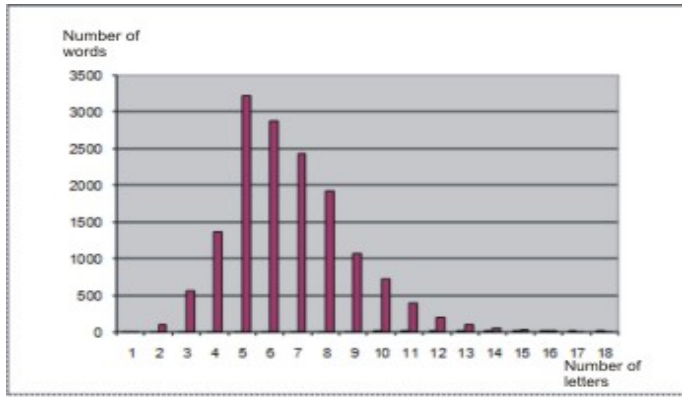


Figure-6: Statistics of Turkish word lengths

In another study, the first and last letters of the words were searched. Accordingly, the result obtained is given in Figure-7 and Figure-8 [7].

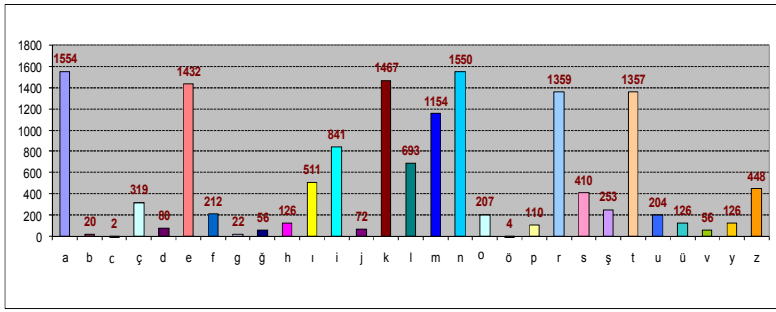


Figure-7: Statistical results of Turkish words according to their first letters

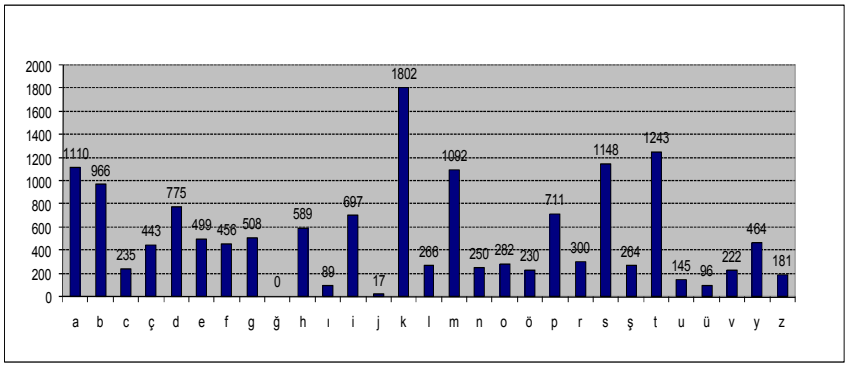


Figure-8: Statistical results of Turkish words according to their last letters

The syllable structures of words consisting of two syllables are as follows:

[S + ZS] (aba), [S + ZSZ] (ocak), [S + ZSZZ] (avurt),

[SZ + ZS] (ayna), [SZ + ZSZ] (aşkın), [SZ + ZSZZ] (aldanç),

[ZS + ZS] (baba), [ZS + ZSZ] (tabak), [ZS + ZSZZ] (kazanç),

[ZSZ + ZS] (bohça), [ZSZ + ZSZ] (bostan), [ZSZ + ZSZZ] (başkurt)

As a result, Turkish words consisting of one and two syllables can be formed in 18 different ways. Meanwhile, in addition to the basic sound features described in this section, the following basic rules should be kept in mind:

- Since three consonants cannot be side by side, the SZZ and ZSZZ syllables cannot form the initial syllable.
- Two vowels are not considered side by side.
- If there are two consonants in the same notation, this consonant pair is one of the following: lç, lk, lp, lt, nç, nk, nt, rç, rk, rp, rs, rt, st, sht
- "j" does not exist in Turkish-origin words, so it does not need to be included in consonants.
- There are no Turkish words that begin with "ğ". If there is a suffix that starts with a consonant and ends with a consonant, then the second notation does not begin with "ğ".
- While forming syllables, the rule of resembling hard consonants should be applied.
- The letter "o" can only be in the first sound of the word.

- When the suffix that begins with a vowel ending with a consonant, they cannot be the same letter.
- The number of words in the dictionary created in accordance with these rules is given in Table-11.

Table 11.

Possible Number of Two Syllables

Monosyllabic words		***Two syllable words***			
struc-ture	***num-ber***	***structure***	***Num-ber***	***structure***	***number***
S	*8*	*S + ZS*	*342*	*ZS + ZS*	*6.498*
SZ	*128*	*S + Z SZ*	*5.202*	*ZS + ZSZ*	*98.838*
SZZ	*112*	*S + ZSZZ*	*4.788*	*ZS + ZSZZ*	*90.972*
ZS	*152*	*SZ + ZS*	*4.914*	*ZSZ + ZS*	*88.794*
ZSZ	*2312*	*SZ + ZSZ*	*74.574*	*ZSZ + ZSZ*	*1.347.534*
ZSZZ	*2128*	*SZ + ZSZZ*	*68.796*	*ZSZ + ZSZZ*	*1.243.116*

The graph of the number of words with one and two syllables is given in Figure-9. If you pay attention to Table-11, there may be 3.039.208 words in Turkish consisting of two syllables. When this artificially created dictionary is examined, the following comments can be made:

Figure-9: Statistical results of syllabletypes

The number of Turkish words using ZSZZ as the second syllable seems high, but words that fit this pattern are rare in our language. Even if all words that fit this pattern are ignored, the number of two-syllable words becomes 1,636,324.

REFERENCES

1. Adalı E. Türkçe Doğal DilİşlemeAkçağyayınevi, 2020, ISBN97860553425519.

2. O. Türk, Ö. Şaylı, A. S. Özsoy, L. M. Arslan, TürkçedeÜnlülerin Formant İncelemesi 18. DilbilimKurultayı, 20-21 Mayıs 2004, Ankara.

3. A. Y. Davutoğlu, Standart Türkçedeki Ünlülerin Akustik Analizi ve Fonetik Altyapı, Doktora Tezi. İstanbul Üniv., 2010.

4. E.Malkoç, Türkçe Ünlü Formant Frekans Değerleri ve Bu Değerlere Dayalı Ünlü Dörtgeni, Dil DergisiSayı: 146 2009.

5. E. Çiçek, A. E. Yılmaz, A new Morse Code Scheme Optimized According to the Statistical Properties of Turkish, Turk J Elec Eng& Comp Sci, 2013, 21: 804.811.

6. /wiki/Letter_frequency, 2014.

7. Adalı E. ve Büyükkuşçu, Y., Heceleme Yöntemiyle Kök Sözcük Üretme, TBV Bilgisayar Bilimleri ve Mühendisliği Dergisi, Sayı:2, 2006. Bilimleri ve Mühendisliği Dergisi, 02, 25.29.

Gataullin R.R.
Institute of Applied Semiotics of the AS RT,
Russia, Tatarstan, Kazan

WEB INTERFACE FOR TATAR NLP PIPELINE

Abstract. This paper describes a web interface for the nlp pipeline of the Tatar language, designed to facilitate and speed up the process of building nlp pipelines from blocks (modules) for processing NL data in the Tatar language. A pipeline (from English, pipeline - a pipe) is a software pipeline, a chain of processes for converting source data into an end result. NLP pipeline is a pipeline for processing natural language data, where text data is usually supplied as input, and annotated text in the required format is received at the output. At the moment, interactive mode is available with entering text data into the form of a web page and the mode of processing a file or an entire archive. There is also an Application Programming Interface (API) for integration into third-party applications. If necessary, the web interface can be redesigned to support the NL data processing in other languages. In this case, it is sufficient to replace the existing modules with modules of the corresponding language. It is also possible to programmatically add a new language to the list of interface languages.

Keywords: *nlp-pipeline, Tatar language, web interface.*

Гатауллин Р.Р.
Институт прикладной семиотики АН РТ,
Россия, Татарстан, Казань

ВЕБ-ИНТЕРФЕЙС ДЛЯ ТАТАРСКОГО NLP-ПАЙПЛАЙНА

Аннотация. В данной работе описывается веб-интерфейс для nlp-пайплайна татарского языка, призванный облегчить и ускорить процесс построения nlp-пайплайнов из блоков (модулей) для обработки ЕЯ-данных на татарском языке. Пайплайн (с англ., pipeline – труба) – это программный

конвейер, цепочка процессов преобразования исходных данных в конечный результат. NLP-пайплайн – пайплайн обработки естественно-языковых данных, где на вход обычно подаются текстовые данные, а на выходе получают аннотированный текст в требуемом формате. На данный момент доступны интерактивный режим с вводом текстовых данных в форму веб-страницы и режим обработки файла или архива целиком. Также имеется программный веб-интерфейс (ApplicationProgramming Interface, API) для интеграций в сторонние приложения. При необходимости веб-интерфейс может быть переделан для поддержки процесса обработки ЕЯ-данных и на других языках. При этом достаточно заменить существующие модули на модули соответствующего языка. Также есть возможность программного добавления нового языка в список языков интерфейса

Ключевые слова: *татарский язык, веб-интерфейс*.

1. Основные принципы работы пайплайна

Пайплайн состоит из блоков/модулей/этапов, которые подключаются последовательно, таким образом, что выходные данные одного модуля в неизменном виде передаются на вход следующего (см. Рис. 1). В данной реализации никакие ветвления процессов не допускаются: каждый модуль имеет один вход и один выход. Единицей (квантом) входных/выходных данных считается документ в определенном формате.

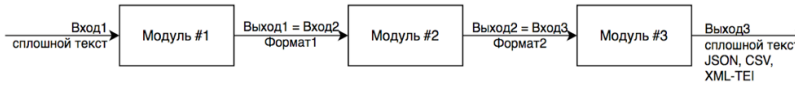


Рис. 1. Основной принцип работы nlp-пайплайна татарского языка

Формат данных входного и выходного потока для каждого модуля может отличаться, хотя и не обязательно. Например,

Модуль #1 (см. Рис. 1) на вход может получать сплошной текст (т.к. является первым этапом обработки), а на выход может выдавать отформатированный и структурированный определенным образом документ, который легче обработать на последующих этапах. Главное, на что нужно обратить внимание – это чтобы последующие этапы/модули могли принимать на вход соответствующий формат данных. В текущей версии веб-приложения не реализована проверка соответствия форматов входных и выходных данных при построении пайплайнов. Задача синхронизации и унификации формата данных остается за разработчиками отдельных модулей пайплайна. Также немаловажно на выходе из пайплайна поставить модуль перевода в требуемый выходной формат, будь то JSON, CSV, TEI (xml) или просто сплошной текст. Хотя и применение таких форматов во внутренних этапах не лишено смысла.

2. Архитектура веб-приложения

Программный интерфейс nlp-пайплайна татарского языка реализует клиент-серверную архитектуру, в которой клиентом выступает браузер, установленный на пользовательской машине, а сервером – веб-сервер. Программный комплекс состоит из следующих компонентов (см. Рис. 2):

- Веб-приложение (на ЯП Python 3.7, веб-фреймворк Flask с расширением Flask Appbuilder [1]);
- База данных (СУБД PostgreSQL);
- Брокер сообщений (REDIS);
- Система очередей задач Celery (на ЯП Python 3.7) [2].

Взаимодействие встраиваемых модулей осуществляется посредством общей реляционной базы данных и системы асинхронного выполнения очередей задач Celery, которая состоит из нескольких взаимосвязанных частей: Celery Client, Message (Queue) Broker, Celery Workers. Внедрение очереди задач позволяет выгружать некоторые задачи в отдельный поток и продолжать работать с пользовательским интерфейсом без каких-либо задержек. Таким образом, избегается блокировка GUI при запуске долгих и сложных вычислений в пайплайнах.

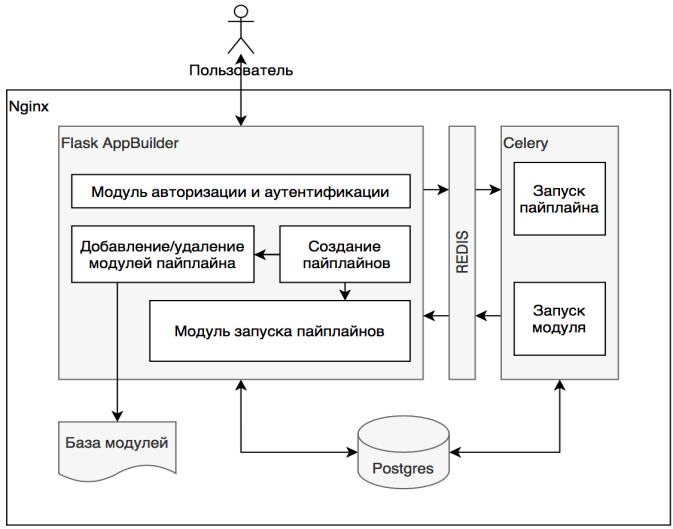


Рис. 2. Клиент-серверная архитектура веб-приложения

Модули пайплайна хранятся в виде файлов в Базе модулей. Пайплайны хранятся в реляционной БД в форме взаимосвязанных записей (см. Рис. 3). Там же хранится история запусков, входные и выходные данные пайплайнов.

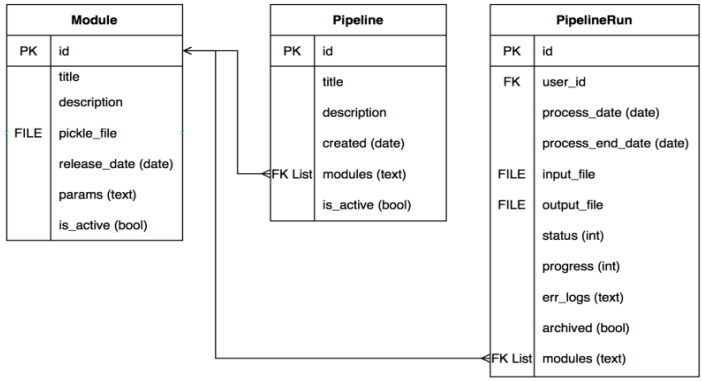


Рис. 3. Схема реляционной базы данных

1.1. Модуль авторизации и аутентификации пользователя

Модуль предназначен для разграничения прав доступа для различных групп пользователей. Группы и уровень доступа определяется администратором системы.

На данный момент определены 3 уровня доступа (роли):

- Незарегистрированный пользователь (Public): доступны просмотр страниц модулей, пайплайнов, документации; доступен интерактивный режим;
- Зарегистрированный пользователь (RunPipeline): к возможностям незарегистрированных пользователей добавляется возможность запускать пайплайны в режиме обработки файлов/архивов, а также в режиме API;
- Администратор системы (Admin) имеет полный доступ к системе, в том числе добавлять новые модули и сохранять пайплайны (для возможности повторного использования другими пользователями).

1.2. Модули пайплайна

При построении nlp-пайплайна используются независимые модули, которые загружаются в систему в виде pickle-файлов, которые в последующем превращаются в исполняемый код. В системе модули пайплайна хранятся в виде файлов в Базе модулей (см. Рис. 4).

Для добавления модуля в Базу модулей необходимо:

- разработать основной функционал модуля на ЯП Python 3.7;
- на программном уровне синхронизировать входные и выходные форматы данных модуля, чтобы модуль мог правильно считать данные с выходного потока предыдущего модуля, обработать его и на выход передать данные в правильным формате для обработки последующим модулем;
- готовый модуль оформить в виде класса с необходимыми обязательными полями и методами;
- с помощью библиотек Pickle (Python 3.7.) или Dill (Python 3.7) подготовить pickle-файл с объектом класса;
- заполнить описание модуля и загрузить pickle-файл в систему;

- при загрузке pickle-файла в систему происходит достаточно простая проверка на соответствие формата pickle-файл и при успешной валидации модуль попадает в Базу модулей, после чего может быть использован при построении пайплайнов.

1.3. Пайплайны

Обычно определенная последовательность этапов обработки (т.е. пайплайны) часто повторяется и появляется необходимость в сохранении данной архитектуры для последующего использования. Например, процесс морфологического анализа может состоять из этапов токенизации, морфологической аннотации, разрешения морфологической многозначности и форматирования выходного потока (JSON, CSV или сплошной текст). При этом возможно, что у некоторых этапов разработано несколько версий, выходные данные которых хоть и незначительно, но отличаются друг от друга. И, например, если есть задача обработать коллекцию текстов, то для всей коллекции важно использовать один и тот же пайплайн с теми же параметрами, чтобы результаты были одинаковыми.

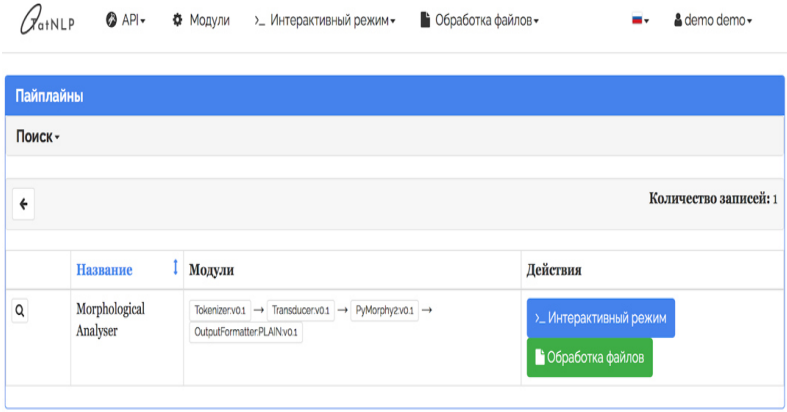


Рис. 4. Основная страница сохраненных пайплайнов

Для таких случаев разработан функционал сохранения и повторного использования сохраненных пайплайнов (см. Рис. 4) (как в интерактивном режиме, так и в режиме обработки файлов/архивов). При этом важный момент: право сохранять пайплайны имеет только пользователь уровня Администратор (при отсутствии таких прав, но необходимости в определенном пайплайне, можно направить соответствующий запрос администраторам системы).

1.4. Интерактивный режим

Интерактивный режим – режим тестирования пайплайна, когда введенный в окно браузера текст проходит все указанные этапы обработки, и результат возвращается в виде сплошного текста на той же веб-странице.

При этом есть возможность менять как сами модули пайплайна, так и последовательность исполнения модулей. Также один и тот же модуль можно запускать несколько раз.

Режим доступен как зарегистрированным, так и незарегистрированным пользователям.

1.5. Режим обработки файла/архива

Режим обработки файла/архива – режим работы, когда на сервер загружается файл (или архив файлов), который проходит все указанные этапы пайплайна, и результат возвращается в виде файла (или архива, соответственно) в требуемом формате.

В отличие от интерактивного режима, в режиме анализа файлов (и архивов файлов) входных данных обычно в разы больше, поэтому весь процесс обработки вынесен в отдельный поток (с помощью библиотеки Celery на Python 3.7). В зависимости от объема входных данных и сложности каждого этапов nlp-пайплайна, для обработки всего пайплайна может потребоваться достаточно много времени. Для отслеживания процесса обработки было введено понятие Запуск пайплайна (eng, PipelineRun), для которого разработана отдельная страница (см. Рис. 5-6).

Режим доступен только зарегистрированным пользователям.

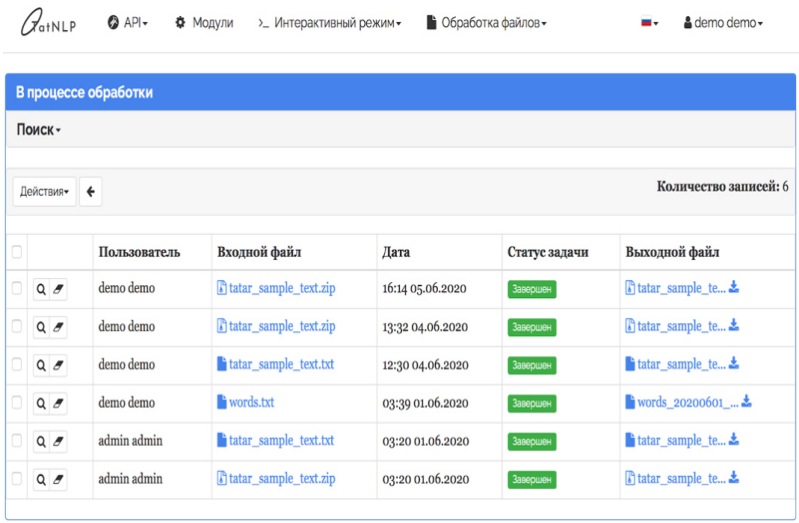


Рис. 5. Страница запуска пайплайна

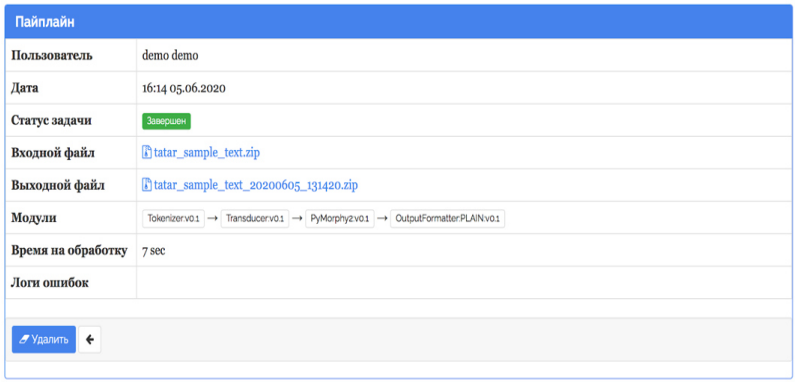


Рис. 6. Страница описания одного из Запусков nlp-пайплайна

1.6. API

Очень часто пайплайны используются в процессах других (сторонних) приложений. Для поддержки таких случаев на основе спецификации OpenAPI [3] был разработан соответствующий программный интерфейс.

Доступны следующие эндпоинты (endpoint – конечная точка, точка обращения):

- **POST** на */api/v1/security/login* – авторизация с получением токена аутентификации;
- **POST** на */api/v1/module/run* – запуск отдельного модуля пайплайна из Базы модулей;
- **POST** на */api/v1/pipeline/run* – запуск определенного сохраненного пайплайна;
- **GET** на */api/v1/pipeline/status/<pipeline_run_id>* – запрос статуса запущенного пайплайна.

API можно протестировать с помощью встроенного инструмента SwaggerTool [4] (доступна по адресу "/swaggerview/v1").

1.7. Мультиязычный пользовательский интерфейс

С помощью библиотеки Babel (Python 3.7) в веб-приложении реализован мультиязычный пользовательский интерфейс. Переключение языка осуществляется с помощью кнопки выбора языка в верхнем правом углу, рядом с кнопкой входа в систему. На данный момент доступны русский и английский языки, к релизу готовится татарский язык интерфейса. Для добавления нового языка интерфейса, достаточно подготовить и, используя библиотеку Babel, скомпилировать файл соответствий языковых отображений для нового языка и добавить в кодовую базу веб-приложения.

3. Заключение

Разработан веб-интерфейс nlp-пайплайна, облегчающий процесс построения и запуска nlp-пайплайнов разной степени сложности на удаленном сервере. Если интерактивный режим позволяет быстро протестировать разные пайплайны с разными версиями модулей и выбрать наиболее подходящую конфигурацию для текущей задачи, то режим обработки файлов (архивов) позволяет относительно быстро обработать большие объемы данных с выбранной конфигурацией. Интеграция с помощью программного интерфейса в сторонние приложения позволяет быстро автоматизировать процессы обработки ЕЯ-

данных, в результате облегчая решение некоторых рутинных задач.

Веб-приложение может быть полезным как для исследователей в области компьютерной лингвистики как инструмент анализа и обработки ЕЯ-данных, так и для инженеров, разрабатывающих системы обработки ЕЯ-данных, как инструмент для автоматизации процессов обработки.

ЛИТЕРАТУРА

1. Документация к Flask-AppBuilder (на англ. языке) [Электронный документ]. URL: https://flask-appbuilder.readthedocs.io/en/latest/ [Дата обращения: 20.10.2020].
2. Документация к системе очередей задач Celery (на ЯП Python 2, 3) (на англ. языке) [Электронный документ]. URL: https://docs.celeryproject.org/en/stable/index.html [Дата обращения: 20.10.2020].
3. Open API pecification (на англ. языке) [Электронный документ]. URL: https://github.com/OAI/OpenAPI-Specification [Дата обращения: 20.10.2020].
4. Программная библиотека инструмента flask-swagger (версия 0.2.14, ЯП Python 2.7, 3.4) [Электронный документ]. URL: https://pypi.org/project/flask-swagger/ [Дата обращения: 20.10.2020].

Yergesh B.G.
Eurasian National University named after L.N. Gumilyov, Kazakhstan, Nur-Sultan

RULE-BASED SENTIMENT ANALYSIS OF COMMENTS IN SOCIAL NETWORKS

Abstarct. Every day, several terabytes of new information appears in the web.Many of them are blo

gs, tweets, articles, and various text, audio, and video information that reflect opinions about various products, companies, movies and etc. manual processing of such large amounts of information becomes impossible.Therefore, the problems of developing formal models and tools for automation and analysis are very actual. This paper describes a rule based approach to polarity detection of kazakh language commentsin social networks. An algorithm for sentiment analysis of Kazakh-language texts is proposed, problems that worsen the result of the analysis are identified.

Keywords: *natural language processing; sentiment analysis; text analysis; Kazakh language*

Ергеш Б.Ж.
Евразийский национальный университет им. Л.Н. Гумилева, Казахстан, Нур-Султан

АНАЛИЗ ТОНАЛЬНОСТИ КОММЕНТАРИЕВ В СОЦИАЛЬНЫХ СЕТЯХ НА ОСНОВЕ ПРАВИЛ

Аннотация. С каждым днем в сети появляется новая информация объемом несколько терабайт. Большинство из них является блогами, твитами, статьями и различной текстовой, аудио- и видеоинформацией, отражающей мнения о различных продуктах, товарах, компаниях, фильмах и т.д. Ручная обработка таких больших объемов информации становится невозможным. Поэтому проблемы создания формальных моделей и программных средств (инструментов) автоматизации

и анализа весьма актуальны. В данной работе описан подход определения тональности текстов комментариев на казахском языке в социальных сетях на основе правил. Предложен алгоритм сентимент анализа казахоязычных текстов, выявлены проблемы ухудшающие результат анализа.

Ключевые слова: *обработка естественных языков, сентимент анализ; анализ текста, казахский язы*

1. Ведение

Сегодня, благодаря широкому распространению интернета, появилась возможность найти необходимую информацию, рекомендации или отзывы людей. Потому что в настоящее время в Интернете люди открыто высказывают и пишут свои мнения. Активное развитие современных социальных сетей, блог-платформ и форумов вызывает большой интерес научного сообщества и различных организаций профессиональных IT-специалистов к задачам автоматической обработки и анализа мнений пользователей Интернета.

Несмотря на долгую историю в области лингвистики и обработки естественных языков, исследования, связанные с мнением и настроением людей, стали проводиться только с начала 2000 года.

Понятие сентимент анализа имеет много различных названий, интерпретаций, задач, таких как, например, сентимент анализ, интеллектуальный анализ мнений, поиск мнений, поиск субъективности, анализ настроения и т.д [0].

Тональность мнения – признак f, который указывает, что мнение будет положительным, нейтральным или отрицательным. Он также называется полярностью мнения, направлением сентимента или семантическим направлением.

Сентимент анализ является одним из новых направлений в области обработки естественного языка. Хотя сентимент анализ во многих работах рассматривается как простая задача классификации, на самом деле это сложная область исследований, которая требует решения многих задач обработки естественного языка. В работе [0] задачу сентимент анализа рассматривают как большой чемодан, где описаны 15 проблем NLP с помощью трехслойной структуры. По мнению авторов

решение этих задач позволит получить наилучшый результат.Много исследовании ведутся по сентимент анализу английского, итальянского, русского и других языков. Из тюркских групп языков в сети есть работы по изучению сентимент казахского [5-10], турецкого [11, 12] и узбекского [13,14] языков.

Многие исследователи занимаются сентимент анализом, соответственно существует множество различных методов и алгоритмов, которые используются в исследованиях. С одной стороны, в прикладном исследовании сентимент анализа могут применяться методы машинного обучения, методы основанные на лексиконе или лингвистические методы. С другой стороны, классификация методов сентимента анализа может зависеть от уровня их классификации, например, уровня документа, предложения или аспекта [1, 2]. Сентимент анализ на уровне документа классифицирует полный документ, как положительный или отрицательный. В этом случае считается, что в документе описывается один объект. Сентимент анализ на уровне предложения разделяет каждое предложение в документе, как субъективное или объективное, и классифицирует субъективные мнения на положительные или отрицательные. А на уровне объекта и аспекта определяется отношение объекта к конкретному аспекту, потому что пользователь может оставить в одном отзыве различные мнения по нескольким аспектам одного объекта.

Методы анализа сентимента можно использовать для различных типов данных, таких как новости, обзоры, блоги или сообщения в социальных сетях.

Каждый вид данных имеет свои особенности, которые необходимо учитывать при сборе, подготовке, предварительной обработке данных и описании объектов.

В настоящее время сентимент анализ находит применение во многих сферах, таких как отслеживание и анализ отзывов о продукте или компании, определение сторонников или противников политической партии или общественного движения, в различных областях, прогнозирование финансовых доходов. Данные, полученные из социальных сетей и

микроблогов (Facebook, Twitter), представляют большой интерес для исследований и приложений, в связи с возможностью публикации в реальном времени отзывов и настроений людей по любым вопросам и доступностью информации в большом количестве.

2. Сентимент анализ текстов на казахском языке

2.1 Данные

Ниже в Таблице 1 приведены количество комментариев, собранные в Интернете из открытых источников (социальных сетей и новостных порталов). Так как пользователи сети не придерживаются определенного языка и правил правописания, встречаются комментарии на смешанных (казахский, русский) языках, а также ненормализованные структуры предложений. Эти показатели значительно ухудшают результат анализа.

Таблица 1.

Количество данных собранные в Интернете

Источники	**Количество комментариев**
Nur.kz	220
Baq.kz	202
tengrinews	397
Facebook	502
Vkontakte	1000
instagram	600
Всего	**2921**

2.2 Определение тональности текста

Для определения тональности казахскоязычных текстов были выявлены признаки, влияющие на определение тональности текстов на казахском языке [00], построены формальные правила синтаксических правил [8, 0].

Для оценки тональности текста используется 5-значная шкала оценки: -2 (очень негативный), -1 (негативный), 0 (нейтральный), 1 (позитивный), 2 (очень позитивный). Построены формальные правила, определяющие сентимент

текстов на казахском языке с использованием продукционной модели. Фрагменты этих правил опубликованы в работах [00].

На основе предложенного метода формальных правил и размеченной базы по тональности лексических единиц разработан алгоритм и программно реализован. В программе сентимент анализа на вход подается текст на естественном языке. Для определения смысла текста требуется предобработка. К этапам предобработки относятся графематический, морфологический и синтаксический анализы.

Сентимент анализ текста использует семантическую базу и работает по следующему алгоритму:

1. Определение тональной оценки слова: к лексическим единицам в тексте, имеющим тональность, присваивается направление сентимента в соответствии с информацией в семантической базе .
2. Определение тональной оценки лексической единицы сентимент анализа: дает тональную оценку лексических единиц в соответствии с формальными правилами.
3. Определение сентимента текста: вычисляет общий сентимент тональных лексических единиц, найденных в тексте

На рисунке 1 показана структура программы.

Точность определения тональности комментариев около 75%. Но для анализа некоторые комментарий были предварительно обработаны, удалены шумы (стоп-слова, знаки припенания), исправлены орфографические ошибки.

При анализе текстов комментариев, собранных в Интернете, возникли следующие проблемы:

– Использование смешанного языка (казахского, русского, иногда английских слов);

– Неправильное написание слов (без использования букв, характерных для казахского языка (қ, ғ, ұ, ү, ө, і));

– Использование неправильных структур предложений (ненормированная структура).

Результат сентимент анализа значительно улучшится при решении данных проблем.

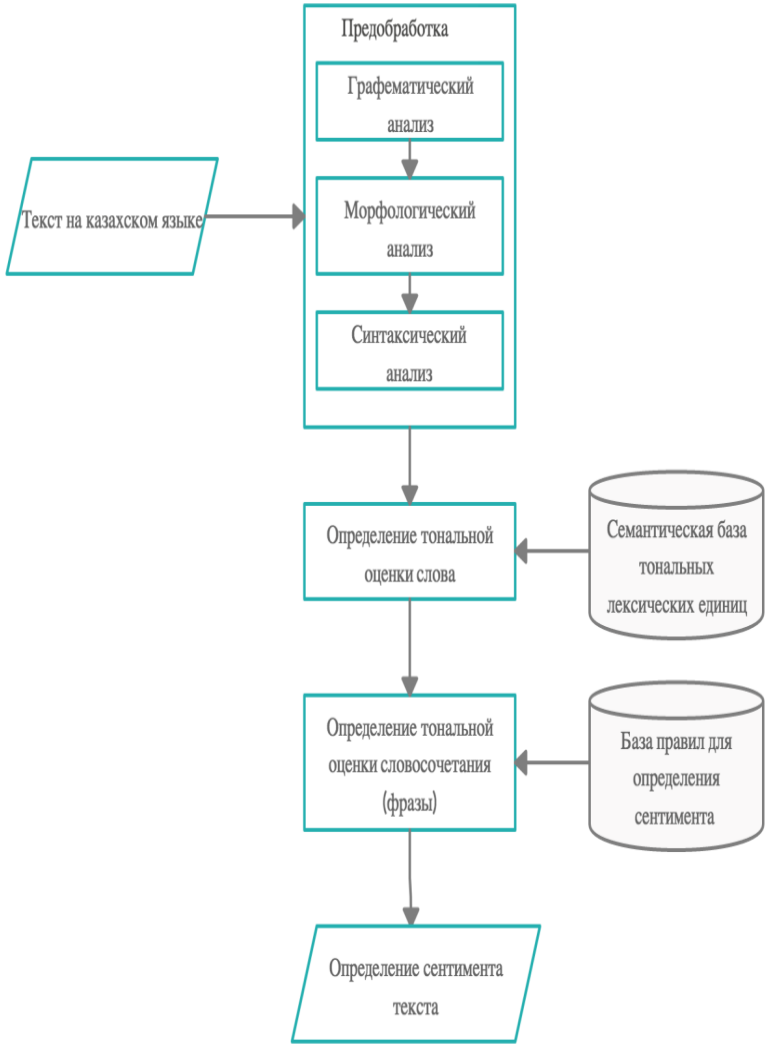


Рис.1. Структура программы сентимент анализа

Заключение

Данная работа посвящена изучению и решению задачи сентимент анализа текстов на казахском языке в социальных сетях. В результате разработки и исследования моделей и методов сентимент анализа текстов на казахском языке были получены семантическая база тональных лексических единиц, предложены и реализованы модели и методы сентимент анализа текстов на казахском языке, точность 75%.

Но тем не менее возникает необходимость предварительной обработки комментариев (проверка правописания, нормализация, определение языка), а также сравнение предложенного метода с методами машинного обучения (ML). В будущем планируется исследование и решение данных проблем.

ЛИТЕРАТУРА

1. Liu B. Sentiment analysis and opinion mining(2012). Synthesis Lectures on Human Language Technologies.Vol. 5(1)

2. Pang B., Lee L. (2008). Opinion mining and sentiment analysis.Foundations and Trends in Information Retrieval (pp.1-135). Vol.2(1-2).

3. Turney P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceed. of the 40th annual meeting on association for computational linguistics (pp. 417-424). Philadelphia.

4. Cambria E., Poria S., Gelbukh A., Thelwall M. (2017). Sentiment Analysis Is a Big Suitcase.IEEE Intelligent Systems (pp. 74-80). Vol. 32, #6.

5. Ергеш Б.Ж., Шарипбай А.А., Бекманова Г.Т. (2016). Роль имен прилагательных в определении тональности текста.Тр. междунар. конф. по компьютерной и когнитивной лингвистике TEL-2016 (С. 85-89). Казань: Изд-во Казан. ун-та.

6. 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 (pp. 39-42). Bishkek. Vol.2(38).

7. Ергеш Б.Ж. (2017). Определение тональности текстов на казахском языке на основе словаря эмоциональной лексики.

Матер. 5-й междунар. конф. по компьютерной обработке тюркских языков «TurkLang 2017» (с. 62-67.). Казань: Издательство Академии наук Республики Татарстан. Т. 1.

8. Yergesh B., Bekmanova G., Sharipbay A. (2019). Sentiment analysis of Kazakh text and their polarity.Web Intelligence (pp. 9-15).IOS Press. Vol.17(1).

9. Ергеш Б.Ж., Шарипбай А.А., Бекманова Г.Т. Модели и методы сентимент анализа текстов на казахском языке. Вычислительная обработка казахского языка: сборник научных трудов / под редакцией Рахимовой Д.Р. Алматы: Қазақ университеті, 2020. Глава 5. ISBN 978-601-04-4698-4.

10. Sakenovich N.S., Zharmagambetov A.S. (2016). On One Approach of Solving Sentiment Analysis Task for Kazakh and Russian Languages Using Deep Learning.In proceed. of the internat. conf. on Computational Collective Intelligence, ICCCI 2016 (pp. 537-545). Sithonia: Springer International Publishing. Vol. 9876.

11. Eryigit G., Çetin F., Yanık M., Temel T., Çiçekli I.(2013). Turksent: A sentiment annotation tool for social media. In proceed. of the 7th Linguistic Annotation Workshop & Interoperability with Discourse (pp. 131-134). Sofia, Bulgary.

12. İşgüder-Şahin G. G., H. R. Zafer and Adali E.(2014). Polarity detection of Turkish comments on technology companies," 2014 International Conference on Asian Language Processing (pp. 136-139). Kuching.doi: 10.1109/IALP.2014.6973514.

13. Kuriyozov E., Matlatipov S, Alonso Pardo M., Gómez-Rodríguez C. (2019). Deep Learning vs. Classic Models on a New Uzbek Sentiment Analysis Dataset.

14. Rabbimov I., Mporas I., Simaki V., Kobilov S. (2020) Investigating the Effect of Emoji in Opinion Classification of Uzbek Movie Review Comments. In: Karpov A., Potapova R. (eds) Speech and Computer. SPECOM 2020. Lecture Notes in Computer Science, vol 12335. Springer, Cham. https://doi.org/10.1007/978-3-030-60276-5_42

15. Sharipbay A., Razakhova B., Mukanova A., Yergesh B, Yelibayeva G. (2019). Syntax parsing model of Kazakh simple sentences.Proceed. of the Second internat. conf. on Data Science, E-Learning and Information Systems (DATA '19).Dubai. #54.

Zhanabekova A.
Institute of Linguistics named after A. Baitursynov,
Kazakhstan, Almaty

LINGUISTIC STATISTICAL FOUNDATIONS OF THE KAZAKH NATIONAL LATINOGRAPHIC KEYBOARD

Abstract. The article describes the ergonomic foundations of the arrangement of the national Kazakh alphabet on the keyboard. For a convenient location of the letters of the Kazakh alphabet on a computer keyboard, statistical studies were carried out to identify frequently used Kazakh letters and their combinations both taking into account their place in a word (at the beginning, at the end of a word), and also without taking them into account. For this purpose, the keyboard itself was conventionally divided into an active (central) section for high-frequency letters and a passive (extreme) section for low-frequency letters of the Kazakh language.

The article also provides statistical data for individual letters of the Kazakh language, provides an ergonomic justification for the choice of places for individual letters of the national Kazakh alphabet.

Keywords: *new alphabet, keyboard, letter frequency, frequency dictionary, statistical analysis.*

Жанабекова А.
Институт языкознания им. А.Байтурсынова,
Казахстан, Алматы

ЛИНГВОСТАТИСТИЧЕСКИЕ ОСНОВЫ СОСТАВЛЕНИЯ КАЗАХСКОЙ НАЦИОНАЛЬНОЙ ЛАТИНОГРАФИЧЕСКОЙ КЛАВИАТУРЫ

Аннотация. В статье описывается эргономические основы расположения национального казахского алфавита на клавиатуре. Для удобного расположения букв казахского алфавита на компьютерной клавиатуре были проведены статистические исследования для выявления часто употребляемых казахских букв и их сочетаний как с учетом их

места в слове (в начале, в конце слова), а также без их учета. С этой целью сама клавиатура была условно разбита на активный (центральный) участок для высокочастотных букв и на пассивный (крайние) участок для низкочастотных букв казахского языка.

В статье также приводятся статистические данные для отдельных букв казахского языка, дано эргономическое обоснование выбора мест для отдельных букв национального казахского алфавита.

Ключевые слова: *новый алфавит, клавиатура, частота букв, частотный словарь, статистический анализ.*

Жаңабекова А.
А.Байтұрсынұлы атындағы Тіл білімі институты,
Қазақстан, Алматы

ЛАТЫН ГРАФИКАЛЫ ҚАЗАҚ ҰЛТТЫҚ ПЕРНЕТАҚТАСЫН ҚҰРАСТЫРУДЫҢ ЛИНГВОСТАТИСТИКАЛЫҚ НЕГІЗДЕРІ

Әріптер мен әріп тіркестерінің жиіліктерін анықтау қазақ әліпбиін латын қарпіне көшірудегі бірнеше теориялық-практикалық бағыттағы мәселелерді шешуге септігін тигізеді. Әсіресе, әріп, әріп тіркестерінің жиілігі мәтінді қолмен теруде жылдамдықты арттыру үшін латын қаріпті қазақ пернетақтасына әріптерді тиімді орналастыруда қажет.

Осы ретте А.Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы лингвистика бөлімі қазақ мәтіндерінің жүйесін құрайтын сөзформалар құрылым бірліктерінің (әріп, әріп тіркесі, буын, қосымша, сөз, сөзформа) жиілік сөздігін құрастырды [1], [2].

Қазақ тіл білімінде әріптерді статистикалық тұрғыдан зерттеудің тарихы 1973 жылы жарық көрген «Қазақ тексінің статистикасы» атты еңбектен басталады деуге болады [3].

Жаңа жиілік сөздікте және оның сипаттамасы жасалған Жинақта [2] қазақ тіліндегі әріп жиілігі бойынша жасалған зерттеулерге шолу жасалып, қазақ тіліндегі жиі кездесетін

дыбыстар (әріптер) мен сирек кездесетін әріптердің жиілігі бойынша сипатталды. Зерттеулердегі жанрлық, екі тілдегі (қазақ және өзбек говорлары) әріптердің жиілігі бойынша салыстырмалы талдаулардан әріп жиілігі бойынша негізінен ұқсас, барабар жиынтық шамалар алынғанын байқаймыз.

Ұсынылып отырған қазақ әріптерінің жиілік сөздігі А.Байтұрсынұлы атындағы Тіл білімі институтының «Қазақ тілінің ұлттық корпусын әзірлеу және жасау» мегажобасы бойынша орындалған «Жалпы білім берудегі қазақ тілінің жиілік сөздігінің» [4] бір түрі ретінде алынған үлкен бір нәтиже – «Сөзформалардың жиілік сөздігінің» мәтінінен алынды. Қазақ әріптерінің жиілік сөздігі екі түрлі нұсқада жасалды. Бірі – «Қазақ тіліндегі әріптердің әліпбилі-жиілік сөздігі», екіншісі «Қазақ тіліндегі әріптердің жиілікті-әліпбилі жиілік сөздігі».

Бұның алдында жасалған әріптердің статистикалық талдауларындағыдай, қазақ тілінде ең жиі кездесетін әріп ***А*** фонемасы. Одан кейінгі орында ***Е, Ы*** фонемалары. Ал Ө, Ұ, Ү дыбыстарының сирек кездесуі қазақ тілінде ерін үндестігінің жазуда сақталмауынан болар деп ойлаймыз. Яғни бұл дыбыстар түбір сөздер құрамында болғанмен, қосымшалар құрамында езулік дыбыстармен жазылғандықтан, аз санды көрсетіп отыр. Ал дауыссыздар ішінде ең жиі кездесетін әріп ***Н*** фонемасы, ***Т, Р, Л*** әріптері жиілігі бойынша екінші орында. 30-орында тұрған Ә-ден төмен тұрған әріптердің барлығы (я, х, ц, ф, в, э, ь, ю, h, ъ, ч, щ, ё) кірме әріптер. Ә әрпінің сирек кездесуі бір жағынан қосымшаларда кездеспеуінен болуы да мүмкін.

Қазақ тіл білімінде әріп тіркестерін статистикалық тұрғыдан зерттеудің тарихы да 1973 жылы жарық көрген «Қазақ тексінің статистикасы» атты еңбектен басталады. Бұл ретте Қ.Бектаев, А.Қ.Жұбановтың «Графема-фонема тіркестерінің жиілік тізімдері туралы», А.Қ.Жұбановтың «Қазақ текстеріндегі әріптер тіркестерінің статистикасы», Е.Агманов пен А.Қ.Жұбановтың «Распределения частот появления сочетаний знаков в орхоно-енисейской письменности», Қ.Бектаевтың «Қарақалпақ тіліндегі фонема тіркестерінің статистикасы» атты еңбектерін атап айтуға болады.

Ұсынылып отырған қазақ әріптер тіркесінің жиілік сөздігі жоғарыда аталған «Сөзформалардың жиілік сөздігінің» мәтінінен алынды.

Қазақ әріп тіркестерінің жиілік сөздігі екі түрлі нұсқада жасалды. Бірі – «Қазақ тіліндегі әріп тіркестерінің әліпбилі-жиілік сөздігі», екіншісі – «Қазақ тіліндегі әріп тіркестерінің жиілікті-әліпбилі жиілік сөздігі».

Екіәріптік жиілік сөздікте әртүрлі тіркестердің рет саны 1160-қа тең, ал осы екіәріптік тіркестер қайталанып қолданыла келіп, олардың абсолютті жиіліктерінің қосындысы 67.604670 мәтін бойындағы екіәріптік тіркестердің 100 пайызын құрайды. Ең жиі тіркесетін әріп тіркестері ретінде ***ар, , ан, ал, ын, да, та, ен, ер, қа, де, ін, ла, ды, ты***... атауға болады. Осы сөздіктің жоғары жиілікті бөлігінен үзінді ретінде берілген 217 екіәріптік тіркес барлық мәтіннің 90 пайызын қамтиды. Статистикалық мәліметтер толық болу үшін біз өз зерттеуімізде сөз басында және сөз соңында кездесетінекі әріптік тіркестерінің де жиілік сөздіктерін құрастырдық.

Әліпби ауыстыруда маңызды мәселелердің бірі – әріптерді пернетақтада орналастыру. Әріптерді пернетақтада орналастыруда қандай ұстанымдарға сүйенеміз деген мәселенің басын ашып алу қажет. Оның *біріншісі* – пернетақтадағы ағылшынша жазатын латын қаріптерінің тұрған орнын сақтап, кирилдегі қазақша әріптер тұрған жоғарғы жақ қатарға жаңа таңбаларды орналастыру. Бұл ағылшын пернетақтасында тұрған латын қаріптерін басуда шатастырмау үшін тиімді деуге болады. *Екіншісі* – ағылшын пернетақтасындағы латын қаріптерін сақтай отырып, жоғары жаққа емес, негізгі үш қатардағы бос орындарға жаңа таңбаларды қою. *Үшіншісі* – әлемдік тіл білімінде ұстанылатын принцип – пернетақтада әріптерді қолданыста жиі кездесуіне қарай орналастыру, яғни жиі кездесетін әріптерді пернетақтада басқанда оңай, икемді басылатын жерлерге қою болып табылады.

Әріптерді пернетақтада тиімді орналастыру мәселесіне қатысты А.Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы лингвистика бөлімі қолданыстағы әріптердің кездесу жиілігін анықтады.Осы жиілік талдау (әріп және әріп

тіркесінің жиілік сөздіктері) нәтижесі бойынша қазақ әріптері мен әріп тіркестерінің жиі кездесуі ескеріліп орналастырылған пернетақтаның пилоттық жобасы ұсынылды.

Пернетақтада әрбір әріптің кирилдегі және латындағы баламалары келтірілді. Пернетақтаның ортаңғы бөлігінде жиілік сөздіктегі ең жиі қолданылған әріптер орын алды, ал олардың оң жақтарында екіәріптік тіркес жиілік сөздігінен алынған әріптер орналастырылды. Пернетақтаның шеткі жақтарына сирек кездескен әріптер орналастырылды. Осы әріптер орналастырылған пернетақтада теруге арналған арнайы компьютерлік бағдарлама жасалды. Жиілік талдау бойынша орналастырылған латын әріптері бірнеше пернетақта батырмасына жапсырылып, бұл пернетақта нұсқасының мәтін теруге ыңғайлы, тиімділігі уақыт пен сапасына қарай сынақтан өткізілуде.

Фонетика саласындағы буын теориясы ең күрделі және маңызды мәселелердің бірі деуге болады. Буынның жиілік сөздігін алу басқа тілдік бірліктерге қарағанда күрделірек. Өйткені қазақ тілінде сөздерді буынға бөлу айтылымда жақсы ажыратылғанмен, кирилл қарпінің әсерінен көптеген сөздерімізді буын жігін сақтамай жазу нормаға түскендіктен, ауызша буынды басқаша дыбыстаймыз, жазуда басқаша жазамыз. Сондықтан сөздерді буынға бөлу жазуға бағына бермейді.

Қ.Жұбановтың “Буын жігін қалай табуға болады?”* атты мақаласында буындардың моделі арқылы “буын жігін оңай, механик түрде табуға болады” – дейді. Қазақ тіл білімінде белгілі ғалым Қ.Бектаев қазақ буындарының жиілік сөздігін 1973 жылы шыққан «Қазақ тексінің статистикасы» атты жинақта береді.

Қазақ буындарының жиілік сөздігін қазіргі заманауи және қазақ тілінің барлық стильдерін қамтитын мәтіндерден алу мақсатында «Қазақ тіліндегі буындардың жиілік сөздігін» ұсынып отырмыз. Бұл жиілік сөздікті жасауда буын жігін

*Мақала алғаш рет “Ауыл мұғалімі” журналының 1934 жылғы 2-санында жарияланған.

автоматты жолмен ажырату қиын болғандықтан, Қолданбалы лингвистика бөлімінің серверде орналастырылған «Қазақ тілінің мәтіндер корпусынан» қазақ тілінің бес стилі бойынша жасалған 6 миллион сөзқолданыстан тұратын мәтіндер жинағынан әр стиль бойынша 10 000 сөзқолданыстан, барлығы – 50 000 сөзқолданыстан тұратын мәтін алынып, «қолдап» буынға бөлінді, яғни «Қазақ тілі буындарының жиілік сөздігін» құрастырудан бұрын көлемі 50 мың сөзқолданыстан тұратын мәтін бойынша сөзформалардың жиілік сөздігі құрастырылды. Бұл сөзформалардағы сөзтізбенің рет саны 32304 сөзформаға тең.

Бұл «Қазақ тіліндегі буындардың жиілік сөздігі» дәстүрлі типі бойынша үш түрлі нұсқада жасалды. «Қазақ тіліндегі буындардың әліпбилі-жиілік сөздігі», «Қазақ тіліндегі буындардың жиілікті-әліпбилі сөздігі» «Қазақ тіліндегі буындардың кері әліпбилі-жиілік сөздігі».

Кітапта берілген «Қазақ тілі буындарының жиілік сөздігінде» қазақ тіліндегі буындардың реттік саны – 1225, олардың қайталана қолданылуы негізінде буындардың қосынды жиілігі 124063-ке тең. Буынның жиілік сөздігі бойынша, қазақ тілінде сан жағынан бітеу буындар (үш және төрт әріпті) көп екен. Жоғарыда 1225 реестрдің 952-сін құрап, 40 пайызын осы бітеу буындар алады. Алайда бітеу буындардың саны көп болғанмен, ашық буындар кездесу жиілігі жағынан бітеу буындарды басып озады. Мәселен, жоғарыдағы буындардың жиілікті-әліпбилі сөздігіндегі жоғары жиіліктегі 26 буынның 25-і ашық буындар. Ашық буындар реестр саны бойынша 173 болғанмен, мәтінде қайталана қолданылып, мәтіннің 55 пайызға жуығын құрайды. Ал тұйық буындар реестр саны бойынша 100-ге жетпейді және жоғары жиіліктегі 26 буын реестрінің құрамында кездеспейді. Мәтінде қайталанып қолданыла келе, мәтіннің 4,24% пайызын ғана құрайды. Үш әріпті тұйық буындарға қарағанда екі әріпті тұйық буындар көп, жиілік сөздікте үш әріпті тұйық буындар 9 ғана болса, 91-і екі әріпті тұйық буындар, екеуі қосылып 100 реестрлік қатарды құрайды.

Қайсібір тілде болмасын сөздер жалғыз өздері тұрып, бір-бірімен байланыспаса, ой дұрыс жеткізілмейді. Сөз бен сөзді

байланыстырудың құралы тілдегі грамматикалық бірліктер болып табылады. Грамматикалық бірліктер (қосымшалар, көмекші сөздер) түбір сөздерге жалғанып не тіркесіп келіп, сөздерді байланыстырады. Тілдегі лексемалар сияқты қосымшалардың да байланысу тәсілдері мен қолданыс аясы бар. Сөз түрлендіруші қосымшалар әдетте сөздерге талғамай жалғанады да, сөздерді түрлендіреді. Осындай сөз түрлендіруші қосымшалардың саны (варианттарын қоса) мен түрленім формалардың санын және олардың мәтіндердегі қолданыс жиілігін анықтаудың маңызы зор.

Ә.Ахабаев пен Қ.Бектаев зат есімдердің, Ә.Ахабаев есімдіктердің морфологиялық құрылымына талдау жасаса, С.Мырзабеков туынды түбір туынды етістіктердің, Қ.Бектаев пен С.Мырзабеков етістіктің форма өзгерткіш аффикстеріне, А.Белботаев ғылыми-техникалық стильдегі сын есім аффиктеріне, А.Жұбанов пен Е.Жұбанов «Абай жолы» романындағы сын есімдердің лексика-морфологиялық формаларына статистикалық талдаулар жасады.

Қазақ тіліндегі түрленім қосымшаларының, яғни форма тудырушы қосымшалардың жиілігі Қолданбалы лингвистика бөлімі құрастырған «Қазақ тілінің мәтіндер корпусы» деп аталатын 6 миллионнан тұратын бес стильден алынған корпус мәтіндерінен алынды.

Мәтіндерден қосымшалар жиілігін алу үшін мәтін алдымен морфологиялық анализатордан өткізілді. Морфологиялық анализатор автоматты түрде мәтіндегі сөздерді түбір мен қосымшаға бөліп, содан соң лемматизацияланған (түбірі бөлінген) бөліктен бөлек екінші жағындағы сөзтүрленім формаларын арнайы қосымшалар кестесі бойынша бөлшектеп және кестеде берілген шартты грамматикалық белгіленімдері бойынша сипаттап шығады. Содан кейін бірдей қосымша бөліктерін грамматикалық сипаттамаларымен бірге топтастырып, бірдей формалардың санын шығарады. Ұсынылып отырған қосымшалар сөздігі осылай автоматты әдіспен алынып отыр. Автоматты әдіс болғандықтан, омонимдерге қатысты кейбір нақты емес жағдайлар кездесуі мүмкін. Алайда 95 пайыз дұрыс ақпарат береді деп ойлаймыз.

«Қазақ тіліндегі қосымшалардың жиілік сөздігі» әдеттегідей үш типте жасалды: «Қазақ тіліндегі қосымшалардың әліпбилі-жиілік сөздігі», Қазақ тіліндегі қосымшалардың жиілікті-әліпбилі сөздігі», «Қазақ тіліндегі қосымшалардың кері әліпбилі-жиілік сөздігі».

Біз қарастырған мәтіндерде түрленім қосымшалардың 26 түрі айқындалды. жоғарыдағы статистикалық мәліметтерді қорытындыласақ: қосымша атауларының 12-сі жоғары жиілікте кездеседі. Олар: ***тәуелдік жалғау, табыс септік, есімше, көсемше, жіктік жалғау, барыс септік, көптік жалғау, жатыс септік, ілік септік, тұйық етістік, шығыс септік, болымсыздық жұрнақ***түрлері қайталана қолданыла келіп, мәтіннің 93,74 пайызын қамтыған. Осы қосымша атаулары ішінде жиілігі жағынан ***тәуелдік жалғау, табыс септік, есімше, көсемше*** түрлері мәтіннің 55,33 пайызын қамтып, қазақ тілінде ең жиі кездесін қосымшалар екендігін көрсетеді.

Түрленім формалар дегеніміз – сөздің түбіріне (негізгі және туынды) жалғанатын қосымшалардың жеке немесе бірінің үстіне бірі жалғануындағы сөздің түрленуі.

Түрленім формалардың жиілік сөздігі де жоғарыда аталған 6 миллиондық корпус мәтінінен автоматты жолмен алынды. Олар: «Қазақ тіліндегі түрленім формаларының әліпбилі-жиілік сөздігі», Қазақ тіліндегі түрленім формаларының жиілікті-әліпбилі сөздігі», «Қазақ тіліндегі түрленім формаларының кері әліпбилі-жиілік сөздігі».

Жоғарыдағы түрленім формалардың жиілік сөздіктерінде түрленім форма саны – 2945. Абсолютті жиілігі ең жоғары түрленім формалар тәуелдік жалғаулары болып тұр. Мәселен -і/ТЖ 6 миллион сөзқолданыстан тұратын мәтінде 335313 рет кездессе, жуан варианты -ы/ТЖ 160419 рет кездесіп, екеуі қосылып, бүкіл мәтіннің 10 пайызын құрайды. Бұл жерде түрленім форманың жиілік сөздігі болғандықтан байқайтынымыз – тәуелді жалғаудың 3 жағының жеке дара қолданысының жиі кездесуі.

Түрленім формалардағы жиілік негізінен жалаң қосымшалар жиілігімен шамалас. Бұдан қазақ тілінде түрленім формалары дара, жалаң тұлғада жиірек кездесетінін байқаймыз.

Сонымен қатар, қосымшалардың түрленім формалардың жиілік сөздігіндегі реестр бойында кездесуіне қарай сан жағынан ең бірінші орынға ие болып тұрған түрленім формасы – ***жіктік жалғау (ЖЖ)*** реестрде 538 рет кездесіп, 2945 реестрлік бірліктің ***18,27*** пайызын қамтиды. Бұдан қазақ тілінде сөздер жіктік жалғауымен жиі түрленіп қолданылатынын көреміз, яғни бұл жіктік жалғауының предикаттық қызметімен тікелей байланысты болса керек.

Сонымен бірге ***жіктік жалғаудан*** басқа да реестр бойында сан жағынан жоғары сатыдағы түрленім формаларына жататындар: ***табыс септік, тәуелдік жалғау, барыс септік, ілік септік, көмектес септік, шығыс септік, жатыс септік, салыстыру форма*** қосымшалары екені анықталды. Бұл қазақ тіліндегі септік жалғауларының сөз байланыстырушылық қызметімен байланысты. Ал тәуелдік жалғаудың реестрдегі түрленім формалар құрамында көп кездесуі оның жалпы жиілік сөздіктегі бірінші орынға шығуына да ықпал еткен. Ал салыстыру формасының реестр бойындағы жоғары жиіліктегі топқа енуі *-дай/дей* қосымшасының кез келген сөзге жалғанып, сөз бойында жалғаулар сияқты еркін қолданысымен байланысты.

29-кестедегі статистикалық мәліметтер бойынша тұжырым жасайтын болсақ, жіктік жалғау түрленім форма реестрінде жиі кездесуі жағынан, яғни реестрді қамту пайызы жағынан ең жоғары жиілікті көрсетіп, тізімде бірінші болып тұрған болса, мәтінді қамту пайызы жағынан жоғары жиілікті тәуелдік жалғауы алып кетеді. Тәуелдік жалғаулары 6 миллион сөзқолданыстан тұратын мәтіннің 18,80 пайызын құрайды. Ал табыс септігі 12, 23 пайызын, жіктік жалғау – 9,48 пайызын, барыс септігі – 8,83 пайызын қамтиды. Осы төрт түрленім қосымша қосылып мәтіннің 50 пайызға жуығын қамтиды екен.

Қазіргі кезде қазақ лексикологиясының аса бір маңызды мәселесінің бірі – қазіргі қазақ тілінің қолданыстағы лексикалық жүйесінің шекарасын айқындау. «Қазақ тілінің жиілік сөздігі», ең алдымен, жиі қолданыстағы сөздік қорымыздың шекарасын анықтап алу үшін өте қажет.

Қазақ тіліндегі сөздердің жиілік сөздіктерінің статистикалық деректері қазақ тілінің лексикалық құрамын жан-жақты зерттеуде салыстырмалы талдауды жүзеге асыруға, сөз табына қатысты нақты фактологиялық мәліметтер беруге, тұтастай алғанда, қазақ тілінің көптеген теориялық мәселелерін шешуге мүмкіндік туғызады.

70-жылдары Қ.Бектаевтың басшылығымен және жетекшілігімен А.Қ.Жұбанов, С.Мырзабеков, А. Белботаев, Ә.Ахабаев т.б. зерттеушілер жиілік сөздіктер құрастыру ісімен айналыса бастады. Абай тілі сөздігінің жиілік сөздігі, М.Әуезовтің 20 томдық шығармалар тексінің жиілік сөздіктері, сонымен қатар, ертегі тілінің, публицистика тілінің, математика терминдерінің жиілік сөздіктері жарық көрді.

Осы жиілік сөздіктерден кейін қазақ тілінде жиілік сөздіктер құрастыру ісі тоқтап қалды. 2012-2014 жылдары А.Байтұрсынұлы атындағы ТБИ Қолданбалы лингвистика бөлімі аз көлемдегі қаржымен «Қазақ тілінің квантитативтік құрылымы (Қазақ тілінің жиілік сөздігі) деген гранттық тақырып аясында бұрын жарық көрген сөздіктердің басын біріктіру арқылы әртүрлі стильдерді қамтитын біртұтас жиілік сөздікті құрастырып, баспаға ұсынды.

2016 жылы Білім және ғылым министрі Сағадиевтің тапсыруымен мемлекеттік тілді оқытудың тиімді әдістемесін жасау мақсатында лексика-грамматикалық минимумдар үшін қажетті қазақ тілінің жиілік сөздігін құрастыру ісі А.Байтұрсынұлы атындағы Тіл білімі институтына тапсырылған болатын. Осы тапсырма бойынша институт қысқа мерзімнің ішінде «Жалпы білім берудегі қазақ тілінің жиілік сөздігі» жарық көрді.

«Жалпы білім берудегі қазақ тілінің жиілік сөздігін» құрастыруға көркем әдебиет, публицистикалық стиль, ғылыми, публицистикалық, сөйлеу стильдер мәтіндері қамтылған, яғни қазақ тілінің 5 функционалдық стилі сөздіктің базасын құрайды. 36 265 сөзтізбеден(сөз, лексикалық бірлік) тұратын бұл сөздік дәстүрлі сөздіктер түрлері болып табылатын үш түрлі құрылымда (Әліпбилі-жиілік сөздік; Кері әліпбилі-жиілік сөздік; Жиілікті-әліпбилі сөздік) жасалған. Сонымен қатар бұл сөздік әрбір стильдер бойынша да «Таралым сөздігінде» (Распределительный словарь) жіктеліп берілген.

Жиілік сөздік жасаудың әдіс-тәсілдері әртүрлі. Соның бірі – мәтінді сөзформалар сөздігіне түсіріп, сөзді түбірге келтіру. Екіншісі – мәтінді морфологиялық анализатор арқылы өткізіп, түбір мен оған қойылатын сөз табы белгісімен бірге автоматты

жолмен алу. Біріншісі көп қол жұмысын керек етеді. Мұндайда қосымша жалғанғанда түбір сөздердің омонимдерін ажырату мүмкін болса, түбір формадағы сөздерге келгенде олардың сөз табына қатысын анықтау қиын болады.

Екіншісі оңай, тез жүзеге асырылғанмен, жүз пайыздық шынайылықты бермейді. Бұл әдісте те омонимдер автоматты түрде ажыратылмайды. Автоматты жолмен омонимдері ажыратылған жиілік сөздіктерді тек омоним ажыратылған мәтіндерден ғана алуға болады.

Осы орайда Қолданбалы лингвистика бөлімі қызметкерлері проза жанры (М.Әуезов, Ә.Кекілбаев т.б.) бойынша 500 000 сөзқолданыстан тұратын мәтіндерге «қолдап» морфологиялық белгіленім қойған болатын. Біз ұсынып отырған жиілік сөздік осы сөз таптары қолмен қойылған, омонимдері ажыратылған, яғни омоним жоқ мәтіннен алынды. Сөздің жиілік сөздігі бойынша әртүрлі сипаттағы жиілік талдаулар жасалды. Мәселен, 500 000 сөзқолданыстан тұратын мәтінде Қ және Т әрпінен басталатын сөз саны 1019 реестрлік сөзді құрап, барлық реестрлік сөздер ұзындығының ***22,539 пайызын*** қамтиды екен. Сөз басында кездесетін әріптердің (осы әріптен басталатын сөздердің) мәтіндегі абсолюттік жиіліктері реестрдегі кездесу санынан басқаша. Бұл статистикалық мәліметтер бойынша, қазақ тілінде мәтін ішінде сөз басында жиі қолданылатын әріп – **Б,** одан кейін – **Қ**. Одан кейін – **К, Ж, А, Т, С, Д, Е, О.** Осы 10 әріп мәтінде сөз басында жиі қолданыла келе, 500 сөзқолданыстан тұратын мәтіннің 80 пайызға жуығын құрайды.

Сөздіктердің басым көпшілігінде сөздер алдыңғы дыбыстар бойынша әліпби тәртібімен орналасса, кері әліпби сөздікте, керісінше, сөз бен сөзтұлғалар соңғы әріп-дыбысынан басталып әліпби тәртібіне келтіріледі және сөздердің соңғы жағынан тегістеліп беріледі. Сөздіктің бұл түрі «Кері әліпби-сөздік» немесе тек ***«Кері сөздік»*** деп аталып жүр. Жиілік сөздіктің кері әліпби түрінде реестрлік сөзге (сөзтұлғаға) қосылған біркелкі жалғаулар мен жаңа сөз жасаушы немесе сөз түрлендіруші жұрнақтар бір жерге жинақталып берілуі тілді зерттеушілер үшін аса құнды тілдік мәліметтер болып табылады.

Сөздердің кері-әліпбилі жиілік сөздігі бойынша сөз таптарының реестр бойын қамтуына қарай зат есім сөздер көп кездеседі. 4521 реестр сөзден тұратын бұл жиілік сөздікте зат есімдер 1753 реестр сөзді, ал етістіктер 1236 сөзді құрайды. Екеуі қосылып, рестр бойынның 40 пайызға жуығын алады екен. Одан кейінгі орында сын есімдер тұр.

Ал сөз таптарының мәтінде қайталана қолданыла келе, абсолютті жиіліктері бойынша жоғарыдағы сөз таптарының жиілік дәрежесі өзгереді. Мәтінде жиі қолданылуы бойынша етістіктер зат есімдерді басып озып, алдыңғы орынға шығып кетеді. Демек, реестр бойын қамту пайызы жағынан зат есімдер алдыңғы орында, мәтінді қамту пайызы бойынша, етістіктер алдыңғы орында, ал екеуінде де сын есімдер үшінші орынды иеленеді. Модаль сөздер, көмекші есімдер, еліктеуіш сөздер, одағай сөздер сирек кездесетіндерге жатады.

Сөздіктердің «жиілік сөздік» түрі әліпбилі-жиілік сөздіктен басқаша, дәлірек айтқанда, әр сөздің (сөзтұлғаның) қолдану жиілігінің дәрежесіне қарай орналасады: ең бірінші ретте орналасатын мәтіндегі ең жиі қолданылған сөз (не сөзтұлға), екінші, үшінші (тағы тағылар) ретте орналасатын сөздер – кездесу жиіліктері бірте-бірте кеміп отыратын сөздер (сөзтұлғалар).

Сөздердің жиілікті-әліпбилі сөздігі бойынша, ең жиі кездесетін «***бол***» етістігі мәтін ішінде ***666 рет*** қолданып, жалғыз өзі-ақ мәтіннің ***2,320*** пайызын қамтыған. *Бол* етістігінің ең жоғарғы жиілікті көрсетуі бұдан бұрын шыққан сөздіктерде де орын алған. ***Бол*** етістігінің жоғары жиілікті көрсетуі оның көмекші етістік болып, грамматикалық мағына үстеуші қызметімен тікелей байланысты деуге болады. Өйткені жиі қолданылу, грамматикалық абстракция жасайтын грамматикалық бірліктерге тән қасиет. Жиілік сөздікте ***бол*** етістігі сияқты жоғары жиілікте тұрған көп сөз осы көмекші етістіктер. Мысалы, *де, еді, кел, ал, тұр, отыр, жатыр, көр, айт, бер, шық, кет, жүр, бар, түс, қыл, сал, қой, баста* т.б. Сол сияқты жиілік сөздікте шылаулардан *да, де, мен, бірақ, ғой, қарай, ғана, және, тағы* шылаулары, сондай-ақ есімдіктер (*бұл, өз, сол, ол, осы, бар, мен*) жиі кездескен. Бұл жиілік сөздік тек көркем шығарма мәтіндерінен алынғандықтан, термин сөздер аз, тіптен кездеспейді десе де болады.

ӘДЕБИЕТТЕР

1. Статистика казахского текста. – Алматы, 1973. – С.614-629.

2. Қазақ сөзформа құрылымыныңжиілік сөздігі. – Алматы, 2017. – 552 б.

3. Жұбанов А., Тоқмырзаев Д., Жаңабекова А. Қазақ жазуын латын қарпіне көшірудің статистикалық негіздері (жинақ). – Алматы, 2017. – 160 б.

4. Жалпы білім берудегі қазақ тілінің жиілік сөздігі. – Алматы, 2016. – 1472 б.

Ishmukhametova A.Sh.
Institute of History, Language and Literature USC RAS,
Russia, Bashkortostan, Ufa

DIALECT VARIANTS OF THE LEXEME '*HAWTHORN*' (using materials from the dialectological base)

Abstract. Dialect is one of the most important sources for studying the historical development and formation of a literary language. the dialectological database created by the laboratory of linguistics and information technologies as part of the machine fund of the bashkir language consists of three separate databases: the lexical database, the database of the dialectological atlas and the textual database. Using the materials of the dialectological base, the lexeme 'hawthorn', which is a medicinal plant, is considered. More than 50 dialect variants have been identified, and the territory of distribution is clearly presented.

Keywords: *dialectological base, Machine Fund of the Bashkir language,the Bashkir language, lexeme, plant vocabulary, hawthorn, etymology.*

Ишмухаметова А.Ш.
Институт истории, языка и литературы УФИЦ РАН,
Россия, Башкортостан, Уфа

ДИАЛЕКТНЫЕ ВАРИАНТЫ ЛЕКСЕМЫ *ЭНӘЛЕК* 'БОЯРЫШНИК' (на материале диалектологической базы)

Аннотация. Диалект является одним из важнейших источников для изучения исторического развития и становления литературного языка. В диалектах отражается особенности национального сознания. Созданная лабораторией лингвистики и информационных технологий диалектологическая база в составе Машинного фонда башкирского языка состоит из трех отдельных баз данных: лексической базы данных, базы данных диалектологического атласа и текстологической базы данных. С

применением материалов диалектологической базы рассмотрена лексема *энәлек ‘боярышник’*, которая является лекарственным растением. Выявлено более 50 диалектных вариантов, наглядно представлено территория распространения.

Ключевые слова: *диалектологическая база, машинный фонд башкирского языка, лексема, растительная лексика, боярышник, этимология.*

Ишмөхәмәтова А.Ш.
РФА ӨФТУ Тарих, тел һәм әҙәбиәт институты,
Рәсәй, Башҡортостан, Өфө

ЭНӘЛЕК ЛЕКСЕМАҺЫ ҺӘМ УНЫҢ ДИАЛЕКТ ВАРИАНТТАРЫ (диалектология базаһы материалдары нигеҙендә)

Милли әҙәби тел – ошо телдә һөйләшеүсе бөтә кешеләрҙе, ошо телдең барлыҡ диалекттарын һәм һөйләштәрен берләштергән тел. Халыҡтың йәнле һөйләштәре, диалекттары хәҙерге башҡорт әҙәби теле менән тығыҙ бәйләнештә тора, артабан да үҫә, тәрәнәйә бара. Диалект материалдары әҙәби телдең тарихи үҫешен һәм формалашыуын өйрәнеүҙә мөһим бер сығанаҡ булып тора. Халҡыбыҙҙың тел байлығын ғилми яҡтан ентекле өйрәнеүҙә башҡорт ғалимдарының: Ғ.С. Амантаев, Т.Ғ. Байышев, Ғ.Й. Дәүләтшин, Т.Ғ. Байышев, Ж.Ғ. Кейекбаев, Н.Х. Мәҡсүтова Н.Х. Ишбулатов, С.Ф. Миржановаларҙың индергән өлөшө баһалап бөткөһөҙ [1; 11; 13; 15; 16; 17]. Улар күп һанлы экспедициялар барышында диалекттарҙы һәм һөйләштәрҙе тасуирлап, таралыу сиктәрен билдәләүҙән тыш, телдең фонетик, лексик һәм морфологик кимәлдәрендә әҙәби тел һәм диалекттарҙың үҙ-ара мөнәсәбәт проблемаларын да тикшерә.

Йәнле һөйләү телен өйрәнеүҙә, халыҡтың һүҙ байлығын бер урынға туплауҙа корпустарҙың әһәмиәтен һәм ролен күҙ уңында тотоп, лингвистика һәм мәғлүмәт технологиялары лабораторияһы коллективы тарафынан Башҡорт теленең машина фонды эсендә диалектология базаһы булдырыла [6]. Ул

3 үҙаллы мәғлүмәт бүлектәренән: лексика, диалектологик атлас һәм текстар базаһынан тора. Базаға индерелгән материалдар һәр бер диалекттың, һәр бер һөйләштең фонетик, лексик һәм грамматик үҙенсәлектәрен сағылдырырға мөмкинлек бирә [9; 12; 19].

Лексика базаһында 1967, 1970, 1987, 2002 йылдарҙа баҫылып сыҡҡан диалект һүҙлектәренән алынған мәғлүмәт тупланған [2; 3; 4; 5] һәм ул 52 000-дән артыҡ берәмекте үҙ эсенә ала. Диалект һүҙ, уның ниндәй һүҙ төркөмөнә, ниндәй диалектҡа, һөйләшкә ҡарауы, әҙәби нормаһы, руссаға тәржемәһе бирелә. Кәрәкле һүҙҙе ошо һанап кителгән билдәләре буйынса эҙләү мөмкинлеге лә бар.

Диалектологик атлас базаһының нигеҙендә 2005 йылда сыҡҡан «Башҡорт теленең диалектологик атласы» материалдары, йәғни башҡорт теле диалекттарының тарихи үҫешендә уларҙың таралыу төбәктәрен күрһәткән системалаштырылған диалектологик карталар йыйылмаһы ята [10]. Атластың нигеҙен 1973–83 йылдарҙа Н.Х. Мәҡсүтова етәкселегендәге Тарих, тел һәм әҙәбиәт институтының тел белгестәре коллективы (С.Ғ. Ғилманова, М.И. Дилмөхәмәтов, У.Ф. Нәҙерғолов һ.б.) тарафынан әҙерләнгән материалдар тәшкил итә. Башҡортостан Республикаһында һәм уға сиктәш Ҡурған, Свердловск, Силәбе, Һамар, Һарытау, Ырымбур өлкәләре, Пермь крайында йәшәүсе башҡорттарҙың теленең фонетикаһы, морфологияһы, лексикаһы һәм синтаксисының территориаль үҙенсәлектәре сағылдырыла, 400 башҡорт ауылынан йыйылған экспедиция материалдарының 250-һе картаға индерелә.

Текстар базаһында халыҡтың йәнле һөйләү өлгөләре урынлашҡан. Бөгөнгө көндә 500-гә яҡын башҡорт теленең төрлө һөйләштәренән текстар индерелгән. Мәғлүмәтте ниндәй диалект, һөйләш, яҙып алынған йылдар, информанттың белем кимәле, йәше, ниндәй енестән, милләттән булыуынан ҡарап эҙләү мөмкинлеге бирелә [7].

Дөйөм алғанда, башҡорт теленең диалект һәм һөйләштәрен фонетик, лексик, грамматик, стилистик йәһәттән өйрәнеү, төрлө

тикшеренеүҙәр үткәреү өсөн диалектология базаһына бик бай материал тупланған.

Һуңғы осорҙа шифалы үҫемлектәр менән ҡыҙыҡһыныусыларҙың һаны артыуын күҙәтәбеҙ. Ҡырағай үҫемлектәр менән файҙаланыу бик борондан килә. Бөйөк рус ғалимы И.П. Павлов әйткәнсә, кешелек донъяһы барлыҡҡа килгәндән алып кеше дауаланыуға мохтаж булған. Сөнки һәр төрлө ауырыуҙар уны шул замандан уҡ һағалай башлаған. Ауырыуҙы дауалау өсөн ғәҙәттә шифалы үләндәр файҙаланылған. Быуаттар ағымында халыҡ үҙ тәжрибәһенән сығып, дарыу үҫемлектәрен өйрәнә килгән. Үләндәр менән эш иткәндә «самаһын белһәң — дарыу, белмәһәң – ағыу», тиҙәр. Башҡортостанда үҫкән ҡырағай үҫемлектәрҙең ике йөҙҙән ашыуы халыҡ медицинаһында ҡулланылыуы билдәле. Дарыу үҫемлектәренең лексикаһын тикшереү, диалекттарҙа нисек әйтелешен өйрәнеү тел тарихын өйрәнеүселәр өсөн дә, халыҡ тарихын һәм этнографияһын өйрәнеүселәр өсөн дә берҙәй әһәмиәтле, сөнки лексиканың был ҡатламы халыҡтың мәҙәниәтен, тик шул ерлеккә генә хас үҙенсәлеклектәрен сағылдыра.

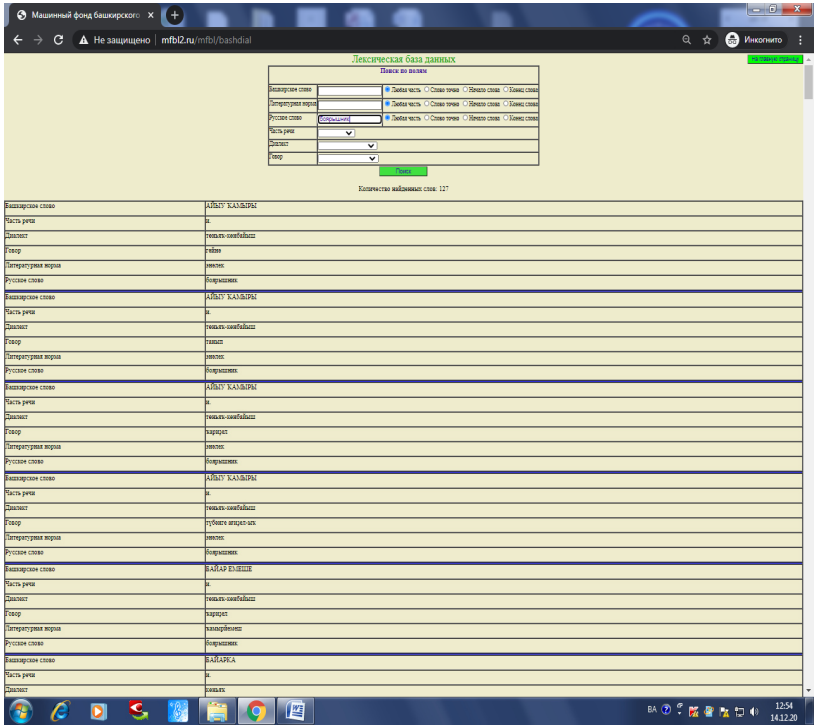
Энәлек, энәте, еҙәй, дунала — энәле ҡыуаҡ йәки ағас. Бейеклеге 4 м-ға етә. Май айында аҡһыл һарғылт сәскә ата, август-сентябрҙә емеше өлгөрә. Энәлек йылға буйҙарында, урман ситендә үҫергә ярата. Республикабыҙҙа киң таралған был үҫемлекте күпселек райондарҙа осратырға мөмкин. Дауалау өсөн энәлектең емеше һәм сәскәһе ҡулланыла. Энәлек емешенән (йәки сәскәһенән) әҙерләнгән төнәтмәне йөрәк эшсәнлеген яҡшыртыу, йөрәк тулауын (йыш-йыш тибеүен) һәм ауыртыуын баҫыу, баш мейеһендә ҡан йөрөшөн яҡшыртыу өсөн, атеросклероз булғанда, ҡан баҫымы күтәрелгәндә (гипертония ауырыуы), йоҡоһоҙлоҡ йонсотҡанда, тын ҡурылғанда ҡулланырға кәңәш ителә [8: 123].

Башҡорт теленең машина фондының лексикография бүлегенә ингән һүҙлектәрҙән был һүҙҙең *энәлек, һыуһар, дунала, энәте, еҙәй, ҡамыраш* булып медицина, умартасылыҡ, урман эше, ауыл хужалығы терминдары буйынса төҙөлгән ике телле һүҙлектәрҙә теркәлеүен асыҡларға була. В.З. Ғүмәров «Урыҫса-

башҡортса медицина терминдары һүҙлеге»ндә емеше йөрәктән тыныслandырғыс ағас тип бирә. «Урыҫса-башҡортса урман эше терминдары һүҙлеге» (З.Ғ. Ураҡсин, З.Ғ. Уразбаева, Н.Ф. Суфьянова) был ҡыуаҡтың бер нисә төрөн билдәләй: ‘боярышник колючий’ *сәнскеле еҙәй*, ‘боярышник восточный’ *көнсығыш еҙәйе*, ‘боярышник кровавокрасный’ *әнәлек, әнәте, еҙәй*, ‘боярышник обыкновенный’ *ғәҙәттәге еҙәй*, ‘боярышник сибирский’ *Себер еҙәйе* [14].

Диалектология базаһының лексика бүлегендә *әнәлек* һүҙенең әҙәби телдә тағы ике варианты *ҡамырйемеш* һәм *дунала* формаһында йөрөүе күрһәтелә. Төньяҡ-көнбайыш диалекттың ғәйнә, танып, ҡариҙел, түбәнге ағиҙел-ыҡ һөйләштәрендә – *айыу ҡамыры*; көнсығыш диалекттың ҡыҙыл һөйләшендә – *беҙ ағасы*; көнсығыш диалекттың әй, салйоғот, арғаяш, урта урал, көньяҡ диалекттың һаҡмар, эйек-һаҡмар һөйләштәрендә – *әнәғас*, көнсығыш диалекттың мейәс, салйоғот, арғаяш, әй, көньяҡ диалекттың дим, урта, һаҡмар һөйләштәрендә – *әналағас/әнәлағас/әнәлеғас*; көнсығыш диалекттың мейәс, урта урал, әй һөйләштәрендә – *әнәде / әнәте / әнәле*; шул уҡ диалекттың ҡыҙыл һөйләшендә – *әнәйемеш*; урта урал, мейәс, әй, салйоғот, арғаяш, ҡыҙыл һөйләштәрендә, көньяҡ диалекттың дим, урта һөйләштәрендә, төньяҡ-көнбайыш диалекттың ғәйнә, танып, ҡариҙел, түбәнге ағиҙел-ыҡ һөйләштәрендә – *әнәлек*; көнсығыш диалекттың мейәс һөйләшендә – *әнәлекәй*, арғаяш һөйләшендә – *әнәтештеғайын*, *әнәтештекәй*, урта урал һөйләшендә – *әнәткес*, төньяҡ-көнбайыш диалекттың ғәйнә, танып, ҡариҙел, түбәнге ағиҙел-ыҡ һөйләштәрендә – *эт емеше* формаларында осрай. *Ҡамыр жиләге / ҡамыр еләге / ҡамыр жиләк* – көньяҡ диалекттың дим һөйләшендә, төньяҡ-көнбайыш диалекттың миңзәлә, ҡариҙел һөйләштәрендә; *ҡамыр жимеше / ҡамырйемеш* – төньяҡ-көнбайыш диалекттың түбәнге ағиҙел-ыҡ, ҡариҙел, танып, ғәйнә, көнсығыш диалекттың урта урал, әй, көньяҡ диалекттың дим һөйләштәрендә, *ҡамырлама* – көньяҡ диалекттың дим, *ҡамырлауыҡ* –төньяҡ-көнбайыш диалекттың танып, ҡариҙел, түбәнге ағиҙел-ыҡ, ғәйнә, көнсығыш диалекттың урта урал, *ҡамырлыҡ* – төньяҡ-көнбайыш диалекттың түбәнге ағиҙел-ыҡ, ҡариҙел, танып, ғәйнә,

көнсығыш диалекттың урта урал һөйләштәрендә ҡулланыла. *Баяр емеше* – төньяҡ-көнбайыш диалекттың ҡариҙел һөйләшендә; *байарка* – көньяҡ диалекттың урта, көнсығыш диалекттың мейәс, арғаяш һөйләштәрендә; *дегәнәк / дегәнәк йемеше* – көнсығыш диалекттың урта урал һөйләшендә; *еҙ емеш / еҙем / емшегән / йөҙөм* – көньяҡ диалекттың эйек-һаҡмар, *еҙағас, еҙәй* –туҡ-соран, эйек-һаҡмар; *йушан* –ырғыҙ; *йәбешкән* – дим һөйләштәрендә; *йәүһәр* – көнсығыш диалекттың әй һөйләшендә әйтелә (*1-се һүрәт*).



1-се һүрәт. Рус һүҙе ‘*боярышник*’ буйынса
лексика базаһынан эҙләү

Әҙәби телдә *дунала* формаһы алынып, башҡорт теленең һөйләштәрендә түбәндәге һүҙҙәр: *ҡара ҡайын* – төньяҡ-көнбайыш диалекттың түбәнге ағиҙел-ыҡ, ҡариҙел, ғәйнә, танып, минзәлә, көньяҡ диалекттың урта, дим; *мәмешкәк* – көньяҡ диалекттың дим, төньяҡ-көнбайыш диалекттың ҡариҙел; *тештекәй* – көнсығыш диалекттың арғаяш, салйоғот; *типкечле ағач* – төньяҡ-көнбайыш диалекттың ғәйнә; *төлкөйемеш, төлкөйемеш ағасы* – көнсығыш диалекттың әй; *һыуһар* – көнсығыш диалекттың әй; *әбей йемеше*, *әбей ҡамыр*– көньяҡ диалекттың эйек-һаҡмар, дим һөйләштәрендә ҡулланыла. Ҡайһы бер һүҙҙәр миҫалдар менән нығытылып кителә: *Байарка* неҙҙең йаҡта үҫәме? (урта урал һөйләше). *Төлкөйемеш* ағасының йемеше бик тәмде була (әй һөйләше). Һыу буйларында үҫә *әнәлек, әнәғас*, ҡарыйәңке талдан башҡа (эйек-һаҡмар һөйләше).

Алда әйтеп кителгәнсә, базала һайлау мөмкинлеге бар, мәҫәлән, көнсығыш диалектты ғына ҡарарға була. Унда был һүҙгә 50-гә яҡын миҫал сыға. Һөйләштәрҙәге әйтелеше бирелә: *байарка* (мейәс, арғаяш), *беҙ ағасы* (ҡыҙыл), *дегәнәк, дегәнәк йемеше* (урта урал), *дунала* (арғаяш, әй, мейәс), *йәүһәр* (әй), *ҡамыр жимеше* (урта урал), *ҡамыр йемеш* (урта урал, әй), *ҡамырлауыҡ / ҡамырлыҡ* (урта урал), *тештекәй* (арғаяш, салйоғот), *төлкөйемеш, төлкөйемеш ағасы* (әй), *һыуһар* (әй), *әнәғас / әнәғас* (әй, салйоғот, мейәс, арғаяш, урта урал), *әнәде / әнәте* (мейәс, урта урал, әй), *әнәйемеше* (ҡыҙыл), *әнәлағас /әнәлеғас* (мейәс, салйоғот, арғаяш, әй), *әнәле* (мейәс, әй), *әнәлекәй* (мейәс), *әнәтештеғайын* (арғаяш), *әнәтештекәй* (арғаяш). Урта урал һөйләшендә энәлектең сәнскеһе *әнәткес* тип йөрөтөүе лә күрһәтелә.

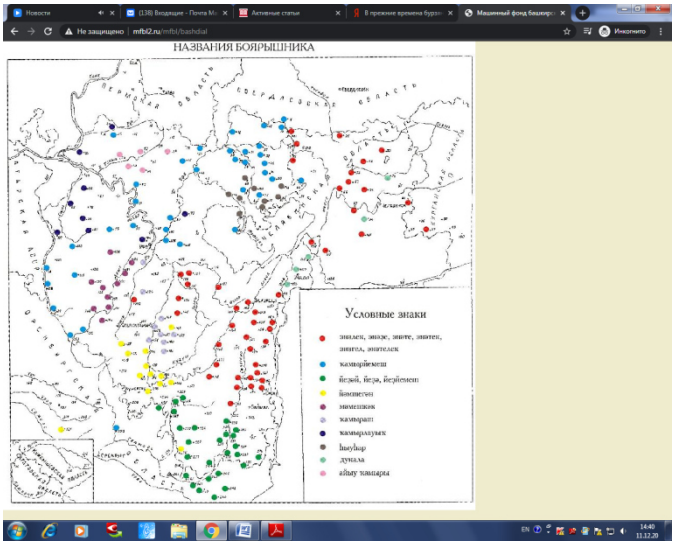
Әнәлек лексемаһы тикшереү исемлегенә ингән 250 ауылдың 35-ендә *ҡамырйемеш* тип йөрөтөлә. Ул Сведловск өлкәһе Түбәнге Һырға, Әртә, Красноуфимск, Ырымбур өлкәһе Октябрьский, Башҡортостан Республикаһының Яңауыл, Асҡын, Мәсетле, Ҡариҙел, Бөрө, Балаҡатай, Салауат, Благовар, Кушнаренко, Нуриман, Иглин, Туймазы, Бәләбәй, Йәрмәкәй, Миәкә, Бишбүләк районы ауылдарында теркәлгән (*2-се һүрәт*).

Тағы бер күренешкә иғтибар итергә мөмкин. *Дунала* – Силәбе өлкәһе Ҡоншаҡ районы Ҡорман ауылында, Сыбаркүл районы Атйетәр ауылында, Башҡортостандың Учалы районы Сәфәр, Һәйтәк, Рәсүл ауылдарында ҡулланыла. Совет тел белгесе, тюрколог, этимолог Э.В. Севортян етәкселегендә сыҡҡан «Төрки телдәренең этимологик һүҙлеге»ндә *әнәлектең*

башҡа телдәрҙә лә таралыуы сағылыш тапҡан: *дǫлəнə* үз.; *долана* уйғ. диал.; *dulane* (көнс.-төрк.); *толона* (шор.); *долоно* (ҡырғ.); *толоно* алт.; *тулана* тат. диал.; *долуна* уйғ.; *ḍɔlaqana* тув., тоф.; *долоҟуна*, *долуҟуна* яҡ.; *долоғоно* яҡ.; *долохоно* яҡ.; *дола*:на тув.; *долоғон* яҡ.; *толан* (шор.); *дунул* (осм.). Тат. диал., үз., уйғ. диал., алт., сағ., тув., тоф., яҡ. – ‘*боярышник*’, ҡырғ. – ‘*деревцо с шипами и желтыми ягодами*’, көнс.-төрк. – ‘*название растения*’, алт. – ‘*ягоды*’, алт., алт. диал., яҡ., тел. – ‘*терновник*’, шор. – ‘*какой-то кустарник с шипами*’ [18: 269–270]. Фин ғалимы, лингвист-тюрколог М. Рясянен soj. *dolaɣana*, *doläna* ‘боярышник’ kkir. *dolono*. осм. *dunul*. sor. *tolan*. *tolana*. tel. *tolono*, jak. *doloɣono* монг. *dologana* һүҙенән сыҡҡан тип бирә [20: 139]. Ҡәрҙәш төрки телдәрендә таралған *дөлəнə / дүлəнə / дулана / тулана* һәм башҡа фонетик варианттар метатеза күренеше һөҙөмтәһендә үҙгәреш кисереп, башҡорт телендә *дунала* булып киткәндер.

2-се һүрәт. Тикшереү исемлегенә ингән ауылдар теҙмәһе

Ҡыҙыҡһындырған һүҙҙең таралыу территорияһын аныҡыраҡ күҙаллау өсөн, мәғлүмәтте картанан ҡарау мөмкинлеге бар (*3-сө һүрәт*).



3-сө һүрәт. Картала *эҙәлек* һүҙенең диалекттарының таралыу территорияһы

Башҡорт теленең машина фондының диалектология базаһы нигеҙендә башҡорттарҙың дауалау үҙенсәләгенә эйә булыусы *эҙәлек, эҙәте, еҙәй, дунала* лексемаһының һөйләштәрҙә 50-нән ашыу варианты ҡулланылыуы асыҡланды. Был һүҙҙең Башҡортостан Республикаһынан тыш, сиктәш Ырымбур, Силәбе, Ҡурған, Свердловск өлкәләре ауылдарында ла осрауы билдәле булды.

ӘҘӘБИӘТ

1. Баишев Т.Г. Башкирские диалекты в их отношении к литературному языку. Уфа, 1953. 132 с.
2. Башҡорт һөйләштәренең һүҙлеге. 1 т. Көнсығыш диалект / Н.Х. Мәҡсүтова һ.б. Өфө, 1967. 300 б.
3. Башҡорт һөйләштәренең һүҙлеге. 2 т. Көньяҡ диалект / Н.Х. Мәҡсүтова һ.б. Өфө, 1970. 327 б.

4. Башҡорт һөйләштәренең һүҙлеге. 3 т. Көнбайыш диалект / Н.Х. Мәксүтова һ.б. Өфө, 1987. 231 б.
5. Башҡорт теленең диалекттары һүҙлеге. Өфө: Китап, 2002. 432 б.
6. Башҡорт теленең машина фонды. URL:http: //mfbl2.ru (инеү ваҡыты: 12.09.2020).
7. Боҫконбаева Л.А., Сиразитдинов З.Ә., Ишмөхәмәтова А.Ш., Шәмсетдинова Г.Ғ. Башҡорт теленең диалект аудиокорпусын төҙөү: проблемалар һәм перспективалар // Ватандаш, №12, Өфө, 93–102-се бб.
8. Ғүмәров В.З. Тыуған яҡтың шифалы үҫемлектәре. Өфө, 1996. 160 б.
9. Диалектологическая база. URL:http://mfbl2.ru/mfbl/bashdial http://mfbl2.ru/mfbl/bashdial(дата обращения: 15.09.2020)
10. Диалектологический атлас башкирского языка. Уфа, 2005. 232 с.
11. Ишбулатов Н.Х. Диалектная система башкирского языка в сравнительно-историческом освещении: Автореф. дисс. … д-ра филол. наук. Уфа, 1974.
12. Ишмухаметова А.Ш., Ибрагимова А.Д. Диалектологический подфонд машинного фонда башкирского языка (состояние и перспективы) // Актуальные проблемы диалектологии языков народов России. Уфа, 2014. С. 264–269.
13. Киекбаев Дж.Г. Башкирские диалекты и краткое введение в их историю. Уфа, 1958. 126 с.
14. Лексикографическая база. URL:http: //mfbl2.ru/mfbl/bashlex (дата обращения: 15.09.2020)
15. Максютова Н.Х. Восточный диалект башкирского языка в сравнительно-историческом освещении. М., 1976. 292 с.
16. Миржанова С.Ф. Южный диалект башкирского языка. М.: Наука, 1979. 272 с.
17. Миржанова С.Ф. Северо-западный диалект башкирского языка. Уфа, 1991. 295 с.
18. Севортян Э.В. Этимологический словарь тюркских языков (Общетюркские и межтюркские основы на буквы «Б»). М.: Наука, 1978. 349 с.
19. Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш.. Об обработке звуковых материалов для диалектологического аудиокорпуса башкирского языка // Turkologia (Казахстан, Туркестан), №4(96), 2019, С. 35–45.
20. Räsänen M. Versuch eines etymologischen Wörterbuchs der Türksprachen, Helsinki, 1969. 533 p.

Kuchkarov Mahmudjon, Kuchkarov Marufjon
New York Department of Education, U.S.A, New York

FROM THE "MOVEMENT LANGUAGE" TO THE HUMAN LANGUAGE

Abstract. The origin of 'Human Language' is still a secret and the most interesting subject of historical linguistics. The core element is the nature of the labeling or coding the things or processes with the symbols and sounds. In this paper, we investigate the human's involuntary Paired Sounds and Shape Production (PSSP) and its contribution to the development of early human communication.

Aimed at twenty-six volunteers who provided many physical movements with various difficulties, the research team investigated the natural, repeatable and paired sounds and shape productions during human activities.

The paper claims involvement of Paired Sounds and Shape Production (PSSP) to the phonetic origin of some modern words and existence of similarities between elements of (PSSP) with characters of classic Latin alphabet.

The results may be used not only as a supporting idea for existing theories but to create a closer look at some fundamental nature of the origin of the languages as well.

Keywords: *Body shape, body language, coding, Latin alphabet, merging method, movement language, movement sound, natural sound, origin of language, pairing, phonetics, sound and shape production, word origin and word semantic.*

Кучкаров Мухамеджон, Кучкаров Маруфжон
США, Нью-Йорк

ОТ "ЯЗЫКА ДВИЖЕНИЯ" К ЯЗЫКУ ЧЕЛОВЕКА

Аннотация. Происхождение «человеческого языка» до сих пор остается тайной и наиболее интересным предметом исторической лингвистики в выяснении характера маркировки или кодирования вещей или процессов с помощью символов и

звуков. В этой статье исследуется непроизвольное образование парных звуков и форм (PSSP) человека и его вклад в развитие раннего человеческого общения.

Была исследована корреляция множества физических движений двадцати шести добровольцев с естественными, повторяющимися и парными звуками.

В статье утверждается причастность парных звуков и форм (PSSP) к фонетическому происхождению некоторых современных слов и существование сходства между элементами (PSSP) с символами классического латинского алфавита.

Результаты могут быть использованы не только в качестве вспомогательной идеи для существующих теорий, но и для более детального изучения фундаментальной природы происхождения языков.

Ключевые слова: *форма тела, язык тела, кодирование, латинский алфавит, метод слияния, язык движения, звук движения, естественный звук, происхождение языка, спаривание, фонетика, производство звука и формы, происхождение слова и семантика слова*.

Introduction

"I cannot doubt that language owes its origin to the imitation and modification, aided by signs and gestures, of various natural sounds, the voices of other animals, and man's own instinctive cries" (Darwin C.) [2].

We all know about Newton's falling apple which caused the law of gravity and other myths about science. That apple was real or myth is not the point. The important thing is Newton's law of gravity is real, and it applies to our actual life however, not for quantum mechanics. In contrast to gravity, the established concept of linguistics is still controversial and needs more research on fundamental aspects like the origin of language.

In 1851 Jacob Grimm stated: "The origin of human language is truly secret and marvelous" [7]. Any significant change forward to find the key of that secret is the most interesting subject among international linguists. Framed with the concept, the learning even under the apple tree does not give anything new, that magic apple will never fall on you. It is about us and our research, as the apple

tree and its related components are essential in this topic. Unlike Newton, his ancestors were more pragmatic; they stretched the hand to get the apple. Maybe, it is not magic but humans had a concept, a real idea about interaction and learning from nature even at that time when a human did not speak yet. They found out how to protect themselves, better organized social life and even how to communicate more efficiently with each other. Perhaps, because of that certain strategy, the humans not only survived but also moved to the next much accelerated condition of development after inventing the tool, namely "The Language". What was their main concept to communicate more efficiently and what is the relation to the Paired Sounds and Shape Production (PSSP). This is the main topic of this study.

At that time, when people could not communicate using speech or words, they communicated with gestures and various clicks. Once the physiological and psychological components of human life became sufficiently developed, people have invented the contents of speech and the general idea of improving the quality of communication with each other. The result had to be understandable and acceptable by local community members only, if it is unique and can be reproduced.

In fact, the connection of sounds with actions or objects is not only the central link in the creation of speech and language, but also it serves as a basis for the development of writing. The necessary factors in association of sounds with actions are: time, sound and movement itself. In the other words, the sound must be heard exactly at the moment when the action occurs. During the extreme physical activity, body organs including speech and breath may function with the particular condition and may affect voluntary or involuntary voice reproduction.

Experiments and Discussion

From the basics of human anatomy, we all know that the voice is the sound that arises from the work of the speech organs as a modulator and the respiratory organs as an energy source and an important component of speaking which production begins with breathing. Air inhales as the diaphragm lowers and expands volume and reduces the pressure of the lungs making air come into fill the space. A human exhales as the muscles of the rib cage lowers, and the diaphragm rises, and the air squeezes out. During high physical activity, body organs including speech and respiratory will be forced to function in such condition, which may affect the voluntary and involuntary sound production.

The first observation shows, when we try to reach the point that is a slightly higher, we stretch out our legs, hands and other parts of our body, we involuntarily exhale the sound like **/ɑ:/**.

Focusing on the above idea, we have provided an experimental research on dependence of the involuntary voices and other natural sounds which took place at ESTA Florida Tennis Academy and Camp [8].

During the summer camp, the research team worked out with twenty-six volunteers with different ages, gender and ethics. The evidence recorded when the volunteers made different physical activities like screaming, lifting weights, jumping, laying down, slow and fast standing up and talking during eating, etc. Experiments provided with each volunteer individually as well as within the group. We have selected six people from the dozen for more detailed investigation as most targeting human physical activity which showed clearer evidence of the pairing of involuntary sounds and signs production. During the experimental activities like stretching to the highest point, a lifting and as well as sounds during the eating competition, between the group members were created natural emotional conditions like motivation and competition. By the end of the experiment, data were obtained on the physical activity of the human in which precise movements and sounds associated with them were detected.

In below example, you may see the results from experimental physical activities of our volunteers and their supporting numbers of associated sounds.

Table1.

Activity	*Noticeable body Signs or Shapes during the action*	*Associated sounds during the activity*	*Amount of supporting volunteers of associated sounds*
Jumping	Both feet landing together making a "B" footprint	Both feet hit the ground together with characteristic sound associated with B	BP-11, BB-9 None of above-3, BOP-1, PT-1, DT-1

Picking up the weight	The process starts and ends with the stretching down the arms parallel. As both hands and arms work together in strong dependence like one, we will see “H” sign, when visually connect them through the elbows	Inhalation with a specific short sound associated with H, and ends the process with relieving long exhalation sound associated with	H HEH-9, HK-7, HAH-4, HIH-2, HQ-2, YEK-1, None-1
Pulling compe-tition	During the competition between two individuals, they try to pull the object from each other. Two bodies and their connected hands makes “H” shape	During the competition inhalation and exhalation processes associated with sound HQ and AH respectively	HQ-8,AH-8, HAH-4, HH-2, H-2, EH-1, None-1
Screa-ming	The head and two elevated arms together makes “W” shape, and the mouth gives “O” sign when screaming	Long exhalation comes after the deep inhalation with a specific length and loud sound associated with WO	WO-9, WOA-5, O-3, WOU-3, OW-2, WA-2, A-1, HO-1
Fast eating competit ion with more relisting sounds	From the side body shaped “E”, ”e” when holding and eating the food respectively	Babbling during the eating with a specific sound associated with EA	E-8, EA-8, EAH-3, M-2, YM-1, YA-1, O-1, None of the above-1

Stret-ching up	Body stretched straight, breathing stopped and sudden exhalation when both hands were down with the shape “A”	Exhalation occurs after holding breath with characteristic sound associated with “A	A-12, AH-8, EH-3, E-1, O-1, I-1
Stret-ching up and jumping	At the moment of stretching and jumping, the hands assume an inverted “A” position. And when hands dropped, they form the usual “A” symbol	In three cases out of twenty-six, activists kept jumping without breathing and exhaled long clear sound ”A” when landing	A-14, AH-9, EH-1, E-1, H-1,

From humans’ all repeatable actions that are shown in the experimental activities above were performed with their own symbols and sounds. This allowed us to make the following hypothesis: The shape of the symbol **“B”** from the Latin alphabet is actually an imprint of a pair of feet. Logically, **“B”** is associated with sound, reaching the ground after a jump. **“B”** - is the sound and the symbol. We have to mention that the study focuses only on the relevant evidence of logical relationships. We ignore the involuntary sound like ***“eh”*** that arises during the jump and the same approach was applied to other actions as well.

The ***/h/***sound is the most commonly originated sound of the “pulling/pushing competitions” and "lifting the weight" activity. The possible reason is that the both operations demand most human energy and power consumption. During the contest, the inhalation and exhalation processes are associated with sound ***/h/*** respectively. In the image of pulling competition, two individuals and their arms connected make an **“H”** shape. The important key elements in two different actions are similar, hands and elbows are necessary to

indicate the visual forms and the sound when pulling, pushing or lifting weights with force that affect the recreation of the **"H"** sound.

Table 2.

Visual characteristics	*Necessary elements involved for creation of visual characteristics*	*Phonetic characteristics*	*Necessary elements for creation of phonetic characteristics*	*Key elements for pairing of visual and phonetic elements*
B	Right and left feet associated action to make specific footprint	/b/	Jumping the distance to make a hearable sound with two feet when they hit the ground together	Foot printing and its sound
H	Collective action of two hands (lifting heavy weight), two bodies, four hands (pulling competition) and tree (pushing) to make specific view	/h/, /he/	Energy demanded physical activities like pulling, pushing and lifting	Inhalation and exhalation in special condition during the activities
h	Hanging one hand, arm, body	/h/	Imbalance, emotionally overwhelmed	Breathing in special condition

W	Two arms, head, high demand to communication	None	None	None
WO, WOW	Two arms, head, high demand to communication	/wɑʊ/	Emotional condition, wondering, surprise, danger	Exhalation with high voice in special condition
O	Wide open mouth, shouting, surprised, fear in critical situation	/ɑʊ/	Life-threatening conditions and its imitation	Constantly screaming and yelling with short break during the breathing
E	Body, arms, the food held straight with both hands	/ɪə/, /əə/	Eating the food in competition and in emotional condition	Breath and created sounds under certain conditions
e	Body, hands holding the food near the mouth	/ɪə/, /əə/	Eating the food in competition and in emotional condition	Breath, chewing and created sounds under certain conditions
A	Hand fingers stretched straight together and the thumb folded	/æ/	Stretching body and hands as high as possible	Interrupted breathing and long exhalation

The study shows that the visual characters of **“B”, “H”, “h”, “O”, “E”, “e”, “A”** are the shapes or signs of the part of the human body. And sounds of **/b/, /h/, /he/, wɑʊ/, /ɑʊ/, /ɪə/, /əə/, /æ/** are sounds or voices during the human’s activities respectively.

An interesting situation can be seen in cases with W and O (Experimental activities 1). When we started to analyze the data during the scream, the dominant “WO” sound and the “O” mouth symbol did not interact until the emphasis was on the position of the body. During the process, the head and two raised hands form the “W” symbol, and the open mouth forms an “O” during the human scream.

Many will agree that our ancestors have communicated through gestures and body language. Verbal communication was not necessary if they visually see each other and the voices were used only in the dark or when the visual contact was blocked by any object. Screams were also used in dangerous situations to call for help. We can imagine that in those days, they often found themselves in situations where they needed immediate communication with each other. However, to communicate using gestures, a clear visual contact is necessary, which, alas, could be blocked by trees or other obstacles. In this case, screaming was the only way to communicate, so they screamed very loud and often.

Table 3.

No	*Symbol (letter)*	*Meaning (Actions)*
1	B	Feet, footprint, stepping, starting, marching, together, shoes, walking, facing, putting the pressure on
2	H	Hand, arm, holding, picking up, lifting, having, gaining, earning

3	W, WO	Attention, alarming, yelling, calling for help, imminent danger, informing, leading to conversation
4	O	Shouting, mouth, top, high, many, big, much, open, empty, hole
5	E	Eating, food, consuming, baiting, learning, earning, absorbing
6	A	Stretching, pulling, pushing, expanding, pointing, touching, directing, inside, outside, surface

The importance of the eating and other processes related has probably left their own fingerprints on the historical landscape of language origination and reformation. In collective eating competition, every individual wanted to consume more food, and they tried to eat as fast as possible and at the same time they attempted to keep food away from each other. As our experiment shows, releasing sound while eating is more likely babbling with the sound **/ɪə/** or **/əə**/. A reasonable question would be: why babbling is releasing the sound **/ɪə**/? The answer is simple: The observation indicates when human eats and at the same time trying to reproduce any voice, the sound involuntarily comes out mostly like **/ɪə/** and not any other. Because, releasing any other sound stops the eating or leads to fallen food from the mouth. Comparing our present experiment and ancient times, producing a sound while eating or making any noise, it is most likely, to show an aggressive attitude and try to become the dominant link while eating, to stay as close to food as possible, taking a position in which it is easy to eat, protect food and attack. Observing from different angles, we found that in most comfortable positions for eating, the human body takes shape **"E"** when holding and eating the food (Experimental activities1).

The body stretching experiment was the last and most discussed subject in this research. Due to the importance of **"A"**, as mentioned earlier, we use all our intellectual, statistical and organizational resources obtained in previous studies to carefully analyze all the fundamental processes associated with **"A"** and its origin.

Another reason for the controversy is the results of studies and hypotheses by Christine Kennelly in the book "The First Word: The Search for the Origins of Language" [1], the book with a picture of the **"A"** symbol on the cover, where picture claimed the **"A"** symbol related to human legs shape. Another top content about the **"A"** symbol's origin is on the Google and Google scholar which shows that the **"A"** originated from the prototype of **"OX"** related to animal's head. Surprisingly, in both materials above there is no information about the phonetic origin of the **"A"** element. In contrast, our study shows that the **"A"** symbol is more likely an actual copy of the human's hand shapes when pointing, reaching the point and in some certain conditions even relaxing.

The study shows, when we try to reach the point that is slightly higher to reach, we stretch out our legs, hands and other parts of our body, we involuntarily end our exhalation of breath with the sound like **"A".** Exhalation and relieving which starts after holding a breath during the process of stretching the hands as high as possible, consequently, the sound production organs makes the **"A"** sound longer and louder. For greater approval, we provide a stretch test in another way, when a person jumps to catch an object hanging over him. The process starts with observing the object above for a few seconds, keep looking and lowering the body to 40% percent with moving hands back during the deep inhalation and sudden jump to catch the thing above with the sharp exhalation. Often, stretching hands up makes the shape of turned over **"A"** and turns back when hands stretch down. Jumping with exhalation makes the sound shorter but louder, sharper and clearer sound associated with **"A".** In three cases out of twenty-six, individuals kept jumping without breathing and exhaled long clear sound **"A"** when landing.

The shape of the outstretched arms in both cases (1.6 & 1.7, Experimental activities 1) is similar to the form of **"A",** but the sound characteristics are different. The sound **/æ/** basically arose at the last stage of the exercise, when after holding the breath a long exhalation and relief begins with a specific long sound associated with **/ɑ:/.** Another situation arises in (1.6, Experimental activities 1), where the sound begins to occur in most cases at the same time when the actual jump occurs with a sharp exhalation and a shorter sound associated with **/ɑ:/.**

The extreme physical or emotional situations were significant in the development of early humans' voice and language reproduction. The results of (1.6 & 1.7, Experimental activities 1) were compared with other recent top viewed internet sources related to the origin of an **"A".** Our interview with internet users shows that today, the

origin of the first words and the alphabet symbols are one of the most interesting subjects not only among the scientists and linguists, but also among the ordinary internet users. Requests such as “origin”, “origin of alphabetic characters” [9] in Google and Google Scholar, do not provide any new information or any reliable evidence of relevant research and experiments on the origin of phonetic sounds, symbols [6] and the relationship with the classic alphabet.

Perhaps there are other studies somewhere, but they are not progressing at all, because, as we think, it does not comply with the established concept of linguistics and will probably never be published or discussed. Either way, the results of our current experimental research needs to be published in order to conduct wider discussions related to it and discussions not only among the linguists and the people with the different scientific backgrounds but also with ordinary internet users. For us, a pragmatic view and discussions are important among Internet users with less academic point of view, in order to move the research in the right direction. Based on this study, we made some conclusions on the reconstruction of early words, however, not all the data introduced online yet.

In this presentation [5], we have limited not only our astonishing hypotheses which can be produced by the small intelligent group of scientists but also, tried to write as simple as possible in order to make them understandable for every reader.

Conclusion

Based on the result of the scientific experiments and its deep data analysis, we made the following conclusion: the research team examined the involuntary sound and shape production of humans, as well as his/her contribution to the development of early methods of communication between people. Analyzing the help of twenty-six volunteers who performed various physical exercises and necessary movements, the research team studied the natural and repetitive Paired Sounds and Shape Productions (PSSP) during human’s physical and emotional activities.

This study aims not only to specify the relationships of some humans’ physical and emotional activities and their involuntary sounds and shape production, it also promotes the idea of the importance of natural sounds of human emotions as a fundamental aspect for establishing the earliest words. The result can be considered as an idea in support of existing theories [3] and hypotheses [8], as well as a new and fresh look at some of the fundamental foundations of the origin of languages.

We would like to note that our current method is free, acceptable and safe that makes easy access for everyone to participate in upcoming experimental projects. Additional supporting results of the research were taken from the Facebook respondents after representation of online video tests [4]. We asked respondents to answer the questions like: "What letter of English alphabet or its combination closely represents the sound in the above video?" There were multi-choice answers from the characters of the English alphabet.

The research continues, and more data needs to be analyzed in the future.

REFERENCES

1. Christine Kennelly (2008). "The First Word: https://lifeclub.org/books/the-first-word-christine-kenneally-review-summary.
2. Darwin, C. (1871). The Descent of Man and Selection in Relation to Sex, 2 vols. London: Murray, p. 56.
3. Flöel A, Ellger T, Breitenstein C, Knechti. 2003 Aug;18**.** Language perception activates the hand motor cortex: implications for motor theories of speech perception. DOI**:** 10.1046/j. https://pubmed.ncbi.nlm.nih.gov/12911767/
4. https://www.facebook.com/groups/EstaFL/permalink/ 913768595424051/
5. I. ICOHL 2020 "International Conference on Origins of Human Language" by World Academy of Science, Engineering and Technology, www.waset.org Sep 24-25, 2020 Istanbul, Tukey. II. The First Online International Conference on "THE IMPORTANCE OF USING INNOVATIVE METHODS IN TEACHING FOREIGN LANGUAGES AND TRANSLATION PROBLEMS" organized by The Republic of Uzbekistan Ministry of Higher and Secondary Special Education, June 15, 2020, Fergana State University, Fergana, Uzbekistanwww.fdu.uz
6. International Phonetic Alphabet http://www. internationalphoneticalphabet.org/
7. Jacob Ludwig Karl Grimm.(1951).On the origin of language. Delivered in the Prussian Academy of Sciences. Berlin, 9 Jan. 1951.
8. Linguistic Hypotheses on the Origins of Language.https://freelanguage.org/general-language-info/linguistic-hypotheses-on-the-origins-of-language
9. The Origin of the Alphabet http://webspace.ship.edu/cgboer /alphabet.html

Kyzlasova I. L.
Khakass State University named after N.F.Katanova,
Russia, Khakassia, Abakan

RECIPROCAL SITUATIONS (MUTUALLY-JOINT PLEDGE) IN THE KHAKASS LANGUAGE ACCORDING TO THE ELECTRONIC CORPS

Abstract. The article highlights typical situations of recycling in the Khakass language: cooperative, subjective distribution, object distribution, comitativity, reciprocity, assistance, depativeness. It is noted that in the Khakass language the dominant meanings are subjective distribution and reciprocity. They have a symmetrical relationship between actants and predicates. It was concluded that other recycling situations are secondary in origin. They were formed by semantic development and disruption of symmetry. The work was carried out on the basis of the data of the Electronic Khakass Language Corps.

Keywords: *Khakass language, voice, mutual-joint pledge, recycling, cooperative, distribution, comitativity, assistance, reciprocity*

Кызласова И. Л.
Хакасский государственный университет
им. Н.Ф.Катанова,
Россия, Хакасия, Абакан

СИТУАЦИИ РЕЦИПРОКА (ВЗАИМНО-СОВМЕСТНОГО ЗАЛОГА) В ХАКАССКОМ ЯЗЫКЕ ПО ДАННЫМ ЭЛЕКТРОННОГО КОРПУСА

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-012-00462

Аннотация. В статье выделены типичные ситуации реципрока в хакасском языке: совместность, субъектная дистрибутивность, объектная дистрибутивность, комитативность,

взаимность, ассистивность, депациентивность. Отмечено, что в хакасском языке доминирующими значениями являются субъектная дистрибутивность и взаимность, в которых наблюдается симметричное отношение между актантами и предикатами. Сделан вывод о том, что другие ситуации реципрока являются вторичными по происхождению и образовались путем семантического развития и нарушения симметричности. Работа выполнена на основе данных Электронного корпуса хакасского языка.

Ключевые слова: *хакасский язык, залоговость, залог, взаимно-совместный залог, реципрок, кооператив, дистрибутив, комитатив, ассистив, взаимность.*

Электронный корпус хакасского языка разрабатывается научными сотрудниками Института языкознания РАН (г. Москва) и Хакасского научно-исследовательского института языка, литературы, истории (г. Абакан): А. В. Шеймович, м.н.с. Института языкознания РАН; И. М. Чебочаковой, к.ф.н., в.н.с. ХакНИИЯЛИ; Э. В. Султрековой, к.ф.н. (в прошлом – м.н.с. ХакНИИЯЛИ); В. С. Мальцевой, м.н.с. Института языкознания РАН. Руководит проектом А. В. Дыбо (д.ф.н., чл.-корр. РАН, зав. Отделом урало-алтайских языков Института языкознания РАН). Программно-техническая разработка и поддержка принадлежит Ф. С. Крылову.

В настоящее время компьютеризирован существенный объем материалов и предоставлен к нему общий открытый доступ через сайт https://khakas.altaica.ru/.

В корпусе содержится 154 текста художественных и фольклорных произведений с морфологической разметкой слов. Все тексты имеют параллельный перевод на русский язык. Для поиска необходимых материалов предоставлены разные ключевые параметры: Словоформа (в хакасской орфографии), Лемма (в хакасской орфографии), Аффиксы (автоматическая разметка), Русский перевод леммы. Можно выбрать дополнительные параметры: Русский перевод предложения (подстрока), Подкорпус, Ширина контекста (число фраз до и после).

Работа по наполнению и развитию корпуса продолжается и служит мощным инструментом для изучения, исследования, сохранения и развития хакасского языка. Хочу выразить огромную благодарность его создателям и разработчикам, его руководителю Анне Владимировне Дыбо.

Для выполнения моего проекта при поддержке гранта РФФИ по теме «Функциональная грамматика хакасского языка: Залоговость. Временная локализованность. Количественность» электронный корпус предоставил огромный практический материал для анализа. Так, поиск по принципу автоматической разметки аффиксов взаимно-совместного залога (реципрока) на *-(ы)с* выдано 10675 словоформ, 8812 фраз на 18 электронных страницах из трех источников (вместо 154). То есть поиск можно продолжать и расширять, задавая системе условия выбора других текстов.

Из полученных данных пришлось отсекать омонимичные формы *-(ы)с*, так как в системе заложены все возможные варианты грамматических значений конкретной словоформы. Так, наряду с реципрокальными формами, проявились:

а) форма *-(ы)с* со значением глагольного процесса: *Ол парасханны тимір тӧгее салып, пазын тимір палтынаң ӱзе сабысчаңнар* «Того беднягу клали на железное бревно и отрубали ему голову железным топором»; *Юраны чи уйат тӱдӱн нимес, харах чібіспес?*«А помнишь Юру — стыд не дым, глаза не выест?»;

б) словообразовательный аффикс *-(ы)с* для существительного: *Мин аныңҷа иртен чуунарға иніс індіре тірлектене ле халчам* «Я через эту калитку утром вниз по пригорку мчусь умываться»; *Суғ ибіріс чірінде ағырин ахча* «На крутом изгибе реки вода спокойная»;

в) понудительный залог (каузатив) на *-(ы)с,* еще не отмеченный в грамматиках хакасского языка: *Кӧр ле салып мындағ харахтарны, эміскен іҷемнең мин хорыххам* «Увидев такие глаза, кормившей меня матери я испугался» (М. Баинов).

Также корректировке подлежали переводы предложений на русский язык, так как заложенный в некоторых текстах

художественный перевод не всегда отражает суть грамматических значений.

Извлеченный из электронного корпуса практический материал был систематизирован по семантике и позволил придти к выводу, что реципрок в хакасском языке – это не только симметрично взаимные действия двух и более субъектов, не только совместное выполнение действия ими, а это по сути - «множественность» в разных вариациях: множество участников действия, множество ситуаций, множество актов в одном действии (мультипликативность), множество объектов.

Рассмотрим типологию реципрокальных ситуаций в хакасском языке:

а) совместность (кооператив; социатив), когда несколько субъектов-агенсов выполняют одно действие. Такая ситуация тестируется обстоятельствами, типа *вместе, совместно, сообща, как один, враз, дружно.*

Примеры:

Піс Кай нанҷымнаң ол хазың көзітчеткен чазыт орынны хайди даа табарға тір чөптес салғабыс «Мы с Каем, моим другом, приняли решение добраться до того таинственного места, на которое смотрит ветка березы»; *Піс күрезербіс, арғыстар, толдыра чиңіске читкенҷе* «Мы будем бороться, товарищи, до полной победы» (В. Кобяков); *Піс, МТС-тің тоғысчылары, сірер дее, колхозниктер, ол чолҷа хол тудыныс салып, хада парчабыстар* «Мы, работники МТС, вы, колхозники, идем по той дороге вместе, взявшись за руки (Г. Топанов);

б) субъектная дистрибутивность – это ситуация, в котором расчлененные действия охватывают много симметричных субъектов.

Общая ситуация, маркированная в высказывании, является гомогенной, поскольку происходит в один период времени [6: 144].

«Мультисубъектные глаголы позволяют выразить идею пространственной дискретизации события как совокупности отдельных микрособытий, происходящих одновременно» [1: 36]

Примеры:

Аалда олар іди көдгісчеткеніне хатхырысчалар «Над их привычкой во всем подражать друг другу в аале посмеиваются»; *Прайзы хайхасчаң* «Все любовались»; *Ікізінең өрініскеннер* «Оба радовались»; *Анда нимеде тыҷырапча, хар-пораан өтіре көгемзік сағыннар тооласча* «А там что-то трещит, и сквозь снежные заряды голубые искры сыплются»; *Киліңер хакас чазыларынзар, сірер хайхас көрерзер анда: суғлар хазынзар, тағлар олиинзар садтар көгерісче чайғыда* «Выходите на поля Хакасии, чудеса вы увидете там: по берегам рек, по склонам гор сады зеленые цветут» (Н. Доможаков);

в) объектная дистрибутивность – это ситуация, в котором один или несколько агенсов выполняют расчлененные действия, охватывающие много симметричных объектов.

Примеры:

Піс пазох ла педальларны матап толғастыр сыххабыс «Мы опять усиленно закрутили педали»; *Көрінче: чабан аңдарыл парча, тура хонча, холларынаң пулғастапча, а хойлар аныңзар күрелісчелер* «Видно: чабан падает, вскакивает, руками размахивает, а овцы к нему бегут»; *— Ууҷа, — пазох саба-суба кірізібіскен Хал Петке, — хайдаң халған пісте пу чатхан?* «— Ууча, — опять не к месту вмешался Хал Петка, — а откуда у нас этот чатхан?»; *«Кізіни тізе, уғаа чоон, мални тізе, мални нимес, та-а,..» — анаң позыныҥ чодазынаң тиңнестір сыххан* « — Ежели человеческая, то шибко большая. На скотскую не похожа. — Потом начала примеривать (кость) со своей ногой»;

г) комитативность: это «прагматическое выделение одного из двух симметричных актантов [5: 278].

Примеры:

Мин палалардаң ам на сынҷых тоғыстаңар чоохтасхам, сині көзідімге турғысхам, а син... «Я только что с ребятами о честности говорила, тебя в пример ставила, а ты...»; *Пу Хал Петкенең сырбалыспас полғам* «Не надо было мне с этим Хал Петкой связываться»; *Чағын пастыр киліп, изеннескем* «Подойдя поближе (к чабану), поздоровался»; *Ууҷам тапсабинчатса, аннаң сарыспасха кирек* «Когда бабушка молчит, с ней не нужно спорить»; *Мындағ чонның ӱчӱн хайди*

ыырҷынаң кӳреспеҷең! «Как же не бороться с врагом ради такого народа!»;

д) взаимность (собственно реципрок): «действие, совершаемое двумя или несколькими субъектами по отношению друг к другу» [4: 160].

Примеры: *Пір-пір ле — олар тудызыбызарлар* «Еще минута — и они раздерутся»; *Піс Кайнаң көгілбей ӧрге — ол аннаң минің не чазыттыг, кізее чоохтабас, киреебіс полар тіп, сӧс тее чох піліс салғабыс* «Мы с Каем без слов друг друга поняли: голубой дворец будет нашей тайной»; *Удур-тӧдір харах көріс полбаабыс* «Друг на друга не смогли посмотреть»; *Удур-тӧдір сӧс пыласчалар, алтын нинҷее турарынаңар сарысчалар* «Перебивают друг друга, спорят, сколько это золото может стоить»;

е) депациентивность: когда «из исходной аргументной структуры удаляется участник с семантической ролью пациенса или цели» [7].

Примеры: *Көрінген, ағаа чоохтазарға сидік* «Было видно, ему разговаривать трудно»; — *Худайбынаң, чаалысханымны таныхтааннарох орден-медальларнаң* «— Слава богу, отметили орденами и медалями мои военные действия (букв. *моё воевание*)»; *Тастаң тастазарға чарабас* «Камнем нельзя бросаться»;

ж) ассистивность - «это ситуация, когда семантические участники имеют несимметричные роли, один из агенсов активен, а другой пассивен и оказывает помощь ему в совершении действия» [3: 95].

Примеры:

Сайанаа азах хаарарға сіренген, че аның ӱчӱн Маля турысхан «Попытался подставить ножку Сайане, но за нее Маля заступилась»; *Ізе, хайы ағамнина турбинчатхан, че ол алыптың чолын чоллас пирген* «Конечно, горловое пение у него получалось не так, как у деда. Зато он привел алып-богатыря, куда сказание велело»; — *Пик сӧӧктіг дее полбаза, чарир,— алныма турызыбысхан іҷем* «— Вовсе не обязательно крепкую кость иметь, — заступилась за меня мама».

При абстрактном употреблении глагольной словоформы на *–(ы)с* проявляется признак полисубъектности, поэтому инвариантным значением реципрока в хакасском языке является «участие в совершении действия одновременно нескольких субъектов (не менее двух)» [2: 177]. В контекстном употреблении семантические роли этих субъектов распределяются по-разному: симметрично и асимметрично.

Доминирует симметричное отношение между актантами, которое можем наблюдать, например, в ситуациях совместности, субъектной дистрибутивности, объектной дистрибутивности, взаимности (см. пп. *а, б, в, д*).

В ситуации совместности наблюдаются «симметричное отношение между актантами с одной и той же семантической ролью (*А и Б поют вместе; я видел А и Б вместе*)» [5: 277].

В ситуации взаимности присутствует симметричное отношение между актантами с различными семантическими ролями (*А целует Б, Б целует А*).

В ситуации субъектной дистрибутивности симметричные агенсы участвуют в симметричных событиях (*А бегает и Б бегает – все бегают*).

В ситуации объектной дистрибутивности агенс/агенсы охватывают действием симметричные объекты (*А сравнивает N_1 и N_2*).

Ассиметричность актантных ролей вторична, развилась от основного реципрока, встречается в ситуациях мультипликативности, комитативности, депациентивности, ассистивности (см. пп. *г, е, ж*).

Комитативность развилась от значения взаимности через понижение синтаксического ранга одного из симметричных субъектов, он становится соучастником действия. В ситуации депациентивности вовсе был удален один из симметричных субъектов. А в ситуации ассистивности, наоборот, добавлен новый участник, оказывающий помощь.

В целом, по частотности употребления превалирует ситуация субъектной дистрибутивности, затем - ситуация взаимности. Это основные значения хакасского реципрока. Другие ситуации менее употребительны и являются

мотивированными от основных значений. Реципрок в хакасском языке оказался совмещен с такими категориями, как «кооператив/социатив», «итератив», «мультипликатив», «дистрибутив», «комитатив», «ассистив». Признаками реципрока являются: «множественность», «кратность», «синхронность», «симметричность / нарушение симметричности».

ЛИТЕРАТУРА

1. Галиева А.М., Замалетдинов Р.Р. Способы выражения мультисубъектности в татарском языке: грамматические и семантические аспекты // ФИЛОЛОГИЯ И КУЛЬТУРА. PHILOLOGY AND CULTURE. 2015. №3(41) С.32-38.
2. Грамматики хакасского языка / под ред. проф. Н.А.Баскакова. - М.: «Наука», 1975.
3. Данилова Н.И. Значение взаимного действия и семантический класс глагола в якутском языке // Гуманитарные и социальные науки. - 2016. - № 5. - С.90-100.
4. Лингвистический энциклопедический словарь / Гл. ред. В. Н. Ярцева, — М.: Сов. энциклопедия, 1990. —685 с.: ил. ISBN 5-85270-031-2.
5. Недялков В.П., Генюшене Э. Ш. Типология рефлексивных конструкций // А. В. Бондарко (ред.). Теория функциональной грамматики. Персональность. Залоговость. - С.Петербург: Наука. 1991, - С. 241—276.
6. Понкратова А.Н. Дистрибутивная глагольная множественность в когнитивно-функциональном аспекте: На материале английского и русского языков: дис. канд. филол. наук. – Кемерово, 2002. -171 с.
7. Привознов Д. К. Семантика одного глагольного показателя в мишарском диалекте татарского языка // Девятая Конференция по типологии и грамматике для молодых исследователей. Материалы. — Институт лингвистических исследований РАН Санкт-Петербург, 2012.

Mammadzada S.
Institute of Information Technology of ANAS,
Azerbaijan, Baku

THE DEVELOPMENT OF THE NATIONAL TRANSLITERATION SYSTEM OF THE AZERBAIJANI LANGUAGE

Abstract. This paper describes the development of the National Transliteration System of the Azerbaijani language. The system performs the transliteration of words, texts and a whole web-site, as well, from Azerbaijani Latin into the scripts of seven languages (for Russian, English, German, French, Farsi, Spanish, and Italian), including into the Cyrillic previously used for the Azerbaijani, and vice-versa (excluding Farsi-Azerbaijani transliteration). The paper also presents the normative and legislative bases for the development of the system and the duties arising from these documents. It highlights the goal and tasks of the system and describes it. The methods and algorithms for the automated transliteration of Azerbaijani language are proposed. It also shows advantages and disadvantages of developed system. The paper touches upon the problems in the field of the Azerbaijani transliteration andthree main transliteration standards adopted for the Azerbaijani language so far. In conclusion, information about the transliteration accuracy is provided.

Keywords: *transliteration; romanization; scripts; languages; conversion tables.*

Мамедзада С.
Институт информационных технологий НАНА,
Азербайджан, Баку

РАЗВИТИЕ НАЦИОНАЛЬНОЙ СИСТЕМЫ ТРАНСЛИТЕРАЦИИ АЗЕРБАЙДЖАНСКОГО ЯЗЫКА

Аннотация. В статье описывается развитие нацио-нальной системы транслитерации азербайджанского языка. Система выполняет транслитерацию слов, текстов и всего веб-сайта с

азербайджанской латыни в алфавиты семи языков (русского, английского, немецкого, французского, фарси, испанского и итальянского), в том числе и с кириллицы, которая ранее использовалась для азербайджанского и, наоборот (за исключением транслитерации фарси на азербайджанский). Автором также представлены нормативные и законодательные основы для развития системы. Предложены методы и алгоритмы автоматической транслитерации азербайджанского языка. Показаны достоинства и недостатки разработанной системы. В заключение приводится информация о точности транслитерации.

Ключевые слова: *транслитерация, романизация, скрипты, языки, таблицы преобразования.*

Introduction

It is challenging to translate names and technical terms across languages with different alphabets and sounds. These words are frequently transliterated, i.e., replaced with approximate phonetic equivalents. For example, a word “computer” appears as “kompüter” in Azerbaijani. In the context of multilingual natural language processes, where the languages using both Latin and other graphics are represented, the lexical resources should include mechanisms for developing words that do not have standard transliteration. This category includes the words with no explicit semantics, such as personal names and places’ names. Those words are not translated but simply transliterated. Transliteration is a linguistic process, which represents the texts written in the graphics other than the Roman (English) languages and has been used for centuries [3]. In this regard, the development of the national transliteration system with standard conversion tables is required.

The following sections of the paper describes the legislative bases for the development of the National Transliteration System for the Azerbaijani language, its goals and duties, advantages and disadvantages.

Normative and legislative bases for the development of the National Transliteration System

After regaining independence, the Azerbaijani language has played exceptional importance as the State language in political-public, socio-economic and scientific-cultural life of the nation. Due to two presidential Decrees, that is “On the Improvement of the Use of the State Language” dated June 18, 2001 [2], and “On the Implementation of the Law of the Republic of Azerbaijan “On the State Language in the Republic of Azerbaijan” dated January 2003[4], the scope of the literary Azerbaijani language has been broadened with the potentiality of its various methods being discovered and entirely new prospects being opened for our language skills and habits.

In addition, the following two clauses of the Presidential Order dated April 09, 2013 on approval of the “State Program on the use of the Azerbaijani language in accordance with the requirements of the time in the context of globalization and the development of linguistics in the country” urges the development of the National Transliteration System for the Azerbaijani language:

6.1.4. “Development of National Transliteration Standards”;

6.3.9. "Development of software for transliteration from the Azerbaijani alphabet to other alphabets on the basis of national transliteration standards".

The duties arising from the order beforerelated institutions of the Azerbaijan National Academy of Sciences include the followings:

- Study of the interaction of the Azerbaijani language with other languages;
- Study of foreign experience in the field of transliteration;
- Compilation of conversion tables of the Azerbaijani scripts with the scripts of other languages;
- Adoption of national transliteration standards based on conversion tables;
- Development of transliteration software for the Azerbaijani scripts with other languages’ scripts on the basis of national transliteration standards;

- Wide application and protection of the Azerbaijani language, scripts and terms, protection of national values (history, territory, culture, etc.) and correct communication with the world community.

Goals and tasks of the system

The goal of the "National Transliteration System" is ensuring the one-to-one conversion of the symbols (scripts) of the writing system of the Azerbaijani language into the symbols (scripts) of the writing systems of other languages.

One-to-one transliteration means that when converting a word from one script into another, one-to-one principle between the elements of two sets is preserved, where each element of one set is paired with exactly one element of other set, and each element of the other set is paired with exactly one element of the first set, i.e., English b is Cyrillic б, and Cyrillic б is English b, English d is Cyrillic д, and Cyrillic д is English d, etc.

Whereas the tasks of the "National Transliteration System" includes the followings:

- ensuring the correct spelling of words on the Internet during the search;
- ensuring the correct transition of the text from one alphabet into another;
- creating a common information environment among Azerbaijanis around the world by eliminating the problem of scripts.

Note that the number of world Azerbaijanis by countries accounts for more than 50 million, including over 32 million Azeris in Iran, 2 million in Russia, 3 million in Turkey, 200000 in Germany and other countries. Unlike the older generation, the youngster may speak Azerbaijani within the family or Azerbaijani speaking community, but may not know the scripts of the Azerbaijani language. Thus, ensuring the correct transition of the text particularly from Azerbaijani into other scripts will provide a common information environment among Azerbaijanis around the world.

Problems in the field of the Azerbaijani transliteration

Due to the lack of unique national transliteration (conversion) tables and standards, the words (names of places, personal names and etc.) are used and distributed in different variants in the Internet, for example:

- Gəncə–Gandja – Gendja – Gandzha – Gendzha – Gence – Genje– Ġänjä

Another reason for the emergence of various transliteration options is related to the adoption of different alphabets for the Azerbaijani language in different periods (Latin, Arabic, Cyrillic and etc.). Consequently, huge number of documents, books, manuals, textbooks, newspapers and numerous publications were published in each of these alphabets. Manual and individual transliteration of all of these publications without any unique standards has led to the wrong transliteration options. For example, many personal names were first transliterated into Cyrillic taking into account the phonetic peculiarities of the Russian language and only then converted into Latin, generating numerous differing transliteration options, as follows:

- Əfəndiyev – Афендиев –Эфендиев – Afendiev–Afendiyev – Efendiev – Efendiyev –Äfändiyev
- Qurbanova– Курбанова – Kurbanova – Gurbanova

Fortunately, there have been adopted three main transliteration standards for the Azerbaijani language so far:

- BGN / PCGN - United States Board on Geographic Names (BG) and the Permanent Committee on Geographical Names (PCGN) adopted in 1993 for the transliteration of the names of places for the Romanization of Cyrillic and modern Latin (for Azerbaijani) [1].
- ISO9 - Transliteration of Cyrillic (for Slavic and non-Slavic languages) into Latin adopted in 1995Romanization of Cyrillic for the Romanization of Cyrillic and modern Latin (for Azerbaijani) [6].
- ALA LC - US Library of Congress conversion tables adopted in 1997 for the Romanization of the books and other publication written in Azerbaijani Cyrillic and ancient Arabic (for Azerbaijani) [7].

Description of the System

The National Transliteration System (http://transliterasiya.az/) runs in eight languages (Azerbaijani, Russian, English, German, French, Farsi, Spanish, and Italian) and includes sections as, home page, about transliteration, transliterations standards (national standards and international standards), transliteration tables, collaboration and contacts. The system performs the transliteration of words, texts and a whole web-site, as well, from Azerbaijani Latin into the scripts of seven languages (for Russian, English, German, French, Farsi, Spanish, and Italian), including into the Cyrillic previously used for the Azerbaijani, and vice-versa (excluding Farsi-Azerbaijani transliteration).

Themethods and algorithms applied in the creation of the system are developed based on the transformation rules, particularly replacement. These transformation rules are worked out in accordance with the conversion tables provided by the experts of related languages. Special method and algorithm are developed for the Azerbaijani-Farsi transliteration, taking into account not only graphic and phonetic characteristics of both languages, but also position (separate, in the beginning, middle and end of the word) of letters within a word in Farsi.

For Azerbaijani-English transliteration ISO9 Transliteration of Cyrillic (for Slavic and non-Slavic languages) into Latin is applied. However, this standard is not applicable for the transliteration of personal names and names of places due to the presence of many diacritics. Moreover, except Azerbaijani-English transliteration, for the transliteration between Azerbaijani and other abovementioned languages, there are no any international standards. Thus, conversion tables for the transliteration between Azerbaijani and other abovementioned languages are proposed by the experts.

Advantage of the developed system is providing transliteration of Azerbaijani Latin with the scripts of other abovementioned languages for the first time, taking into account not only graphic characteristicsof each language, but also phonetic one. In addition, any changes can be made to the system and additional conversion tables can also be added. Along with the words and texts, a whole web-site can also be transliterated.

Disadvantage of the system is that proposed conversion tables are applicable for the transliteration of the general information, not the personal names and names of places. However, the section "Transliteration tables" of the system includes extended tables for the manual transliteration of the personal names and names of places.

Future research

To improve the system, in the future, the inclusion of training data for the transliteration of personal names and names of places from Azerbaijani into the scripts of mentioned 7 languages and vice-versa. Future research will also embrace support of Farsi-Azerbaijani transliteration and Arabic-Azerbaijani, as well as Azerbaijani-Arabic transliteration.

Conclusion

This paper highlighted the development of the National Transliteration System of the Azerbaijani language jointly by the institutes of ANAS (Institute of Information technology and Institute

of Linguistics after Nasimi), which performed the transliteration of words, texts and a whole web-site, as well, from Azerbaijani Latin into the scripts of seven languages (for Russian, English, German, French, Farsi, Spanish, and Italian), including into the Cyrillic previously used for the Azerbaijani, and vice-versa (excluding Farsi-Azerbaijani transliteration). The developed system is the first system performing automated transliteration for Azerbaijani scripts. Since, ISO9 standard was applied for Azerbaijani-English transliteration, transliteration accuracy was 100%. Transliteration accuracy for Azerbaijani-Russian transliteration was accounted for 95%, due to some uncertainties related to the conversion of the Azerbaijani letter "ə". Transliteration accuracy for the transliteration of other scripts was under 90%.

REFERENCES

1. BGN/PCGN Romanization System for Azerbaijani, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/881069/TABLE_OF_CORRESPONDENCES_FOR_AZERBAIJANI.pdf.
2. Decree of the President of the Republic of Azerbaijan "On the Improvement of the Use of the State Language" dated June 18, 2001, https://aztc.gov.az/az/posts/id:190.
3. Lehmann W.P. 1992. An introduction in W.P. Lehmann (Ed.), Historical linguistics, 3rd ed., pp.46-64.
4. On the Implementation of the Law of the Republic of Azerbaijan "On the State Language in the Republic of Azerbaijan" dated January 2003, https://www.aztc.gov.az/en/posts/id:12.
5. Presidential Order dated April 09, 2013 on approval of the "State Program on the use of the Azerbaijani language in accordance with the requirements of the time in the context of globalization and the development of linguistics in the country", www.president.az.
6. Transliteration of Cyrillic (for Slavic and non-Slavic languages) into Latin, https://www.iso.org/standard/3589.html
7. US Library of Congress conversion tables, https://www.loc.gov/catdir/cpso/romanization/azerbaij.pdf.

Muzafarova A.I., Minullin D.A., Gafarova V.R.
KFU, Institute of Computational Mathematics and Information Technologies, Institute of Applied Semiotics of the AS RT, Rossia, Tatarstan, Kazan

CLUSTER ANALYSIS OF TEXTS OF LESSONS PLANNING SYSTEM "ELECTRONIC EDUCATION OF THE REPUBLIC OF TATARSTAN"

Abstract. The article is devoted to the application of BigData methods in the school education system, using the example of analyzing the big data on the activities of teachers and student success, collected and continuously updated in the "Electronic Education in the Republic of Tatarstan" system. This work describes the development of a system for clustering texts of lesson planning to determine their belonging to the corresponding teaching materials. Based on the programs developed by the authors, the texts of the lesson planning are divided into 8 clusters, of which 6 clusters are in Russian, 2 clusters are in the Tatar language of instruction. Also, an analysis of the average marks of students was carried out, depending on the teaching materials used by teachers.

Keywords: *big data; data analysis in education; information processing; clustering of texts; comparison of texts.*

Музафарова А.И., Минуллин Д.А., Гафарова В.Р.
Казанский федеральный университет, ИВМиИТ, Институт прикладной семиотики АН РТ, Россия, Татарстан, Казань

КЛАСТЕРНЫЙ АНАЛИЗ ТЕКСТОВ ПОУРОЧНОГО ПЛАНИРОВАНИЯ СИСТЕМЫ «ЭЛЕКТРОННОЕ ОБРАЗОВАНИЕ РЕСПУБЛИКИ ТАТАРСТАН»

Аннотация. Статья посвящена применению методов больших данных (BigData) в системе школьного образования, на примере анализа массив больших данных о деятельности

педагогов и успешности учащихся, собранного и непрерывно обновляемого в системе «Электронное образование в Республике Татарстан». В этой работе описана разработка системы кластеризация текстов поурочного планирования для определения принадлежности их к соответствующему УМК. На основе разработанных авторами программ, тексты поурочного планирования выделены в 8 кластеров, из которых 6 кластеров на русском языке, 2 кластера на татарском языке преподавания. Также проведен анализ средних оценок учеников в зависимости от используемых педагогами УМК.

Ключевые слова: *большие данные, анализ данных в образовании, обработка информации, кластеризация текстов.*

1. Введение

Аналитика больших данных — это процесс изучения больших объёмов данных с целью выявления скрытых закономерностей, рыночных тенденций, предпочтений клиентов и другой полезной информации для принятия правильных решений [1]. Она была принята самыми различными отраслями и стала самостоятельной отраслью [14]. Оперирование большими данными (BigData) в образовании – это технология аналитики образовательной системы, включающей измерение, сбор, анализ и представление структурированных и неструктурированных данных огромных объемов об обучающихся и образовательной среде с целью понимания особенностей функционирования и развития образовательной системы [20]. Работа с объёмными массивами данных требует не только наличия современных аппаратных средств, но также и математических алгоритмов, которые позволили бы сократить необходимое число вычислительных операций для компьютера. Для успешного управления образовательным процессом необходимо оперативно обрабатывать многочисленные разнообразные поступающие данные в онлайн режиме, поэтому применение технологий BigData становится необходимостью [7].

Кластеризация — это задача поиска групп похожих документов в коллекции документов. Алгоритмы кластеризации являются одними из самых популярных методов

интеллектуального анализа данных, который широко применяется для обработки текстовых данных. Они имеют широкий спектр приложений, таких как классификация [2; 3], визуализация [4] и организация документов [6]. Существуют различные методы кластеризации текстов, наиболее популярные из них это LSA/LSI – Latent Semantic Analysis/Indexing [13], Suffix Tree Clustering [15], Scatter/Gather [5]. В последнее время популярность получили методы, основанные на использовании нейронных сетей [10] совместно с классическими методами кластеризации, такими как алгоритм k средних [8]. Алгоритмы, основанные на нейронных сетях, применяются также и в задачах классификация текстов: например, борьба со спамом, распознавание эмоциональной окраски текстов, разделение сайтов по тематическим каталогам, персонализация рекламы [19].

В данной работе решена задача кластеризация большого объёма текстовых данных, накопленных в системе «Электронное образование в Республике Татарстан» с 2014 по 2020 годы. На первом этапе с использованием метода косинусного расстояния тексты поурочного планирования проверены на схожесть. На втором этапе на основе использования метода агломеративной кластеризации тексты сгруппированы в 8 кластеров. Статья устроена следующим образом. В первом разделе описываются возможности гибкой библиотеки Dask для проведения параллельных вычислений в кластерных вычислительных системах, во втором разделе рассматривается оценка схожести текстов с использованием метода косинусного расстояния. В третьем разделе описан алгоритм агломеративной кластеризации, и его применение для кластеризации текстов поурочного планирования на основе матрицы схожести. Далее приводится заключение, подводящее итог статьи.

2. Обработка больших данных с использованием системы Dask

Основу исследования составили данные, собранные через государственную информационную систему «Электронное

образование в Республике Татарстан». Система включает в себя базы данных образовательной информации по всем учащимся и всем педагогам общеобразовательных организаций РТ. Для исследования наборы данных предоставлялись в формате Comma-separated values (CSV файлов). Общий объём данных составляет более 60 Гб. Такой объём данных невозможно эффективно обработать стандартными средствами. Следовательно, нужны технологии, которые позволяют обработать большой объем неструктурированных данных, систематизировать их, проанализировать и выявлять закономерности.

Для проведения расчетов был развернут вычислительный кластер, состоящий из 4-виртуальных машин, каждая из которых имеет по 1ТБ постоянной памяти, 32 ГБ оперативной памяти, 16 вычислительных ядер. На этом кластере была установлена система для параллельных вычислений - Dask. Dask – гибкая библиотека параллельных вычислений для аналитики, предназначенная главным образом для обеспечения масштабируемости и расширения возможностей существующих пакетов и библиотек [11]. Данная система позволяет производить параллельные вычисления на данных, размер которых превышает доступный объём памяти, на нескольких ядрах или нескольких машинах. Можно даже сконфигурировать Dask для использования ресурсов тысячи машин – каждой с несколькими ядрами (Зятев, 2019).

В обработку поступили данные, характеризующие профессиональную деятельность педагогов, включая темы уроков, задаваемые домашние задания, и оценки учеников. Процесс загрузки данных из CSV-файлов в Dask Dataframe осуществлялся с помощью функции dask.dataframe.read_csv(). Для переименования столбцов использовалась функция dask.dataframe.rename(). Для объединения фреймов данных применялся метод dask.dataframe.merge(). С помощью описанных инструментов все нужные для исследования данные были считаны и объединены в общий dataframe, который использовался для дальнейшей обработки.

Далее для дисциплины «математика» к сгруппированному по классам (1-11 классы) dataframe была применена функция параллельной обработки, которая для каждого учителя выделила темы занятий урока и заданные им домашние задания. После обработки исходных данных на кластере с установленной системой Dask, мы получили тексты поурочных занятий, сгруппированные по педагогам и предметам (Табл. 1)

Таблица 1.

Представление данные после обработки в системе Dask

worker_id	homework	theme
594697	РТ стр. 42 43	Связи между скоростью, временем и расстоянием
594697	РТ стр. 44 45	Письменное умножение двузначного числа на двузначное
594697	РТ стр. 46 47	Письменное умножение двузначного числа на двузначное

3. Определение схожести текстов

Следующей задачей исследования является построение матрицы схожести текстов. В настоящее время поиск сходства между текстами имеет большое практическое применение. Существует много методов и программ для сравнения текстов [16]:

1. Методы сравнения текстов, основанные на коэффициенте подобия Джакарда, косинусное подобие и расстояние Левенштейна.
2. Метод латентно-семантического анализа.
3. Программы обнаружения плагиата, основанные на поиске скрытой семантики текста и т.д.

Тексты поурочного планирования каждого преподавателя, полученные на первом этапе, соединяются в единую строку, которая подвергается следующей обработке:

1. Строка разбивается на токены.

2. Из массива токенов удаляются: знаки препинания, пустые строки и стоп-слова, которые не придают особого значения предложению.

Это необходимо для повышения точности сравнения.

На основе полученных данных осуществляется анализ схожести сравниваемых текстов. В качестве способа сравнения был выбран метод косинусного сходства [12]. Косинусное сходство – это мера сходства между двумя векторами пространства внутренних произведений, которое измеряет косинус угла между ними. Если даны два вектора признаков, A и B, то косинусное сходство, cos(θ), может быть представлено, используя скалярное произведение и норму [17]:

$$\cos(\theta) = \frac{\vec{a}\vec{b}}{\|\vec{a}\|\|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i}\sqrt{\sum_1^n b_i}}, \quad (1)$$

где a и b вектора, $\|\vec{a}\|$ – координаты первого вектора (количество вхождений для первого текста), $\|\vec{b}\|$ – координаты второго вектора (количество вхождений для второго текста).

Сравнивая попарно вектора строим матрицу схожести текстов.

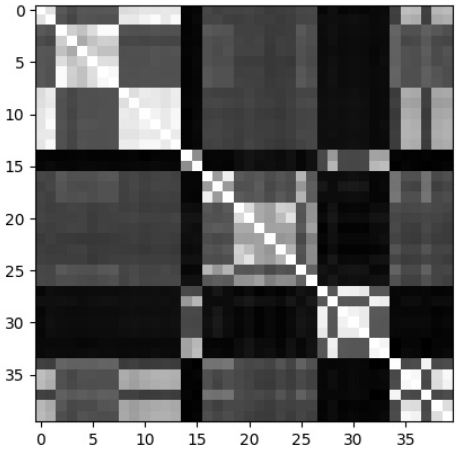


Рис. 1. Матрица похожести для 40 векторов

На рисунке 1 представлено изображение матрицы схожести текстов тематического планирования занятий для 40 преподавателей. По диагонали 100% совпадение, так как сравниваются одинаковые тексты. Чем ярче цвет, тем выше степень похожести между двумя текстовыми данными.

4. Кластерный анализ

Для обучения учеников начальной школы используется 11 различный УМК. Такие как: "Школа России", "Перспектива", "Начальная школа XXI века", "Гармония", "Перспективная начальная школа", "Планета знаний", "РИТМ", "Начальная инновационная школа", система Л.В. Занкова, система Д.Б. Эльконина-В.В. Давыдова, **"Учусь учиться (математика Л.В. Петерсон)" [18].** Однако в Республике Татарстан о основном используется только около 8-ми УМК, 2 из которых на татарском языке. Следовательно, было принято решение полученные данные по тематическому планированию разделить на 8 различных кластеров. Для осуществления этой задачи был выбран алгоритм, который относится к классу *агломеративных:* основной операцией является слияние нескольких уже имеющихся кластеров в один более крупный кластер.

Суть алгоритма заключается в следующем. Изначально все объекты считаются отдельными кластерами. Предполагается, что кластеризация образует покрытие исходного множества. При необходимости можно считать, что изолированные (не принадлежащие ни одному из кластеров) элементы образуют тривиальные кластеры из одного элемента. Затем начинается процесс последовательного слияния кластеров, на каждой итерации которого выбираются два наиболее близких кластера и объединяются в один новый [9]. Алгоритм заканчивает работу, когда остается только один кластер, совпадающий с исходным множеством. Таким образом создается дерево от листьев к стволу.

Расстояния между объектами (таблица 2) исходной выборки рассчитываются на основе матрицы схожести (таблица 1).

В итоге получается дерево кластеров, из которого можно выбрать кластеризацию с требуемой степенью точности.

Таблица 2.

Матрица расстояний для 6 векторов

0	0.1157	0.6469	0.6954	0.6555	0.6590
0.1157	0	0.6693	0.185	0.6624	0.6703
0.6469	0.6693	0	0.1880	0.2390	0.1294
0.6954	0.7185	0.1880	0	0.3266	0.2314
0.6555	0.6624	0.2390	0.3266	0	0.1807
0.6590	0.6703	0.1294	0.2314	0.1807	0

В идеальном раскладе уже на двух кластерах получилось бы разбиение на кластеры содержащий тексты русском языке и на татарском языке. Но этого не произошло, так как в заданиях, которые указывают учителя содержатся не только татарские слова, но и русские. Следовательно, однозначного разбиения на русский и татарский звенья на двух кластерах получить невозможно. Увеличивая количество кластеров уже на трёх кластерах, мы получаем один кластер на татарском языке, и два кластера на русском языке. При разделении на восемь кластеров отчетливо выделились два кластера на татарском языке и 6 кластеров на русском языке. Таким образом, мы предположительно получаем 6 различных групп УМК на русском языке и 2 на татарском языке (таблица 3). Принадлежность кластеров к определенным УМК было

определено на основе сравнения текстов поурочного планирования, введенных в систему педагогами с текстами из учебников по соответствующим УМК.

Таблица 3.

Результаты кластерного анализа

№ кластера	Количество учителей в кластере	Язык кластера	Средняя оценка учеников
0	314	Русский	4.01
1	503	Русский	3.95
2	154	Русский	3.97
3	502	Русский	3.90
4	223	Русский	3.97
5	117	Русский	4.03
6	124	Татарский	3.93
7	63	Татарский	4.00

Анализ результатов кластеризации показал, что разработанные программные средства позволяют корректно сгруппировать отдельные группы тексты поурочного планирования в соответствующие УМК. Также для каждой группы учителей (рис. 2) была рассчитана средняя оценка учеников, обучающихся по соответствующему УМК.

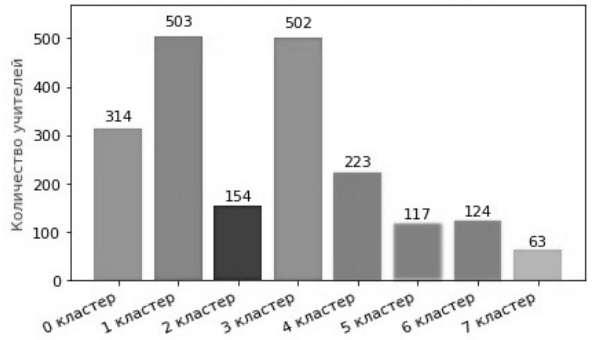


Рис. 2. Диаграмма распределения педагогов по кластерам

5. Заключение

В результате проведенного исследования, посвященного классификации текстов для определения принадлежности к соответствующему УМК, в 4 классе по дисциплине математика были выделены 8 кластеров, два из которых на татарском языке. Для каждого кластера был подобран предполагаемый учебно-методический комплекс и проведён сравнительный анализ успеваемости учеников. Разработанный программный комплекс можно использовать для автоматической обработки текстов поурочного планирования в целях определения УМК для всех предметов и классов.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта «Цифровая модель формирования индивидуальной траектории профессионального развития учителя на основе больших данных и нейросетей (на примере Республики Татарстан)», № 19-29-14082.

ЛИТЕРАТУРА

1. Ajah, I.A. (2019) Nweke, H.F. Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications. Big Data Cogn. Comput. 3, 32.

2. AnickP., Vaithyanathan S. (1997) Exploiting Clustering and Phrases for Context-Based Information Retrieval. ACM SIGIR Conference

3. Angelova R., Siersdorfer S. (2006) A neighborhood-based approach forclustering of linked document collections. CIKM Conference.

4. Chakrabarti K., Mehrotra S. (2000) Local Dimension reduction: A newApproach to Indexing High Dimensional Spaces, VLDB Conference.

5. Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. (1992). Scatter/Gather: a cluster-based approach to browsing large document collections. In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '92). Association for Computing Machinery, New York, NY, USA, 318–329.

6. Fisher D. (1987) Knowledge Acquisition via incremental conceptual clustering. Machine Learning, 2: pp. 139–172.

7. Javidi G., Rajabion L. and Sheybani E.(2017) "Educational Data Mining and Learning Analytics: Overview of Benefits and Challenges," 2017 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1102-1107.

8. Manning et al. (2008) Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. Introduction to information retrieval, volume 1. Cambridge university press Cambridge.

9. Mullner D. (2011). Modern hierarchical, agglomerative clustering algorithms, ArXiv, abs/1109.2378

10. Mikolov et al. (2013) Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. Proceedings of ICLR 2013, pages 1–12.

11. Rocklin M. Dask: Parallel Computation with Blocked algorithms and Task Scheduling, Proceedings of the 14th Python in Science Conference, 130 - 136 ,2015.

12. Singhal, A. (2001). "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43.

13. Susan T. (2005). "Latent Semantic Analysis". Annual Review of Information Science and Technology. 38: 188–230.

14. Youssra Riahi (2018). Big Data and Big Data Analytics: Concepts, Tyoes and Technologies – Moroko.

15. Zamir O., Etzioni O., Web Document Clustering: AFeasibility Demonstration, Proc. ACM SIGIRconference on Research and development in informationretrieval, New York, USA, pp. 46-54, 1998.

16. Бермудес С.Х.Г., Керимова С.У. (2016). О методе определения текстовой близости, основанном на семантических классах // Инженерный вестник Дона. № 4(43)

17. Гришин В.Д. (2018). Метода анализа и поиска заимствований в тексте // Проблемы науки. №7 (31)

18. Школьный гид (2018). Программы начальной школы [Электронный ресур]. URL: https://schoolguide.ru/index.php/progs.html

19. Попова Е.С., Спицын В.Г., Иванова Ю.А. (2019). Использование искусственных нейронных сетей для решения задачи классификации текста. Труды международной конференции по компьютерной графики и зрению "Графикон", - С. 270-273

20. Утёмов В.В., Горев П.М. (2018). Развитие образовательных систем на основе технологии BigData // Научно-методический электронный журнал «Концепт». - №6 (июнь). – С. 449-461.

Orujova S.A.
Azerbaijan State University of Economics,
Azerbaijan, Baku

NEW APPROACHES TO THE ENGLISH LANGUAGE TEACHING AT AZERBAIJAN STATE UNIVERSITY OF ECONOMICS

Abstract. New syllabus of the department of "Foreign Languages" of Azerbaijan State University of Economics is analyzed in the article. This new syllabus reflects quite a new approach to the English language teaching. Presenting a syllabus, which reflected all four (reading, listening, writing and speaking) language skills will surely be able change foreign language knowledge and mindset of undergraduates of the university. State-of – the-art technologies of ASEU (UNEC) will also provide conditions to apply changes shown in syllabus. Teaching of different foreign languages at secondary schools and higher educational institutions has always been priorities at different parts of the republic of Azerbaijan. That's why year by year Azerbaijan University of Languages (formerly called Azerbaijan State Institute of Languages named after M.F.Akhundov, then Azerbaijan State University of Languages) has provided a lot of educational institutions with future foreign language teachers.

Keywords: *teaching, state-of-the-art technologies, writing skill, syllabus, business English.*

Оруджева С.А.
Азербайджанский государственный экономический университет, Азербайджан, Баку

НОВЫЕ ПОДХОДЫ К ПРЕПОДАВАНИЮ АНГЛИЙСКОГО ЯЗЫКА В АЗЕРБАЙДЖАНСКОМ ГОСУДАРСТВЕННОМ ЭКОНОМИЧЕСКОМ УНИВЕРСИТЕТЕ

Аннотация. В статье анализируется новый учебный план кафедры "Иностранных языков" Азербайджанского государ-

ственного экономического университета. Эта новая учебная программа отражает совершенно новый подход к преподаванию английского языка. Учебный план включает в себя четыре языковых навыка: чтение, прослушивание, писание и разговор — которые, безусловно, изменят как знание иностранного языка, так и мировоззрение студентов университета. Современные информационные технологии Азербайджанского государственного экономического университета (UNEC) могут создать условия для применения изменения отраженных в учебной программе. На наш взгляд, новая учебная программа станет новым шагом в преподавании английского языка не только в Азербайджане, но и в Кавказском регионе.

Ключевые слова: *обучение, современные технологии, навыки письма, учебный план, деловой английский.*

During the post-Soviet period foreign language teaching differed more than it does nowadays. Lack of state-of – the-art technology, bad equipped auditoriums had a great impact on this falling behind. While taking part at English lessons you could observe only thin hardcover books on the desks of the classrooms and small tape-recorders playing some listening tracks in the English language. But this case met the requirements of those times. Foreign language teaching, especially English language teaching was not of a great importance: Azerbaijan Republic had not gained its state independence, political and economic relations with other foreign countries hadn't been developed. In other words, the dependent country hadn't integrated to different and foreign world, not opened its doors to investors.

But the unforeseen moment came. In 1992, on October the 18th Azerbaijan got its independence and everything changed in the minds of the population: the decisions, the steps the people were going to take in further development of economy, politics and even in education system of the country. Development of economic relations of the country with foreign ones meant improvement of English language teaching. It was stimulated by " Contract of the Century". The Agreement was signed in Gulistan Palace of Baku on September 20, 1994 which was later named as the Contract of the

Century due to its tremendous importance. 13 companies from 8 countries participated in signing of the Contract of the Century. This contract was a flow of English speaking people to Azerbaijan, new workplaces where people spoke the global language – English. So preparation of army of foreign language teachers and the teaching process itself changed a lot. If we take a glance at the lessons at Azerbaijan University of Languages, we can see new textbooks and belles-lettres in the languages taught, especially in English. Certainly it changed the atmosphere of the lessons at the University. The students were able to hear native speakers on CDs attached with books, watch DVDs in English; moreover, see the lively dialogues performed by actors and actresses abroad. While teaching a foreign language many problems can be met. For choosing methods of foreign language teaching attention should be paid on aims of teaching a foreign language. Nowadays compared to a teaching of Azerbaijani language to native auditorium varies from teaching of the English language to the same auditorium. Teaching English at Azerbaijan State University of Economics has a great and interesting history. "As known, this international foreign language has being taught at Azerbaijan universities for many years. A bright example of it is UNEC. For about 9 years ago future economists of Azerbaijan could learn Business English for three years at the university" [1; p.167]. Teaching Business English was quite a new approach for Azerbaijani learners. Because they had got used to simple general words and word-combinations. Firstly, Business English had been taught to the students for three years. The first textbook on Business English was "English for businessmen" in six volumes by I. F. Jdanova, O. E. Kudryacheva, N. S. Popova and others. Then teachers had to take different range of words and make up business situations. There were sample of formal business letters in the all volumes of the book. But it was of great importance to take those books.

As new oil contracts were signed between the country and other foreign ones, the lecturers had to prepare future economists being able to communicate with foreign businessmen abroad and within the country. Then other textbooks were "English for Businessmen" by O.I.Antonov, "Business English" by A.Abbasova, "English on Economics" by N.Nabiyev, A.Jafarova. In spite of all these attempts,

it was impossible to teach students English comprehensively. In order to possess a foreign language thoroughly, the person teaching it should try to develop all language skills simultaneously: reading, writing, speaking and listening. It should be noted that all of these four skills should be taught to learners while they are at school. You should learn all four skills if you want to have full access to the language as native speakers do. There are some reasons in which you may not want to develop all four major language skills. One of them is when your target language doesn't allow for use of the four skills. But this is quite different for Azerbaijani learner. This language allows the student to use all four skills. But how to be able to do it here at the university majoring on quite a different specialty? This was the most challenging matter for the department of "Foreign Languages" of Azerbaijan State University of Economics. But the head of the higher institution had aimed it and wanted to meet its newly enrolled students with a new syllabus. If as a lecturer of a university wants his or her student to possess a foreign language which is taught by him or her perfectly, the lecturer should to be able to share the time of the lesson effectively; to get the students acquainted with a wide vocabulary, get them ready to listen to the material and to do other activities. This surely was the most challenging one. Nowadays the syllabus, the lessons of which are taken by English teachers of "Foreign languages" department satisfies the demands of today's education. The goals of the teaching of the subject are mastering the students reading, writing, listening and speaking skills, developing a skill of usage of new words and expressions, enhancement of resource of synonyms, comprehending listening tracks, teaching students habits of expressing their ideas freely in English. The tasks of the subject are:

- Development of dialogue and monologic speech
- Introduction to the development of writing speech
- Increase of vocabulary
- Explanation of context of texts
- Correction of pronunciation

As it's seen, innovation made to a new syllabus was mainly of addition of audio and writing materials. "Students can improve their listening skills- and gain valuable language input-through a

combination of extensive and intensive listening material and procedures"[2;p.303]. Being engaged in listening and comprehending a foreign language material, may further develop the speaking skill of learners. As first of all, language learners hear the listening material, write down unknown words they catch while listening to it. Passionate students are eager to look up those words in dictionaries. So they get a wide vocabulary and it helps their speaking skills to improve. Most of students of universities are able to correct their pronunciation by listening the right form and repeating them after a speaker. Listening shouldn't be limited with an independent textbook on listening. Moreover, it includes audio form of the texts to be taken for reading and discussion. It's advisable to make students ready to listen to the text firstly, then to do reading exercises. These may have effective feedback. Listening shouldn't be only intensive one, a lecturer may tell his learners to listen to a definite TV or radio program and for the next lesson to analyze and discuss it on their own or in group. The main textbook taken in order to improve listening skill of students of Azerbaijan State University of Economics for the first term of 2019-2020 academic year is "Basic Tactics for Listening" (3rd edition) by Jack C. Richards and Grant Trew . The book is very interesting and diversified and presents listening exercises of different kinds. These exercises include matching, true (false) exercises, filling in the blanks with suitable words and word-combinations. Before doing listening exercises the lecturers should give some instructions. What are they? Which of them are given to the first year students of UNEC (Azerbaijan State University of Economics) before starting listening?

1. An efficient technique is to look at the questions and answers. Work out why the answer is correct without listening to the recording
2. Pay particular attention to analyzing multiple-choice questions
3. Don't hurry to choose right answers. Try to listen to the end. Because mostly key to right answers are at the end of the tracks

One of the most interesting features of the book is total review of four units at the back pages. It gives a lot of opportunities to the lecturers to check their listening skills on each four lessons.

Practicing listening skills is very important for Azerbaijani auditorium. Because at listening lessons a foreign language teacher make the students to listen to a foreign speech, practice and correct their pronunciation. According to the new syllabus compiled by the department of "Foreign languages" of ASEU, listening skills of students of the university are evaluated three times a term. Listening shouldn't be limited only listening to CDs. As it is done at English lessons of the university, during and also after the lessons conversations on different topics are held at the university. That's why most of them do their bests to learn new words and also use these words in their conversations. One of the way in foreign language teaching is to make a lively situation where a teacher invites her co-worker to the lesson and they may deliberately speak on the weather, attendance of students at the university, etc. In this way student listen to teachers speaking English a speech that doesn't greatly differs from one from native speaker′s. In order to increase vocabulary of students, the lecturer should also make the students to do reading exercises. That's why a new textbook "Aim High" (Elementary) 3-rd edition by Tim Falla, Paul A.Davies, Paul Kelly for the first semester and the same textbook but pre-intermediate level are valuable resources for learners. In this way an instructor also try to improve reading skills of her or his students. One of the distinguished feature of the textbook is that an author gives the sound track of the texts which are given to be analyzed and discussed. Writing skills are taught to students by "Writing book", with the help of which undergraduates firstly learn how write paragraphs, then they are explained the rules of writing academic essays.

REFERENCES

1. Orujova S. The process of English teaching and its history in UNEC. Filologiya məsələləri, №2, Bakı "Elm və təhsil" , 2016 səh.166-171.

2. Jeremy Harmer.The Practice of English language teaching. UK, "Pearsons-Longman", 2007, pp.448.

Ochirova N.G.
Southern Scientific Center RAS,
Russia, Rostov-on-Don

APPLICATION OF COMPUTER TECHNOLOGIES IN PRESERVATION, STUDY AND DEVELOPMENT OF THE KALMYK LANGUAGE

Abstract. In the context of the processes of globalization, the real threat of the disappearance of the Kalmyk language, new forms and means are required for its preservation. The article highlights the experience of Kalmyk scientists in the use of information and communication technologies as one of the most important ways to study, preserve and develop the Kalmyk language.

Keywords: *Kalmyk language, preservation, information and communication technologies, electronic resource, globalization.*

Очирова Н.Г.
Южный научный центр РАН,
Россия, Ростов-на-Дону

ПРИМЕНЕНИЕ КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ В СОХРАНЕНИИ, ИЗУЧЕНИИ И РАЗВИТИИ КАЛМЫЦКОГО ЯЗЫКА[1]

Аннотация. В контексте процессов глобализации, реальной угрозы исчезновения калмыцкого языка требуются новые формы и средства для его сохранения. Статья освещает опыт ученых Калмыкии в области использования информационно-коммуникационных технологий, как одном из важнейших способов изучения, сохранения и развития калмыцкого языка.

Ключевые слова: *калмыцкий язык, сохранение, информационно-коммуникационные технологии, электронный ресурс, глобализация.*

[1] Публикация подготовлена в рамках реализации ГЗ ЮНЦ РАН № гр. проекта АААА – А19 – 119011190182-8.

Тюркские и монгольские этносы контактируют вот уже многие тысячелетия на достаточно ограниченной территории Центральной Азии, поэтому не удивительно, что в их языках накопилось заметное количество общих элементов, причем как в лексике, так и в грамматике. Подсчет Вл. Л. Котвича, произведенный им в процессе изучения взаимодействия алтайских языков, показал, что между тюркскими и монгольскими языками выработалось около 50% общих элементов в грамматике и около 25% в лексике [1]. Это очень большие проценты сходства. В связи с тем, что взаимодействие монгольских языков уходит своими корнями в глубокую древность, в монгольских языках имеются явные тюркизмы, носящие общемонгольский характер. Они проникли в монгольские языки еще на уровне общемонгольского праязыка и представлены во всех монгольских языках, в том числе калмыцком. Поэтому представляется, что вопрос с состоянием калмыцкого языка на современном этапе, несомненно, будет представлять интерес на форуме.

Сила языка, как средства общения и воздействия на формирование личности, в настоящее время особенно возрастает. Современные процессы демократизации оказывают благотворное влияние на подъём национального самосознания, способствуют новому осмыслению истории своего народа, его культуры, традиций, обычаев, проблем национального языка. И это не случайно. Ведь в языке заключён дух этноса, отражено его национальное сознание. В языке отражается и философия народа, пройденный им социальный путь, его история и культура. В этом смысле язык следует рассматривать как уникальное средство воспитания человека. Язык, как инструмент, используется для передачи интеллектуальных знаний, морально-этических норм, эстетических ценностей народа. Изучение родного языка в таком ракурсе, т. е. во взаимосвязи с философией, историей, литературой, культурой будет более эффективно способствовать сохранению и сбережению национального языка и культуры.

Язык является одним из важнейших элементов культуры любого народа и важнейшим фактором сохранения этнического

самосознания. Неразрывное единство национального языка и национальной культуры обретает в жизни народа многообразные воплощения. Языковые же факты и феномены культуры не поддаются строгому разграничению, потому что изменения, происходящие в сфере культуры, отражаются в языке, а языковые эволюции в свою очередь стимулируют ход культурных процессов. Культура, представляя собой наивысший семиотический уровень, принадлежит в равной степени, как к миру языка, так и к миру действительной жизни людей. Наиболее полно диалектика понятий «национальный язык» и «национальная культура», основательно изученная Вильгельмом фон Гумбольдтом, раскрывается в связи с общефилософской проблемой соотношения мышления и языка. По образному выражению Гумбольдта: «мир, в котором мы живём, есть ... именно тот мир, в который нас помещает язык, на котором мы говорим» [2]. Складывающаяся в процессе развития национальной культуры система понятий отражает присущее данной нации восприятие действительности и выражает его в языке. При этом, чем сильнее воздействие национальной культуры на язык, тем богаче и специфичнее развитие последнего; с другой стороны, от богатства языка зависит глубина и полнота выражения им национальной самобытности.

Политика государства в области языка является одним из социальных факторов, оказывающих влияние на функционирование и развитие языков и определяющих его общественно-политическую жизнь. В настоящее время приоритетными направлениями в языковой политике являются установление межэтнической толерантности в обществе, сохранение и развитие языков, формирование общероссийской идентичности. Руководство и исполнительные органы власти, муниципальные образования и общественные организации Калмыкии вот уже в течение трех последних десятилетий проводят целенаправленную работу по развитию языков народов, проживающих на территории республики, сохранению национально-языкового согласия в регионе. В Калмыкии действует государственная программа «Развития государственных языков Республики Калмыкия», рассчитанная

до 2020 года включительно, направленная на повышение роли государственных языков в степной республике, разработку и внедрение современных технологий развития и сохранения государственных языков. Министерством образования и науки Республики Калмыкия разработаны и реализуются проекты "Сохранение, развитие и совершенствование калмыцкого языка", "Развитие национальной школы", была принята и реализовывалась республиканская целевая программа "Калмыцкий язык и языки народов Республики Калмыкия» и др.

Вместе с тем, несмотря на предпринимаемые меры, реализация языковой политики в Калмыкии сталкивается с множеством проблем объективного и субъективного характера, среди которых можно выделить следующие: 1) этническая функция калмыцкого языка как национального символа превалирует над коммуникативной, являющейся основной функцией языка; 2) наличие у калмыцкого языка кодифицированной нормы; 3) наличие в калмыцком языке традиционных диалектов; 4) высокий статус (престижность) русского языка и низкий статус (непрестижность) калмыцкого языка для основной массы населения, не владеющих родным языком; 5) неконкурентоспособность калмыцкоязычных СМИ по отношению к русскоязычным в количественном и качественном планах; 6) непоследовательная языковая политика: с одной стороны, поддержка калмыцкого языка, с другой — ограничение (не использование в практике властных структур калмыцкого языка как государственного и пр.), что сказывается на его престиже; 7) отсутствие четкости в распределении функций государственных языков [3].

На сегодня калмыки в силу объективных и субъективных причин практически не владеют родным языком, вполне удовлетворяя свои потребности при помощи русского языка, который обслуживает все сферы социальной жизни. Калмыцкий язык остается практически не востребованным социальной средой, у большей части населения в силу различных причин отсутствует мотивация общения на калмыцком языке. Развитие процесса языковой аккультурации, зародившейся в годы

советской власти, обусловлено многими причинами, среди них унификация образования на базе русского языка, как языка межнационального общения. Этот курс породил нигилизм по отношению к собственным языкам у части народов, особенно у молодежи [4]. Постепенно исчезала мотивация общения на калмыцком языке, предпочтение продолжает отдаваться русскому языку, которым на сегодня владеют практически все калмыки, в отличие от своего родного. Так сложилось, что у них нет надобности в общении на калмыцком языке, так как русский язык обслуживает все коммуникативные потребности. Калмыцкий язык не функционирует ни в сфере государственного управления, ни в общественно-политической деятельности, частично он используется в сферах образования и массовой коммуникации [5]. Отрицательную роль в современном критическом состоянии калмыцкого языка сыграла и 13-летняя ссылка калмыцкого народа в Сибирь в годы сталинских репрессий. Данные социологических исследований последних лет свидетельствуют о том, что почти треть выборки (31,7 %) не удовлетворена теми мерами, которые предпринимаются в республике по сохранению и развитию калмыцкого языка. В трех рассматриваемых группах (жители города, райцентров и сел) количество недовольных оказалось одинаковым. На наш взгляд, это свидетельствует о том, что существует устоявшееся мнение, выражающее определенную степень неудовлетворенности тем, как на деле реализуется Программа сохранения и развития калмыцкого языка. Противоположную точку зрения высказали 37,4% опрошенных (31,1 % живущих в райцентре, 42,4% горожан; 43,4% сельчан). Их вполне устраивает то, как реализуются меры, направленные на сохранение и развитие языка. Что касается прогнозов, то основная часть опрошенных не верит в улучшение языковой ситуации, считая, что через некоторое время калмыцкий язык исчезнет. И только 21 % респондентов полагает, что положение выправится, калмыцкий язык укрепит свои позиции. При этом более оптимистично настроены женщины (24 % против 17 %; отрицательный ответ дали 39 % женщин и 45 % мужчин).

Интересным оказалось то, что чем старше респондент, тем в большей степени он верит в изменение языковой ситуации к лучшему и такого же мнения 18 % опрошенных до 20 лет. 50,8 % юных респондентов не выразили уверенности в улучшение состояния языка (самое максимальное количество ответов «нет»). Следующая возрастная группа, 30-39-летние респонденты, проявили полное согласие в мнении о том, что многое в процессе улучшения состояния калмыцкого языка зависит от финансирования (66 %). Из этой же возрастной группы 29 % респодентов отметили необходимость разработки функциональной учебной программы с учетом незнания детьми языка. Четверть из опрошенных в возрасте 40-49 лет считают, что обслуживание в различных ведомствах, учреждениях, в сфере услуг необходимо организовать на калмыцком языке. Они также уверены, что общение в семье на родном языке также будет способствовать реальному улучшению состояния калмыцкого языка. Почти треть респондентов из возрастной группы 50-59 лет, предлагая те же меры, отмечает, что нужны современная литература, новые учебники, оформление вывесок на калмыцком языке, создание мультироликов, использование передовых технологий по обучению языку и др. Поэтому одним из эффективных мер сохранения и овладения языком на современном этапе является применение информационно-коммуникационных технологий, без которых сегодня в мире практически не обходится ни одна область человеческой деятельности.

В Калмыкии, начиная с 2000-х гг. в работу министерств и ведомств, учреждений и организаций начали активно внедряться новые современные технологии. Сегодня уже очевидно, глубокое проникновение информационных технологий во все сферы жизнедеятельности региона: в управленческую, учебно-воспитательную, образовательную и др. Активно внедряются информационные технологии в деятельность научных учреждений республики, в систему высшего и среднего образования республики [6]. В условиях процессов глобализации, реальной угрозы утраты родного языка

требуются новые формы и средства по поддержке и сохранению языков. Одним из таких мер явилась компьютерная обработка. Важно перевести "древнее слово", заложенное в национальных языках и письменности в современность, используя возможности цифровизации. Работа в этом плане началась в республике с 2000–х гг. С этого времени калмыцкий язык начинает обретать новую жизнь в сети Интернет. Во всех направлениях науки региона и, прежде всего в гуманитарной, создаются различные программные продукты (н-р, TextAligner), информационно-справочные системы, электронные словари различного типа и др. В 2010 году в Калмыцком институте гуманитарных исследований РАН была начата работа над созданием Национального корпуса калмыцкого языка по проекту Программы фундаментальных исследований Президиума РАН (рук. Н.Г.Очирова). В 2012 году в КИГИ РАН (ныне КалмНЦ) был создан Отдел теоретической и экспериментальной лингвистики. В течение трех лет работники отдела занимались дальнейшей разработкой программы, наполнением Нацкорпуса. В него были включены почти 7 тыс. произведений калмыцкой художественной литературы, в том числе прозаические, поэтические и фольклорные произведения, начиная с 1950-х гг. по 1980-е гг. прошлого столетия. Кроме того, вошел архив газеты «Хальмг унн» (Калмыцкая правда) за последнее десятилетие. Затем ученые приступили ко второму этапу — разработке подкорпусов. Сначала была проведена работа по созданию морфемного подкорпуса, что давало возможность с помощью его провести морфемный анализ любого слова. Более сложной стала разработка старокалмыцкого подкорпуса. В 2013 году Национальный корпус калмыцкого языка был запущен во всемирную сеть. Первый крупный электронный ресурс был создан в целях сохранения и развития калмыцкого языка при финансовой поддержке гранта РГНФ (рук. В.В. Куканова) и ориентирован, в первую очередь на молодежь, ученых, учительство, студентов и школьников, на всех, кто изучает и исследует калмыцкий язык. Объем Национального корпуса калмыцкого языка составлял около 10 миллионов словоупотреблений,

последующая работа должна была этот показатель увеличить до 30 миллионов слов. Национальный корпус калмыцкого языка является фундаментальным проектом, направленым на фиксацию и сохранение калмыцкого языка, создание условий для его возрождения и развития как полноценного средства коммуникации [7]. Он позволит решать как фундаментальные теоретические, так и практические задачи: впервые был введен текстовый материал большого объема в научные исследования и в процесс составления словарей различного формата и характера, что в свою очередь дало возможность расширить базу научно-теоретических исследований и углубить имеющееся описание лексики и грамматики калмыцкого языка. Кроме того, ученые Института разработали и другие электронные продукты для изучения, сохранения и развития национального языка. Это «Частотный словарь калмыцкого языка», двуязычные словари и другие электронные базы данных [8].

Необходимо отметить, что калмыцкий язык все активнее находит свое место в глобальной сети. Появились энтузиасты, продвигающие язык в интернет-пространстве. Так, с начала 2014 года в интернет-магазине AppStore стало доступно новое бесплатное приложение по обучению калмыцкому языку под названием «Хальмг келн» (Калмыцкий язык). Его разработал уроженец Калмыкии Алексей Зунов, выпускник Московского физико-технического института (МФТИ). В созданном им Приложении — два словаря: русско-калмыцкий и калмыцко-русский. Кроме того, есть определенные категории слов для изучения, как дни недели, животные и т.д. За основу А.Зуновым приняты опубликованные в Интернете калмыцко-русские словари, которые были оптимизированы и отредактированы им для ускорения работы Приложения. После создания своего аккаунта, Приложение было отправлено разработчиком в компанию Apple для прохождения проверки. В течение трех недель специалисты проводили испытания Приложения и затем его одобрили. Как отмечает сам разработчик, в Приложении постоянно доступны обновления. К примеру, если пользователь сделает соответствующие настройки, то он может получать

сообщения с новыми словами. В планах разработчика — добавление калмыцкой азбуки, перевод калмыцких имен и йорялей (благопожеланий) и социальной составляющей, то есть связи с популярными социальными сетями. Приложение «Хальмг келн» (Калмыцкий язык) было доступно только для iOS-устройств в AppStore. А.Зунов продолжил работу и запустил Приложение не только для платформы iOS, но и для Android. В приложении «Хальмг келн» есть раздел обратной связи, где каждый пользователь может сообщить о найденной ошибке или предложить свою идею. В настоящее время ведется активная работа над иллюстрированным алфавитом калмыцкого языка [9]. Приложение «Хальмг келн» – это попытка заинтересовать молодое поколение калмыков в изучении родного языка. Сегодня оно включает калмыцко-русский и русско-калмыцкий словари оффлайн (более 30000 слов), встроенную калмыцкую клавиатуру, режимы обучения на выбор, благопожелания, пословицы, стихотворения на калмыцком языке и обратную связь, доступно по ссылке https://play.google.com/store/apps/details?id=ru.zuno. Параллельно с А.Зуновым успешно работает в сети руководитель общественной организации «Центр по развитию калмыцкого языка» В. Манджиев, знаток калмыцкого языка и активный его популизатор. Число энтузиастов с каждым годом растет. Так, становится популярным в Сети и другой ресурс, созданный не лингвистами, но людьми, также обеспокоенными судьбой родного языка и культуры. Это «Калмыцкая электронная библиотека» (halmglib. org), задуманная для онлайн-чтения произведений калмыцких авторов и народного фольклора. Создали ее выпускники физического факультета Санкт-Петербургского госуниверситета М.Аинов и А.Цаган-Манджиев. Созданный ими сервис тестируется с помощью нескольких брошюр, разработан и новый дизайн сайта с использованием элементов национальной культуры. Создателями «Калмыцкой электронной библиотеки» являются поколения Next, не мыслящего себя без интернета, уверены, что в деле сохранения калмыцкого языка IT технологии — верные и

надежные помощники. На сайте библиотеки размещены произведения и биографии калмыцких писателей и поэтов, а сама Калмыцкая электронная библиотека доступна по ссылке http://halmglib.org/. Благодаря сети Интернет растет число полезных Приложений для изучения калмыцкого языка. К примеру, все пользователи смартфонов могут скачать эти Приложения в магазинах Google Play и App Store. Приложение «Хальмг келн» (Калмыцкий язык) сегодня включает калмыцко-русский и русско-калмыцкий оффлайн-словари. В нем более 30 тыс. слов, имеется встроенная калмыцкая клавиатура. Для желающих общаться в соцсетях на калмыцком языке, клавиатура с нужной раскладкой также доступна к скачиванию. В «Русско-калмыцком разговорнике» можно выучить слова по разным темам: дом, предметы, еда, родня и др. После изучения темы сразу же можно пройти тест и узнать, насколько хорошо вы знаете язык. Например, скачав «Кроссворды на калмыцком» — это проект фонда содействия развитию калмыцкого языка «Сәәхн келн» (Красивый язык). Авторы представили 63 кроссворда разной степени сложности. Есть вариант и для тех, кто легче запоминает новые фразы с помощью музыкальных исполнителей. Пользователи ПО «Мана дун» (Наши песни) могут послушать народные, современные песни и танцевальные мелодии. Из понравившихся треков можно создать плейлист. Чтение на досуге тоже возможно. Так, весь текст калмыцкого героического эпоса «Джангар» на калмыцком и русском языках находится в одноименном Приложении и доступен для всех желающих ознакомиться с ним.

Таким образом, информационно-коммуникационные технологии стали составной частью жизнедеятельности региона, сообщество Калмыкии пришло к осознанию необходимости внедрения их как способа повышения эффективности управленческой деятельности, как способа распространения знаний, их ротации и популяризации и как одного из способов изучения, сохранения и развития калмыцкого языка.

ЛИТЕРАТУРА

1. Котвич В. Исследование по алтайским языкам. Перевод с польского. М., 1962. С. 351.

2. Кёрнер Э.Ф. Вильгельм фон Гумбольдт К. и этнолингвистика в Северной Америке от Боаса (1894) до Хаймса (1961) // ВЯ. 1992. № 1. С. 105-113.

3. Омакаева Э.У., Горяев А.Т. Язык как фактор национально-культурного возрождения (Республика Калмыкия сквозь призму реформы калмыцкой орфографии) // Народы Калмыкии: перспективы социокультурного и этнического развития. Элиста: Калм.ГУ, 2000. – С.45-47.

4. Губогло М.Н. К изучению перспектив развития двуязычия у народов СССР // История СССР. 1978. №1.

5. Катушов К.П. Калмыкия в геопространстве России / КИГПИ, Элиста: АПП«Джангар», 1998. 336 с.

6. Очирова Н.Г. Проблемы развития региона в условиях трансформации российского общества (на примере Республики Калмыкия) // Известия высших учебных заведений. Серия: Общественные науки. Ростов-на-Дону. 2011. № 6 (166). С. 68-72.

7. Куканова В.В., Бембеев Е.В., Мулаева Н.М., Очирова Н.Ч. Национальный корпус калмыцкого языка: архитектура и возможности использования // Вестник Калмыцкого института гуманитарных исследований РАН. 2012. № 3. С. 138–150.

8. Куканова В.В., Каджиев А.Ю., Бембеев Е.В. Электронные двуязычные словари калмыцкого языка // Исследование проблем исчезающих языков в условиях глобализации (на примере калмыцкого языка) (VIII Волковские чтения): Мат-лы Международной научно-практической конференции (23–26 октября 2013 г.) Элиста: Изд-во Калм. ун-та, 2013. С. 53–56.

9. Электронный ресурс: http://journal.bashkort.org/ru/tag/tsifrovizatsiya-yazyka/ (последнее обращение – 15.04.2020).

Tukeyev U.A.
Kazakh National University. al-Farabi,
Kazakhstan, Almaty

KAZAKH LATINICA WITH FULL ENGLISH ALPHABET

Abstract. The article discusses a variant of the Kazakh Latin alphabet, which includes 26 letters of the Latin alphabet. This allows when typing on the keyboard texts mixed with formulas, usually using Latin letters, to perform typing without changing the alphabet. It is also convenient when the text is interspersed with Latin letters, for example, names in English. For comparison: the Turkish alphabet without letters Q, W, X; Azerbaijani alphabet without the letter W; Uzbek alphabet without letters C, W; Kazakh alphabet in Latin, approved in February 2018 without letters C, W, X. That is, if the text contains formulas or English words that include missing letters, then you need to switch to the English alphabet, which causes inconvenience when typing the text.

Keywords: *Kazakh language, Kazakh alphabet, phonetic system of the Kazakh language , Latin alphabet*.

Тукеев У. А.
Казахский национальный университет им. аль-Фараби,
Казахстан, Алматы

КАЗАХСКАЯ ЛАТИНИЦА С ПОЛНЫМ АНГЛИЙСКИМ АЛФАВИТОМ

Аннотация. В статье рассматривается вариант казахской латиницы, включающий 26 букв латинского алфавита. Это позволяет при наборе на клавиатуре текстов смешанных с формулами, обычно использующие латинские буквы, выполнять набор текста без смены алфавита. Это удобно также когда текст имеет вкраплины из латинских букв, например, названия на английском языке. Для сравнения: турецкий алфавит без букв Q, W, X; азербайджанский алфавит без буквы W; узбекский алфавит без букв C, W; казахский алфавит на латинице, утвержденный в феврале 2018 года без букв C, W, X. То есть, если в тексте встречаются формулы или английские слова,

которые включают отсутствующие буквы, то надо переключаться на английский алфавит, что вызывает неудобства при печатании текста.

Ключевые слова: *казахский язык, фонетическая система казахского языка, казахский алфавит, латинский алфавит.*

В предлагаемом нами варианте алфавита казахского языка на латинице устранены недостатки второго утвержденного варианта латинского алфавита казахского языка.

При разработке данного варианта алфавита казахского языка на латинице мы придерживались следующих критериев:

1) буквы алфавита должны отражать фонемы казахского языка (одна буква – один звук);

2) фонемы и буквы заимствованные из русского языка отображаются диграфами;

3) буквы латинского алфавита должны максимально соответствовать английскому (международному) произношению.

Так, в последнем варианте алфавита казахского языка буква ‘**U’** с мягким произношением [ju:] на английском языке представляет в казахском алфавите букву ‘**Ұ’** казахской кириллицы**,** являющейся твердым гласным звуком. С нашей точки зрения, правильней будет, если буква ‘**U’** будет представлять букву ‘**ы**’ казахской кириллицы. Соотвественно, правильней будет, если диакритика **Ŷ** будет представлять букву ‘**Ұ’** казахской кириллицы, так как английская буква ‘**у’** обычно читается как «ы» в кириллице, а казахская буква ‘ұ’ близка по произношению к ‘ы’, чем к букве ‘у’ кириллицы, которая в английском языке представлена буквой ‘**u’** латиницы. Например, написание английского слова «university» было представлено на казахской латинице как «ýnıversıtet». В предлагаемом варианте алфавита это слово будет записано как «unıversıtet».

В предлагаемом латинском алфавите казахского языка предлагается оставить все 26 букв латиницы. Спорным представлялось оставление буквы ‘X’, однако, в казахском техническом языке он также может использоваться, например, X-saule, X-koordinata. Поэтому буква ‘X’ оставлена в латинском алфавите казахского языка. Она также может быть использована для представления буквы ‘Xx**’** казахской кириллицы, таким образом, устраняется разработка специальных орфографических правил для применения и обучения при использовании буквы ‘Hh**’** в последнем варианте казахской латиницы.

Таблица 1.

Алфавит казахского языка на латинице

№	казахская латиница	казахская кириллица
1	Aa	Аа
2	**Áá - álem**	Әә - әлем
3	Bb	Бб
4	Cc - cıkl	Цц - цикл
5	Dd	Дд
6	Ee	Ее
7	Ff	Фф
8	Gg	Гг
9	**Ǵǵ- aǵa**	Ғғ-аға
10	Hh	Нһ
11	Ii	Іі
12	**Íı – Ínstıtut, qoı**	Ии - институт; Йй -қой
13	Jj	Жж
14	Kk	Кк
15	Qq	Ққ
16	Ll	Лл
17	Mm	Мм
18	Nn	Нн
19	**Ńń - meniń**	Ңң - менің
20	Oo	Оо
21	**Óó - ózen**	Өө - өзен
22	Pp	Пп
23	Rr	Рр
24	Ss	Сс
25	**Śś- qaraśy**	Шш - қарашы
26	Tt	Тт
27	Uu	Уу
28	**Úú - kún**	Үү - күн
29	Vv	Вв
30	Ww - Wali	Уәлі
31	Xx	Хх
32	Yy	Ыы
33	**Ýý - qýlyp**	Ұұ- құлып
34	Zz	Зз

Специфическое использование сделано для латинской буквы C. Это осуществлено с целью сокращения диграфов.

Весьма важным является использование в казахской латинице буквы ‘W’, потому что в казахском языке существует отдельная фонема [w].

Весь список новых диакритических символов составляет всего 8 единиц. Полный список букв нашего алфавита представлен в таблице 1. Там же представлены примеры написания слов для новых диакритических букв и букв ‘C’, ‘W’.

Таблица 2.

Новые диакритические буквы алфавита казахского языка на латинице

1	2	3	4	5	6	7	8
Áá	Óó	Ýý	Úú	Ǵǵ	Ńń	Íı	Śś
ə	ө	ұ	ү	ғ	ң	ий	ш

Итого в предлагаемой казахской латинице 34 букв: 26 + 8.

Предлагается использовать несколько диграфов, принятые и во многих других языках. Это ch – буква ‘ч’, и очень редкие для казахского языка (для заимствованных слов из русского языка): ıo- буква ‘ё’, ıa – буква ‘я’, ıu – буква ‘ю’, śś – буква ‘щ’. Ниже представлена таблица диграфов:

Таблица 3.

Диграфы казахского языка на латинице

1	**2**	**3**	**4**	**5**
ıo	**ch**	**śś**	**ıu**	**ıa**
ё	**ч**	**щ**	**ю**	**я**

Новые диакритические буквы: они обозначены символами, которые по звучанию являются близкими. Так, буквы ‘**ы’** и ‘**Ұ’** являются «жуан» (твердыми) звуками, соотвественно, их желательно представить одной буквой: ‘**Y’** (без и с диакритикой). И также, буквы ‘**у’** и ‘**ү’** являются «жінішке» (мягкими) звуками, соотвественно, их желательно представить одной буквой: ‘**U’** (без и с диакритикой).

Yy	**Ýý**	**Uu**	**Úú**
ы	**ұ**	**у**	**ү**

Для сравнения рассмотрим текст гимна Республики Казахстан на казахской кириллице и предлагаемой казахской латинице.

Таблица 4.

Гимн Республики Казахстан

Действующй алфавит кириллица	Предложенный латинский алфавит
Алтын күн аспаны, Алтын дән даласы, Ерліктің дастаны, Еліме қарашы! Ежелден ер деген, Даңқымыз шықты ғой. Намысын бермеген, Қазағым мықты ғой	Altynkún aspany, Altyn dán dalasy, Erliktiń dastany, Elime qaraśy! Ejelden er degen, Dańqymyz śyqty ǵoı. Namysyn bermegen, Qazaǵym myqty ǵoı!
Қайырмасы: Менің елім, менің елім, Гүлің болып егілемін, Жырың болып төгілемін, елім! Туған жерім менің – Қазақстаным!	**Qaıyrmasy:** Meniń elim, meniń elim, Gúliń bolyp egilemin, Jyryń bolyp tógilemin, elim! Tuǵan jerim meniń – Qazaqstanym!
Ұрпаққа жол ашқан, Кең байтақ жерім бар.	Yrpaqqa jol aśqan, Keń baıtaq jerim bar,

Бірлігі жарасқан, Тәуелсіз елім бар. Қарсы алған уақытты, Мәңгілік досындай. Біздің ел бақытты, Біздің ел осындай!	Birligi jarasqan, Táuelsiz elim bar Qarsy alǵan waqytty, Máńgilik dosyndaı, Bizdiń el baqytty, Bizdiń el osyndaı!

Ниже рассмотрен вариант представления предложенного латинского алфавита казахского языка на клавиатуре QWERTY.

Вопрос расположения дополнительных букв на клавиатуре может быть еще дополнительно исследован с точки зрения эргономики.

Расположение 8 дополнительных букв на клавиатуре:

~	**!** **1**	**@** **2**	**#** **3**	**$** **4**	**%** **5**	^ **6**	**&** **7**	* **8**	**(** **9**	**)** **0**	**_** **-**	+ =	**Back-space**
Tab	**Q**	**W**	**E**	**R**	**T**	**Y**	**U**	**I**	**O**	**P**	**{Á** **[** **á**	**}** **Ó** **]** **ó**	**| İ** **\ı**
Caps Lock	**A**	**S**	**D**	**F**	**G**	**H**	**J**	**K**	**L**	**:** **Ý** **;** **ý**	**“** **Û** **‘** **ú**		
Caps Lock	Z	X	C	V	B	N	M	< **Ś** **,ś**	>**Ń** . **ń**	?**Ǵ** / **ǵ**			

Khakimov B.E.
Institute of Applied Semiotics of the AS of the RT,
Russia, Tatarstan, Kazan

USING THE TATAR LANGUAGE IN YANDEX SEARCHING: SOME OBSERVATIONS

Abstract. This paper presents some observations on web search queries in the Tatar language. The Republic of Tatarstan is an asymmetric bilingual community with elements of diglossia. Search queries are good at showing how ordinary people use language on the web, and reflect how much the language is involved in the digital reality. The article analyzes the reasons for the low frequency of search queries in the Tatar language in comparison with the number of active speakers. The development and increase of thematically diverse Tatar-language content, as well as the inclusion of Tatar language models in search algorithms, will also help to overcome sociolinguistic and psycholinguistic barriers, and change the behavior of Tatar-speaking users when searching for information on the web.

Keywords: Tatar language, low-resourced languages, bilingualism, diglossia, language-related search queries.

Хакимов Б. Э.
Институт прикладной семиотики АН РТ,
Россия, Татарстан, Казань

ИСПОЛЬЗОВАНИЕ ТАТАРСКОГО ЯЗЫКА В ПОИСКОВЫХ ЗАПРОСАХ ЯНДЕКСА: НЕКОТОРЫЕ НАБЛЮДЕНИЯ

Аннотация. В данной статье представлены некоторые наблюдения по запросам веб-поиска на татарском языке. Республика Татарстан — асимметричное двуязычное сообщество с элементами диглоссии. Поисковые запросы хорошо показывают, как обычные люди используют язык в Интернете, и отражают насколько язык задействован в

цифровой реальности. В статье анализируются причины низкой частоты поисковых запросов на татарском языке по сравнению с количеством активных носителей. Развитие и увеличение тематически разнообразного татароязычного контента, а также включение татарских языковых моделей в поисковые алгоритмы также помогут преодолеть социолингвистические и психолингвистические барьеры и изменить поведение татароязычных пользователей при поиске информации в Интернете.

Ключевые слова: *татарский язык, языки с ограниченными ресурсами, двуязычие, диглоссия, языковые поисковые запросы.*

1. Введение

Анализ статистики поисковых запросов в Интернете является эффективным инструментом маркетинговых, социологических, социолингвистических и других исследований. В данной статье представлена попытка применения данной методики для исследования языковой ситуации в отдельно взятом регионе.

Республика Татарстан часто характеризуется как билингвальное сообщество, но специфика местного билингвизма требует отдельного изучения. Проблеме двуязычия в Татарстане посвящено значительное число научных работ, в большинстве из которых билингвизм характеризуется как асимметричный, с элементами диглоссии [1-4]. С точки зрения владения языками, в обществе можно выделить монолингвальную (русскоязычную) и билингвальную (татарско-русскую) часть.

Татарский язык – государственный язык Республики Татарстан, юридически равноправный с русским [5]. Используется в республиканских официальных документах, средствах массовой информации, в деятельности органов государственной власти (зачастую декларативно, как «церемониальный» язык), наиболее распространен в сфере культуры и искусства, религиозной жизни и в быту.

Хотя функциональные различия между русским и татарским языком очевидны и многие выводы можно было бы сделать без специальных исследований, более детальное изучение статистики поисковых запросов помогает сделать это точнее и заметить возможные интересные тенденции. Поисковые запросы хорошо показывают, как используют язык обычные люди в неформальных ситуациях. Они также отражают, насколько язык «проник» в цифровую действительность, от этого зависят и перспективы его будущего развития.

2. Методология и материалы исследования

Для анализа поисковых запросов был использован сервис Яндекса «Подбор слов» (wordstat.yandex.ru). Данный сервис позволяет получить статистику частотности того или иного поискового запроса, динамику за последние годы по месяцам и неделям. Информация может быть представлена по географическому принципу, в частности, отдельно для Республики Татарстан или определенного города, района, населенного пункта. Кроме того, сервис предоставляет данные о связанных и частично совпадающих по составу запросах, что расширяет возможности поиска.

В исследовании приняли участие студенты Казанского федерального университета, обучающиеся по специальности «татарская филология и журналистика» (всего 20 человек). Каждый участник исследования определил 10 тем (ситуаций), мотивирующих поисковые запросы пользователей. Необходимо было выделить 5 ситуаций, интересующих лично данного участника, и 5 ситуаций, которые, по мнению участника, часто вызывают потребность поиска в Интернете у большинства людей. С учетом повторов у разных участников, всего было выделено около 30 тем/ситуаций.

Далее по каждой ситуации были сформулированы ключевые слова и словосочетания для потенциальных поисковых запросов параллельно на русском и татарском языке. Каждый студент работал со своим списком.

Была исследована следующая информация:

- Количество поисковых запросов за последний месяц
- Слова и словосочетания, встречающиеся вместе с искомым словом или фразой
- История запросов (динамика с конца 2018 г.).

3. Результаты

В результате анализа, все темы были поделены на группы в зависимости от сравнительной активности поисковых запросов на татарском языке. Эти группы хорошо коррелируют с теми сферами общественной жизни, в которых татарский язык считается более распространенным. При этом, даже в самой высокочастотной группе запросы на татарском языке уступают по употребительности русскоязычным аналогам.

1. Темы со сравнительно высокой частотностью запросов на татарском языке.

- Концерты
- Исполнители
- Фильмы, телепередачи
- Телеканалы, радиостанции
- Праздники
- Дни рождения, поздравления
- Учеба, школа
- Религия

2. Темы со средней частотностью запросов на татарском языке.

- Пища, еда
- Рецепты
- Гороскоп
- Сны, толкование
- Литература, стихи (часто привязано к школьной программе)

3. Темы с низкой частотностью запросов на татарском языке.

- Здоровье
- Природа

- Животные
- Растения
- Соцсети
- Путешествия

4. Темы с очень низкой частотностью запросов на татарском языке.

- Погода
- Жилье
- Автомобили
- Работа
- Финансы
- Ремонт
- Мебель

Состав данных групп демонстрирует явно выраженную градацию и постепенное снижение уровня использования татарских поисковых запросов по мере продвижения от тематики, связанной с культурой, искусством, досугом к более «прагматичным» темам и жизненным ситуациям. Данный факт демонстрирует функциональное разграничение языков и закрепление за татарским языком определенного набора «нишевых» функций.

Первичный анализ статистики и содержания поисковых запросов также позволяет выявить следующие общие тенденции:

- большое количество поисковых запросов написано с орфографическими ошибками, специфичные татарские буквы заменяются на схожие по начертанию основные кириллические;
- поисковые запросы на татарском языке часто используются в сочетании с русскими словами, которые несут функциональную нагрузку;
- татарские слова и фразы чаще встречаются в функции цитат при поиске песен, их текстов и т.п.
- количество запросов в форме вопросительных предложений очень незначительно, что свидетельствует о

недостаточной реализации функции поиска необходимой актуальной информации;

- история запросов показывает сезонные колебания (напр., день учителя – укытучылар көне).

Можно выделить следующие причины низкой частотности поисковых запросов на татарском языке:

- отсутствие информации по многим темам на татарском языке (*проблема контента*);
- в системах информационного поиска не учитываются модели татарского языка, что приводит к снижению релевантности результатов поиска и мотивирует пользователей выполнять поиск на русском языке (*техническая проблема*);
- сформировавшаяся в среде билингвальных носителей, в их сознании диглоссия, функциональное разграничение татарского и русского языка, несмотря на официальный статус и потенциальные возможности (*социальная, фундаментальная проблема*).

Подобная ситуация, свою очередь приводит к негативным последствиям:

- возникновение своеобразного «замкнутого круга», когда сниженный спрос на татароязычный контент не стимулирует расширение подобного контента, что, в свою очередь, дополнительно демотивирует пользователей искать информацию на татарском языке;
- ограничение в доступе к информации для татароязычных пользователей
- дальнейшее вытеснение языка из цифровой реальности.

4. Заключение

Развитие и увеличение тематически разнообразного татароязычного контента, а также учет татарских языковых моделей в поисковых алгоритмах будет способствовать также и преодолению социолингвистических и психолингвистических барьеров, изменению поведения татароязычных пользователей при поиске информации в Интернете. Более высокий спрос на

информацию на татарском языке стимулирует увеличение объема и разнообразия такой информации.

Важную роль в этих процессах играют прикладные разработки в сфере автоматической обработки языка, например, голосовые интеллектуальные помощники, интеграция машинного перевода в поиск и другие решения, которые способны смягчить проблему недостаточности контента и создать комфортные полноценные условия для татароязычных пользователей в информационной среде.

Схожие процессы, вызовы и потенциальные решения в той или иной степени релевантны и для других национальных языков, подверженных негативному воздействию глобализации, вне зависимости от официального статуса и количества носителей.

ЛИТЕРАТУРА

1. Байрамова Л.К. Татарстан: языковая симметрия и асимметрия. Казань: Изд-во Казан. ун-та. 2001
2. СалимоваД.А. Двуязычие в республике Татарстан: критерии определения уровня владения языками и пути к амбилингвизму // Муниципальное образование: инновации и эксперимент. 2012. №5.
3. Гузельбаева Г.Я. Практики использования государственных языков жителями Татарстана в ситуации официального двуязычия// Вестник ТГГПУ. 2013. №4.
4. Кириллова З.Н. Двуязычие без диглоссии в Татарстане: сравнение двух периодов // *Вестник ТГГПУ. 2014. №4 (38)*.
5. Яндекс. Подбор словwordstat.yandex.ru
6. Конституция Республики Татарстан https://minjust. tatarstan.ru/ konstitutsiya.htm?pub_id=1084014.

Hakimov B.A., Shaekhov M.R.
Institute of Applied Semiotics of the AS of the RT,
Russia, Tatarstan, Kazan

ON THE QUESTION OF CREATING A PARALLEL TEST CASE FOR THE PROBLEM OF MACHINE TRANSLATION IN A RUSSIAN-TATAR PAIR

Abstract. The TatSoft neural machine translation systemwas designed specifically for the Russian-Tatar language pair and is focused on the use of language models that take into account the specifics of the Tatar language. The accuracy evaluation based on the BLEU metric shows better results in comparison with analogues. However, when using a test set from the same sources as the training data, the evaluation results may be overestimated and not reflect the actual quality of the translation, especially for individual language styles. For this reason, the developers faced the need for a more accurate and detailed evaluation using both quantitative and qualitative methods. The task was to create a balanced parallel test corpus. Reaching the representativeness of a variety of linguistic structures is more complex than stylistic representativeness. On the other hand, the variety of styles presented also affects the variety of language structures in the corpus.

Keywords: *Machine translation; Tatar language; evaluation; parallel corpus; quantitative and qualitative methods*.

Хакимов Б. Э., Шаехов М.Р.
Институт прикладной семиотики АН РТ,
Россия, Татарстан, Казань

К ВОПРОСУ СОЗДАНИЯ ПАРАЛЛЕЛЬНОГО ТЕСТОВОГО КОРПУСА ДЛЯ ЗАДАЧИ МАШИННОГО ПЕРЕВОДА В РУССКО-ТАТАРСКОЙ ПАРЕ

Аннотация. Система нейронного машинного перевода TatSoft была разработана специально для русско-татарской языковой пары и ориентирована на использование языковых моделей, учитывающих специфику татарского языка. Оценка точности на основе метрики BLEU показывает лучшие

результаты по сравнению с аналогами. Однако при использовании тестовых выборок из тех же источников, что и обучающие данные, результаты оценки могут быть завышены и не отражать фактическое качество перевода, особенно для отдельных языковых стилей. По этой причине разработчики столкнулись с необходимостью более точной и детальной оценки с использованием как количественных, так и качественных методов. Задача состояла в том, чтобы создать сбалансированный параллельный тестовый корпус.

Ключевые слова: *машинный перевод, татарский язык, параллельный корпус, оценка качества перевода.*

Введение

В настоящее время общедоступные системы нейронного машинного перевода с пользовательским онлайн-интерфейсом в русской-татарской языковой паре представлены следующими переводчиками:

1) TatSoft (https://translate.tatar/);

2) Яндекс.Переводчик (https://translate.yandex.ru/translator/Russian-Tatar);

3) Google Translate (https://translate.google.com/?sl =ru&tl =tt&op =translate);

4) Promt.One (https://www.translate.ru/).

Первые три из перечисленных систем в целом сопоставимы по точности. Последняя система появилась в августе 2020 года, и на данный момент пока отсутствуют результаты ее сравнительной оценки с остальными русско-татарскими машинными переводчиками.

В отличие от переводчиков Яндекс и Google, где татарский язык является одним из десятков языков и не относится к наиболее часто используемым, переводчик «TatSoft» разработан специально для русско-татарской языковой пары. Не предоставляя возможности многоязыкового перевода, тем не менее, он ориентирован на использование языковых моделей, учитывающих специфику татарского языка. Для создания переводчика «TatSoft» использовался параллельный корпус объемом около 1 млн. предложений, источниками для которого стали переведенные книги, двуязычные материалы из Интернета, ручной перевод текстов. Также активно применялись CAT-системы и метод обратного перевода [10].

Сравнение трех русско-татарских машинных переводчиков по тестовой выборке из 1000 случайных предложений по общепринятой методике BLEU [11] представлено в таблице 1.

Таблица 1.

Сравнение русско-татарских переводчиков

Переводчик	С русского на татарский язык	С татарского на русский язык
Yandex	15.59	18.16
Татсофт	35.39	39.21
Google	17.00	22.64

Как видно из таблицы, переводчик «TatSoft» превосходит аналоги по точности перевода в количественном измерении. Однако при использовании тестовой выборки из тех же источников, что и обучающие данные, показатели по метрике BLEU могут оказаться завышенными и не отражать действительное качество перевода, особенно по отдельным стилям. По этой причине перед разработчиками встала необходимость более точной и детализированной оценки точности машинного перевода с использованием как количественных, так и качественных методов. Для достижения данной цели была поставлена задача создания сбалансированного параллельного тестового корпуса.

Методология, источники и сбор данных

При создании параллельного тестового корпуса мы придерживались следующих основных принципов:

- ориентация на многообразие;
- представленность стилей/жанров;
- представленность языковых явлений;

Объем создаваемого тестового корпуса составляет около 2000 пар предложений, распределенных по нескольким стилям и

жанрам. Используются как тексты с готовыми переводами, так и непереведенные тексты с последующим переводом.

Следует отметить, что обеспечение представленности разнообразных языковых структур является более сложной задачей по сравнению с достижением стилистической репрезентативности. С другой стороны, разнообразие представленных стилей влияет и на разнообразие языковых структур в корпусе.

Можно использовать комбинированный подход: по результатам проведенных тестов выявляются типичные ошибки, и тексты с соответствующими языковыми структурами добавляются как в обучающую, так и в тестовую выборку. Конечно, это не является исчерпывающим решением – параллельно необходимо исследовать и повышать разнообразие языковых структур в тестовой выборке с учетом их частотности и типичных стилевых маркеров.

На основе вышеуказанных принципов для включения в состав тестового корпуса были выбраны тексты различных стилей языка – официально-делового, литературного, публицистического, научного, а также тексты специализированной тематики (кулинарные рецепты). Более подробная характеристика состава корпуса представлена в таблице 2.

Таблица 2.

Состав параллельного тестового корпуса

Источник	Кол-во предло-жений	Стиль
Официальные документы	200	Официально-деловой
Саша Денисова. «Второй месяц весны – это апрель»	200	Художественный (современная проза)
Татар-информ	500	Информационно-публицистический
“Гыйлем”	800	Научно-популярный
“Татарская кухня”	300	Специализированный (кулинарные рецепты)
Всего	2000	

Процесс пополнения тестового корпуса осуществлялся в следующей последовательности. Сначала выполнялся тестовый (выборочный) перевод текстов разной тематики по каждому стилю машинным переводчиком «TatSoft». Данные результаты

сравнивались с эталонным переводом этих текстов. Выявленные ошибки автоматического перевода были систематизированы и распределены по типам. На следующем шаге производился подбор текстов с типичными ошибками, новыми словами и конструкциями, которые отражают наиболее распространенные ошибки машинного перевода.

Таким образом, тексты для тестового корпуса отбирались по следующим критериям:

1. Содержание типичных случаев, выявленных на этапе изучения ошибок.
2. Наличие в открытом доступе: выбирались тексты, незащищенные авторским правом и/или доступные для использования без указания авторства.
3. Наличие параллельного текста на русском или татарском языке, а также возможность выполнения качественного профессионального перевода при его отсутствии.

Языковые особенности отдельных стилей

1. Официально-деловой стиль представлен текстами приказов, постановлений и других нормативных документов в объеме около 200 предложений. Переводы этих текстов уже выполнены профессиональными переводчиками.

На этапе тестового перевода типичные ошибки машинного переводчика «TatSoft» заключались в построении предложений, порядке слов и подборе терминов. Например:

Таблица 3.

Примеры ошибок при переводе официальных документов

Текст в оригинале	Результат автоматического перевода	Правильный перевод
в соответствии с требованиями, установленными законодательными и иными нормативными правовыми актами	законнар чыгару һәм башка норматив хокукый актларда билгеләнгән таләпләр нигезендә	Законнарда һәм башка норматив хокукый актларда билгеләнгән таләпләр нигезендә
возможность посадки в транспортное средство и высадки из него перед входом в объект	транспорт чарасына утырту һәм объектка керү алдыннан аннан төшерү мөмкинлеге	объектка кергәндә транспорт чарасына утыру һәм төшү мөмкинлеге

2. Литературный стиль представлен повестью Саши Денисовой «Второй месяц весны – это апрель» в объеме около 200 предложений. Тестовый перевод данного текста машинным переводчиком ожидаемо выявил ошибки в обработке новых оборотов разговорного стиля, передаче современной тематики, молодежного сленга, метафор и др. выразительных средств. Например:

Таблица 4.

Примеры ошибок при переводе литературного текста

Текст в оригинале	Результат автоматического перевода	Правильный перевод
Пахнет мокрой улицей	Юеш урамнар исе килә	Юеш урам исе килә
Она протёрла мобильник, врубила случайную группу и, воткнув наушники, отправилась к центру города	Ул мобильникны сөртте, очраклы группаны алдады һәм, наушникларын кадап, шәһәр үзәгенә китте.	Ул кесә телефонын сөртте, очраклы группаны кушты һәм, наушникларын тыгып, шәһәр үзәгенә китте.

3. Публицистический стиль представлен текстами новостей информационного агентства «Татар-информ» в объеме 500 предложений [3]. Для включения в тестовый корпус были выбраны новости разнообразной тематики, потенциально сложной для русско-татарского машинного переводчика, а именно технологии, религия, медицина, спорт, астрономия. Перевод некоторых новостей дублируется в татарской версии сайта [3], но требовалась их проверка на эквивалентность. Осуществлялся также и перевод тех новостей, которые по

разным причинам были опубликованы только на русском языке. На этапе тестового перевода были выявлены следующие типичные ошибки машинного переводчика в текстах данного стиля: неправильная передача разговорных оборотов, отсутствие новых слов, терминов, современной и специализированной лексики. Например:

Таблица 5.

Примеры ошибок при переводе публицистических текстов

Текст в оригинале	Результат автоматического перевода	Правильный перевод
Верующие Татарстана пройдут с чудотворным образом Казанской иконы Божией Матери	Татарстан дин тотучылары могҗизалы рәвештә Казан Изге Ана иконасы үткәреләчәк	Татарстан дин тотучылары могҗизалы Казан Изге Ана иконасы белән үтәчәк
Наша Земля постоянно пересекает какие-то сгущения метеорных частиц – метеороидных частиц, маленьких частиц в виде различных камешков	Безнең Җир даими рәвештә метеор кисәкчекләре - метеороид кисәкчекләр, төрле ташлар рәвешендәге кечкенә кисәкчекләр куертып тора.	Безнең Җир даими рәвештә метеор кисәкчекләре - метеороид кисәкчекләре тупланмалары, төрле ташлар рәвешендәге кечкенә кисәкчекләр аркылы үтеп тора.

4. Научно-популярный стиль представлен текстами статей образовательного проекта «Гыйлем» [4] в объеме 800 предложений. Эти записи взяты с популярных русскоязычных сайтов с указанием источника и переводчика, поэтому дополнительный перевод не требуется. По итогам анализа в этих текстах были выявлены типичные ошибки в переводе современной тематики, специализированной лексики и

терминов, а также в построении предложений на татарском языке. Например:

Таблица 6.

Примеры ошибок при переводе научно-популярных текстов

Текст в оригинале	Результат автоматического перевода	Правильный перевод
Поскольку любая система стремится к полному равновесию, ее энергия, то есть тепло, постепенно рассеивается	Теләсә нинди система тулы тигезлеккә, аның энергиясенә, ягъни җылыга омтыла, акрынлап тарала	Теләсә нинди система тулысынча тигезләнешкә омтылганга күрә, аның энергиясе, ягъни җылылыгы, акрынлап тарала бара
Помимо обычной материи, из которой мы с вами состоим, есть еще и антиматерия	Гади материядән тыш, без сезнең белән яшибез, антиматерия дә бар әле	Безне тәшкил иткән гадәти материядән тыш антиматерия дә бар

5. Специализированная тематика в тестовом корпусев данный момент представлена текстами кулинарных рецептов блюд татарской кухни в объеме 300 предложений [1]. Для тестового корпуса была взята часть полного перевода данной книги. Типичные ошибки в этом источнике наблюдаются в передаче названия продуктов, особенного построения предложений.

Отобранные для тестового корпуса тексты отличаются по своей стилевой принадлежности, наблюдаются особенности по следующим характеристикам:

1. Длина предложения сравнительно больше в официально-деловых текстах; самые короткие неполные предложения наблюдаются в текстах кулинарных рецептов.

2. Лексический состав каждого стиля является по своему особенным, так как отражает специфичную сферу деятельности: наука, кулинария, медицина, религия, жаргон и т.д. Редкие и новые слова чаще встречаются в научных и публицистических текстах, которые посвящены узкоспециальным темам.

3. Наличие заимствований наблюдается во всех текстах, в зависимости от жанра и оригинала. Если в религиозных и кулинарных текстах преобладают арабско-персидские заимствования, то научные и публицистические тексты пестрят терминами из западноевропейских и латинских источников.

4. Наличие многозначных слов, переносных значений более характерно для литературного текста, тогда как в остальных стилях требуется точная передача значения в переводе.

Заключение

Тестовый параллельный корпус, сбалансированный по стилям и языковым явлениям, позволит проводить более точную оценку качества машинного переводчика «TatSoft» как по автоматическим метрикам, так и экспертным методом.

В свою очередь это дает возможность «точечного» пополнения обучающей выборки в условиях ограниченных двуязычных данных.

В дальнейшие планы входит пополнение текстового корпуса, исследование частотных и типичных ошибок на тестовой выборке, изучение значимых стилевых маркеров в татарском языке.

ЛИТЕРАТУРА

1. Адиатулин Ф.З. (2011) История народа и его кухни. Рецепты национальных блюд. - СПб.:«Издательство «ДИЛЯ», 2011 — 160 с.

2. Денисова Саша. Второй месяц весны – это апрель. – ЛитРес: Самиздат. – 550 с. [Электронный ресурс]. URL: https://aldebaran.ru/author/denisova_sasha_1/kniga_vtoroyi_mesyac_vesnyi_yeto_aprel/read/ (дата обращения: 01.11.2020).

3. ИА Татар-информ. https://www.tatar-inform.ru/.

4. «Гыйлем» проекты http://giylem.tatar/.

5. Официальный портал правовой информации Республики Татарстан https://pravo.tatarstan.ru/.

6. Русско-татарский переводчик TatSofthttps://translate.tatar/

7. Переводчик Яндекс. https://translate.yandex.ru/translator /Russian-Tatar.

8. Переводчик Google https://translate.google.com/?sl=ru&tl =tt&op =translate.

9. Переводчик Promt.Onehttps://www.translate.ru/.

10. Khusainov, A., Suleymanov, D., Gilmullin, R., Gatiatullin, A. (2018) Building the Tatar-Russian NMT System Based on Re-Translation of Multilingual Data. InProceedings of the 21st International Conference TSD 2018 (Brno, Czech Republic, September 11-14, 2018) (pp. 163-170).

11. Khusainov Aidar, Suleymanov Djavdet, Gilmullin Rinat (2020) The Influence of Different Methods on the Quality of the Russian-Tatar Neural Machine Translation. InRussian Conference on Artificial Intelligence, Springer, Cham(pp. 251-261).

СОДЕРЖАНИЕ

УДК 004.8+81’32
ББК 81.1

Восьмая Международная конференция по компьютерной обработке тюркских языков «TurkLang-2020». (Труды конференции) –Уфа: ИИЯЛ УФИЦ РАН, 2020. – 296 с.
ISBN 978-5-91608-199-2

НАУЧНОЕ ИЗДАНИЕ

Подписано в печать 29.12.2020. Бумага офсетная.
Формат: 60 х 84 1/8; Гарнитура Таймс.
Усл.-печ. л. 38,00.
Издат. л. 35,51. Тираж 500.

Изготовлено с готового оригинал-макета на ризографе.
450054, г. Уфа, пр. Октября,71.
Институт истории, языка и литературы УФИЦ РАН.
Тел.(347) 235-60-50.