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

Труды конференции

Том 2

КАЗАНЬ
2017

УДК 004.8+81’32
ББК 81.1

Организаторы:

Академия наук Республики Татарстан
Институт прикладной семиотики

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

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

Международная Тюркская академия

Российская ассоциация искусственного интеллекта

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

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

Пятая Международная конференция по компьютерной обработке тюркских языков «TurkLang 2017». – Труды конференции. В 2-х томах. Т 2. – Казань: Издательство Академии наук Республики Татарстан, 2017. – 327 с.
ISBN 978-5-9690-0407-8

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

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

УДК 004.8+81’32
ББК 81.1

ISBN 978-5-9690-0407-8

ПРЕДИСЛОВИЕ

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

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

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

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

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

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

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

ПРОГРАММНЫЙ КОМИТЕТ

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

ОРГАНИЗАЦИОННЫЙ КОМИТЕТ

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

СЕКЦИЯ 4

ТЕХНОЛОГИИ, МОДЕЛИ И СИСТЕМЫ ДЛЯ МОРФОЛОГИЧЕСКОГО АНАЛИЗА ТЮРКСКИХ ЯЗЫКОВ

УДК 81'33

MORPHOLOGICAL ANALYSIS SYSTEM OF THE TATAR LANGUAGE BASED ON THE TWO-LEVEL MORPHOLOGICAL MODEL

D. Suleymanov, R. Gilmullin, R. Gataullin
Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic, Kazan, Russia
dvdt.slt@gmail.com, rinatgilmullin@gmail.com, ramil.gata@gmail.com

This paper presents the description of the morphological analysis system for the Tatar Language based on a two-level morphology model. The morphological system is used for grammatical annotation of the Tatar national corpus. This paper shows the results of evaluation of completeness of the system using statistical information that was obtained from the corpus data and describes the ways to improve this system.

Keywords: Morphological analysis system; FST; alphabet; phonological rules; morphotactical rules; the Tatar language; Tatar National Corpus; Completeness of the system.

МОРФОЛОГИЧЕСКИЙ АНАЛИЗАТОР ТАТАРСКОГО ЯЗЫКА НА ОСНОВЕ ДВУХУРОВНЕВОЙ МОДЕЛИ МОРФОЛОГИИ

Д.Ш. Сулейманов, Р.А. Гильмуллин, Р.Р. Гатауллин
Институт прикладной семиотики АН РТ, Казань
dvdt.slt@gmail.com, rinatgilmullin@gmail.com, ramil.gata@gmail.com

В данной работе представлено описание морфологического анализатора татарского языка, основанного на двухуровневой модели морфологии. Морфологический анализ является одним из ключевых этапов в задачах автоматической обработки текстов, в особенности для татарского языка,

в силу его мощной и богатой по природе морфологии, проявляющейся на всех языковых уровнях. Модуль морфологического анализа, реализованный на основе данного подхода, используется в составе многих научно-исследовательских и прикладных программных приложений, таких как, УИС Россия (МГУ) – для поддержки поискового функционала в русско-татарской коллекции текстов; Интернет-сервис Яндекс.Перевод – для поддержки машинного перевода в русско-татарской языковой паре, а также национальный корпус татарского языка «Туган тел» – для грамматической аннотации контента. В работе также представлены результаты оценки полноты модуля морфологического анализа, полученных на статистической информации с использованием корпусных данных «Туган тел», а также предложения по улучшению мофроанализатора.

Ключевые слова: морфологический анализатор, двухуровневая модель, FST, алфавит, фонологические правила, морфотактические правила, татарский язык, национальный корпус татарского языка, полнота системы.

Введение

Как известно, морфология татарского языка это семантически сложное и, одновременно, конструктивно элегантное, регулярное, практически автоматное, языковое явление, и соответственно, морфологический анализ представляет собой важный этап при решении задач, связанных с обработкой естественного языка. Это также актуально и для других агглютинативных языков с богатой морфологией, таких как тюркские языки, к которым относятся татарский, турецкий, казахский, кыргызский, крымскотатарский, туркменский, уйгурский, узбекский, чувашский, якутский и другие.

В настоящее время существует множество работ, посвященных описанию моделей и систем морфологического анализа для тюркских языков [1-6]. Морфологическая модель для большинства этих систем, как и для татарского языка, описывается на основе фонологических и морфологических правил с использованием формализмов FST (Finite State Transducer). Данный выбор весьма логичен в силу особенности тюркских языков, обладающих высокой степенью регулярности и практически автоматной морфологией. Однако при описании морфологической модели татарского языка авторы столкнулись с большим количеством нарушений в морфологии татарского языка, основные причины которых – это боль-

шой процент (до 30-35%) не ассимилированных заимствований, в основном, с русского и арабского языков, а также неадекватность кириллического алфавита для отображения татарских фонем [7]. Кроме того, в имеющихся работах [8-10], посвященных описанию системы морфологического анализа для татарского языка, отсутствуют или слабо представлены данные об использовании системы и о соответствующих экспериментальных результатах. В данной работе мы также постарались восполнить этот пробел.

Разработка и практическое внедрение модели морфологии татарского языка нами и нашими коллегами осуществляется практически с середины 1990-х годов. За этот период были разработаны татарские морфологические анализаторы, основанные на трех типах моделей – генеративная модель, парадигматическая модель [11] и двухуровневая модель [8], отличающиеся по таким параметрам как точность, полнота и скорость обработки. На основе генеративной модели был разработан первый корректор татарских текстов [12], который проверял правильность написания татарских слов без ограничений на длину. Эта же модель включена в состав распознавателя тестов фирмы ABBYY, начиная с версии OCR FineReader 4.0. Парадигматическая модель используется в текстовом процессоре фирмы Microsoft, начиная с MS Word 2007.

В данной работе представлено описание текущей версии морфологического анализатора татарского языка, основанной на двухуровневой модели морфологии и реализованной с помощью технологии трансдьюсеров конечных состояний (FST) [13]. Практически за 20 лет с начала разработки татарского морфологического анализатора на основе двухуровневой модели в рамках совместного проекта с турецкими коллегами по технологии PC-KIMMO [14], продукт претерпел ряд существенных изменений, как в содержательном плане, так и в плане уточнения описания модели в соответствии с требованиями к современным системам морфологического анализа, использования новых технологических возможностей для реализации модели. Модуль морфологического анализа, реализованный на основе данного подхода, используется в составе многих научно-исследовательских и прикладных программных приложений, таких как, УИС Россия (МГУ) для поддержки поискового функционала в русско-татарской кол-

лекции текстов; в Интернет-сервисе Яндекс.Перевод для поддержки машинного перевода в русско-татарской языковой паре, а также в национальном корпусе татарского языка «Туган тел» для грамматической аннотации контента. Для оценки полноты системы морфологического анализа на основе двухуровневой модели собрана статистическая информация, полученная в процессе морфологического аннотирования корпусных данных «Туган тел» [15].

На основе этих данных в работе представлены наши последние результаты по повышению полноты морфологической модели.

1. О татарском языке

Татарский язык относится к тюркской группе, которая образует подсемейство алтайских языков. На татарском языке разговаривают в западно-центральной России (в Поволжье) и южных частях Сибири. Количество татар в России в 2010 году составило 5,31 миллиона человек [16]. Выделятся различные диалекты татар: западные, казанские (средние) и восточные. Более 2 миллионов татар проживают в диаспорах за пределами страны, главным образом, в бывших странах СНГ, а также Финляндии и Турции. В 2013 году существующие языковые классификации [17, 18] описали татарский язык как язык с ограниченными ресурсами. Однако результаты последних лет, связанных с разработкой систем машинного перевода [19, 20], систем речевого анализа и синтеза [21], а также электронного корпуса татарского языка [15] способны изменить эту ситуацию.

2. Описание морфологического анализатора татарского языка на основе двухуровневой модели морфологии

Одна из первых версий морфологического анализатора татарского языка была создана на основе двухуровневой модели морфологии с использованием инструмента PC-KIMMO [8, 14]. Информационная база PC-KIMMO, с точки зрения разработчика морфологического анализатора, состоит из двух файлов, созданных пользователем. Первый файл – файл правил (Rules), который описывает алфавит и фонологические правила. Второй файл – лексикон, содержащий словарь лексических единиц (корневых

и аффиксальных морфем) и их толкования, а также описание морфотактических правил. Лексикон состоит из подлексиконов (sublexicons), разделенных по селективным признакам и парадигматическим классам. Структура подлексиконов образует связанный граф, в вершине которого стоит корневой (root) лексикон, начинающий анализ входного слова [22]. Все правила второго компонента морфологии записываются на языке регулярных выражений. Технология анализа построена на разновидности конечных автоматов FST (finite-state transducer) (ТКС – трансдьюсер конечных состояний). ТКС называется автомат, в котором каждый переход между состояниями в сети (network) имеет выходную помету в дополнение к входной [23]. Исходный морфологический лексикон компилируется в lexicon transducer, а компонент правил – в two-level rule transducer. Результирующий лексический конечный автомат, т.е. полное морфологическое представление языка – lexical transducer, получается композицией lexicon transducer и rule transducer [22]. Символьный алфавит конечного автомата называется Sigma [23]. Sigma лексического ТКС состоит из алфавита анализируемого естественного языка и специальных грамматических помет (tags), выражающих значение селективных признаков и граммем (например, Noun – существительной, +Poss – притяжательность, P1 – 1-ое лицо, +Sg – число и т.д.).

Корневой лексикон осуществляет вызов подлексиконов. Выражения в лексиконах представляют собой пару форм: лексическая (lexical) и поверхностная (surface) формы, разделенные двоеточием. Строящий ТКС компилятор интерпретирует такую пару как регулярное отношение. Решетка ‘#’ маркирует конечное состояние. Уникальность пути переходов в сети конечного автомата дает однозначность морфологической интерпретации. Приводящие в конечное состояние варианты пути в сети ТКС задают множественность интерпретаций для поверхностной формы, что соответствует морфологической многозначности.

При двухуровневом подходе фонология определяется как связь между лексическим уровнем глубинного представления слов и их реализации на поверхностном уровне, в силу чего теоретическая модель фонологии PC-KIMMO называется двухуровневой фонологией. PC-KIMMO включает две функциональные компоненты – генератор и распознаватель.

Генератор на входе получает лексическую форму, применяет

правила фонологии и возвращает соответствующую поверхностную форму. При этом лексикон не используется.

Распознаватель получает на входе поверхностную форму, применяет правила фонологии, обращается к лексикону и возвращает соответствующие лексические формы с их комментариями (толкованиями). На рисунке 1 показана структурно-функциональная схема двухуровневого морфологического анализатора.

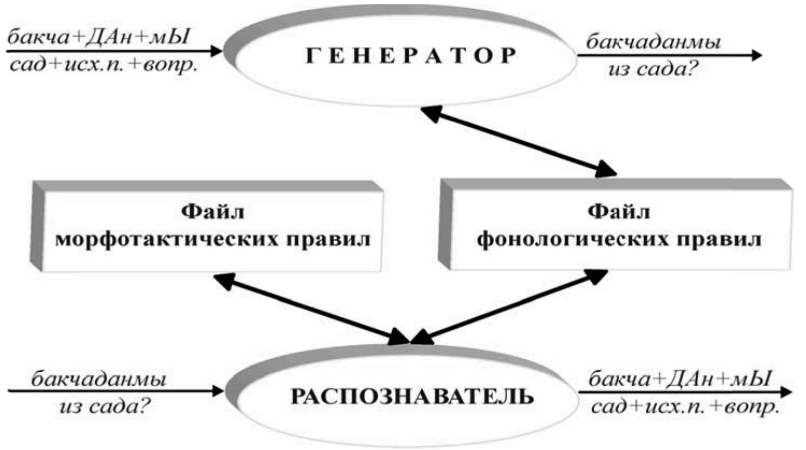


Рис. 1. Структурно-функциональная схема двухуровневого анализатора

Генератор, используя файл фонологических правил, лексическую запись *бакча+ДАн+мЫ* переводит в поверхностную форму – *бакчаданмы 'из сада?*'. Распознаватель, используя оба файла – файл фонологических и файл морфотактических правил, словоформу (поверхностную форму) *бакчаданмы* раскладывает по составляющим и соответствующим им содержательным описаниям: *Сущ.(бакча)+[исх.пад.(ДА)] +[вопр.(мЫ)]*.

Данный инструментарий имел заметные технические ограничения, поэтому в последующем была осуществлена новая реализация с использованием инструментария и технологии HFST (Helsinki Finite-State Transducer Technology), что позволило существенно повысить технические параметры анализатора. Основные технические параметры систем представлены в таблице 1.

Таблица 1

Технические параметры	**PC-KIMMO**	**HFST**
Морфотактические правила	49	49
Фонологические правила	42	55
Словарь	26500	28000
Скорость обработки	~500-1000 (слов/сек.)	~5000-20000 (слов/сек.)
Полнота	80-85%	93-95%
Алфавит	Латиница	Кириллица
Язык программирования	C/C++, Delphi	C/C++, Python
Наличие API	Нет	Да
Интерфейс	Десктопное приложение	Веб-приложение

Как показано в таблице 1, количество правил в текущей версии морфологического модуля увеличелось на 13 правил. Это прежде всего связано с реализацией правил орфографии в кириллице, в которой пристутсвуют символы “е”, “ю”, “я”, представляющие разные звуки в языке в зависимости от гармонии гласных в слове: например, для “е”: “ка**е**н” – [ка**ый**н] (берёза), “киб**е**т” – [кибэт] (магазин), “ти**е**н” – [ти**ий**ен] (белка, копейка); для “ю”: “а**ю**” – [а**йу**] (медведь), “**ю**кә” – [**йү**кә] (липа); and for “я”: “**я**л” – [**йа**л] (отдых), “яшел” – [**йә**шел] (зелёный). Для описания этих символов, в файле правил системы морфологического анализа были созданы дополнительные правила. Кроме этого, множество нарушений в морфологии возникают из-за не ассимилированных заимствований из русского и арабского языков, например, слова, в которых нарушена гармония гласных: “бәл**а”** (беда) – вместо “бәлә”, “кабин**е**т+**ы**” (его/её/их кабинет) – вместо “кабин**е**т+**е**”, “әдәби**я**т+**ы**” (его/её/их литература) – вместо “әдәби**я**т+**е**”. Для учета этих особенностей языка также были созданы дополнительные исключительные правила и словарь исключительных слов.

В настоящее время татарский язык официально использует кириллицу. Кириллический алфавит был принят в 1939 году и состоит из 39 символов: 12 гласных и 28 согласных. Правило гар-

монии гласных в татарском языке гласит, что в пределах одной словоформы гласные буквы обладают определенными характеристиками, а именно, они относятся к гласным переднего ряда или заднего ряда произношения. Это называется законом сингармонизма и является основной лингвистической особенностью практически всех тюркских языков и используется при создании файла правил двухуровневого морфоанализатора.

В своей истории татарский язык в разные периоды был представлен также арабской и латинской графикой. При этом, как это показано в [7], наиболее эффективным с точки зрения описания и обработки татарского языка в информационных технологиях является использование латинской графики.

Все особенности в языке, как уже было отмечено, могут быть описаны в HFST помощью двух компонент – файла фонологических правил и файла морфотактических правил [24].

2.1. Файл фонологических правил морфологического анализатора татарского языка

С точки зрения нотации HFST, Файл фонологических правил состоит из служебных слов: Alphabet (Алфавит), Set (Множество) и Rule (Правило).

Alphabet *а б в г д е ё ж з и й к л м н о п р с т у ф х ц ч ш щ ъ ы ь э ю я ə ө ү җ ң һ Л:л Л:н Д:д Д:т Д:н А:я А:а А:ə а:ы А:ы ə:0 А:0 Ы:ы Ы:е Ы:0 ы:е ы:0 С:с С:0 Г:г Г:к Г:н Г:0 Й:й Й:а Й:ə Й:и Й:я У:у У:ү У:в У:ю ь:0 й:0 к:г п:б 0:ы е:0 е:и а:я ə:я у:ю ү:ю*

Sets – объявление множеств.

CS = *һ х б ц җ д ф г ж к л м н ң п р с ч ш щ т й з ъ ь в Л Д С Г Й* – множество всех согласных букв;

NASAL = *м ң н* – nasalization;

VOWEL = *э а е ы и о ө ү у ə я ю Ы А* – множество всех гласных букв;

BACKV = *а ы о у* – множество задних («твердых») гласных;

FRONTV = *э ə и ө ү е* – множество передних («мягких») гласных;

ZVONKCS = *р н ң ж л м й з* – множество звонких согласных (особый случай);

ZVOPARE = *б д г ж л р й з м н ң* – множество всех звонких согласных;

GLUPARE = *п т с к ф ш һ ч щ* – множество всех глухих согласных;

GLUHCS = *х б ц ч д ф г һ к п с ш т* – множество глухих и звонких согласных (особый случай);

WUY = *у ү ю*

Алфавит содержит 39 символов из современного кириллического татарского алфавита, а также 43 пары соответствий, реализуемые в правилах для описания морфонологических проявлений в языке.

Некоторые заглавные символы, такие как *А, Ы, С, Л, Д, Г, Й, У*, определены на промежуточном уровне и невидимы для пользователя. Эти представления используются для замещения поверхностных символов. Например, символ *А* вместо *а, я, ә* или *0*, а символ *С* вместо *с* и *0*. Здесь *0* – это пустой символ. Символы используется для реализации множества аффиксов и их проявлений в виде соответствующих алломорфов, зависящих от контекста. Так, например, аффиксы с символом *А* и их алломорфами приведены в таблице 2. При этом некоторые аффиксы могут содержать более чем один заглавный символ. Для каждого такого символа, называемого лексическим символом, реализуется отдельное правило, которое описывает его соответствующее проявление, называемое поверхностным символом.

Таблица 2

Лексический символ	**Грамматический тип**	**Поверхностный символ**
-КАлА	Модальность	*-кала, -гала, -кәлә, -гәлә*
-ЫргА	Инфинитив	*-ырга, -ергә, -рга, -ргә*
-мАк	Инфинитив	*-мак, -мәк*
-мА	Негатив	*-ма, -мә*
-сА	Условие	*-са, -сә*
-КАн	Категория времени	*-кан, -ган, -кән, -гән*
-АчАк	Категория времени	*-ачак, -әчәк, -ячак, -ячәк*
-АУ		*-ау, -әү*
-КА	Директив	*-га, -ка, -гә, -кә, -а, -ә*
-ДА	Локатив	*-да, -дә, -та, -тә, -нда, -ндә*

Лексический символ	**Грамматический тип**	**Поверхностный символ**
-ДАн	Аблатив	*-дан, -дән, -тан, -тән, -ннан, -ннән*
-ДАгЫ	Посессив	*-дагы, -дәге, -тагы, -тәге, -ндагы, -ндәге*
-чА	Компаратив	*-ча, -чә*

Для полного описания морфологической модели татарского языка было реализовано 55 правил. Детальное описание этих правил приведены в работах [8, 25]. 12 правил из 55 реализуют исключительные ситуации, вызванные несовершенством татарской орфографии и из-за неассимилированных заимствований, нарушающих закон сингармонизма. Все эти правила реализованы с помощью инструментария HFST. Рассмотрим некоторые из них. Например, для лексического символа ***A***, представленного в аффиксах таблицы 2 и проявляющегося в виде поверхностного ***а***, правило в нотации HFST [24] могла бы иметь следующий вид:

“rule #1”

A:a => :BACKV (CS) (:0*) (CS)_;

Правило утверждает, что символ ***A*** реализуется как ***а***, если ему могут предшествовать согласные буквы, либо пустой символ, встречающийся ноль и более раз, перед этим снова может встретиться согласная буква, которой обязательно должна предшествовать любая буква из множества задних гласных.

В принципе этого было бы достаточно с точки зрения соблюдения закона сингармонизма при условии отсутствия следующих факторов их нарушения:

– использование несовершенного кириллического алфавита с символами *е, ю, я*, которые не могут относится ни к заднему, ни к переднему ряду гласных.

– множество заимствований в основном из русского, арабского языков, которые из-за неэффективной академической грамматики современного татарского языка остаются в языке в нессимилированной форме. Для таких слов был создан исключительный словарь со специальной меткой, сигнализирующей для правила появление такого рода контекста и о невозможности применения регулярных правил.

Для реализации модели морфологии с учетом всех этих факторов нарушений в правило были введены соответствующие изменения в виде различных исключительных контекстов. Так, например, «rule #1” пришлось преобразовать в следующей громоздкой форме:

```
A:a =>   :BACKV (e) (CS) (%>:0) (CS) ( :0*) (CS)_;
[#ю|BACKV й:0 %>:0 У:ю] %>:0 CS_;
:BACKV (й:0) %>:0  :я (%>:0) CS_;
 Ы:ы %>:0 0:н CS_;
 BACKV й:0 %>:0 С:0 Ы:е %>:0 Г:н_ ;
  :BACKV (CS) (%>:0) (Ы:0) (CS) %>:0 (CS)_;
:BACKV %>:0 Ы:0 (CS) (%>:0) (CS*) _; !абайла+ ЫрГА
      :BACKV (й:0) (%>:0)  :е (CS) (%>:0) (CS) ( :0*) (CS)_;
      [а:я|у:ю|ы:е] (е) (CS) (%>:0) (CS) ( :0*) (CS)_;
      [а:я| :BACKV|у:ю|ы:е] %>:0 Ы:0 (CS) %>:0 Г:0_;!дөньяма
      [а:я|у:ю|ы:е] %>:0 Ы:0 (CS)* (%>:0) (CS)*_;!дөньямда
      BACKV й:0 %>:0 Ы:е (CS*) (%>:0) (CS) _;!тоелма
```

2.2. Файл морфотактических правил морфологического анализатора татарского языка

Файл морфотактических правил разработан на основе морфотактических схем для именных (рисунок 2) и глагольных групп (рисунок 3) и определяет взаимосвязи между основой и аффиксальными группами.

Лексикон корневых лексем построен на основе современного татарского языка и состоит из ряда лексиконов, заполненных согласно соответствующих требований инструментария HFST. Подлексиконы содержат строки лексических входов, состоящих из следующих трех частей: **первая часть** – лексический атом (татарское корневое слово); **вторая часть** – класс присоединения (или продолжения); т. е. то, что может присоединяться, следовать как продолжение непосредственно за этим атомом – подлексикон, который может иметь другие лексические единицы. Классы присоединений могут следовать за множеством других морфемных единиц. Лексикон **ALTERNATION** в PC-KIMMO – это список названий подлексиконов, порядок которых определяет – какой класс за каким может следовать, притом, возможно только одно его определение, то есть это ограничение, свойственное подлек-

сикону; **третья часть** – его трактовка (описание грамматических признаков). Как правило, здесь записываются любые морфологические, грамматические, лексические, или семантические свойства лексической единицы. При обработке слова распознавателем трактовка каждой избранной морфемы добавляется в строку результата.

(1) Имена существительные (Nouns). Лексикон включает около 15984 корневых имен существительных.

(2) Глаголы (Verbs). Лексикон содержит около 7281 глагольных корней.

(3) Прилагательные (Adjectives). Как известно, татарский язык является агглютинативным регулярным языком, подчиняющимся строгим правилам. Вместе с тем, как и в любом естественном языке, имеются исключения, чаще всего, также подчиняющиеся определенным закономерностям. Так, прилагательные превосходной степени имеют префиксы, записываемые через дефис '-'. Например: корневое слово 'красный' в превосходной степени записывается как *кып-кызыл* (*'очень красный'*). Лексикон **Прилагательные** содержит 3646 базовых корней и дополнительно включает лексикон, состоящий из 136 прилагательных превосходной степени с префиксами. Определены также следующие Лексиконы, составляющие небольшую долю в общем словаре, включающем 28411 корневых слов, имеющие особые морфотактические правила, присущие выделенным группам слов: **(4) Наречия (Adverbs). (5) Местоимения (Pronouns). (6) Числительные (Numerals). (7) Послелоги (Postpositions). (8) Союзы (Conjunctions). (9) Междометия (Exclamations).**

Параметр **ALTERNATION** включает 8 входов для словоформ (т. е. в данном описании определено, что имеет место 8 разных возможностей для начала татарского слова): **VERB (Глагол)** – подлексикон для глаголов; **NOUN (Имя существительное)** – подлексикон для существительных; **ADJECTIVE (прилагательное)** – подлексикон для прилагательных; **ADJECTIVE2 (прилагательное2)** – подлексикон для прилагательных; **NUMERAL (число)** – подлексикон для чисел; **PRONOUN (местоимение)** – подлексикон для местоимений, послелогов; **ADVERB (наречие)** – подлексикон для наречий; **SPECIAL (специальное)** – подлексикон для союзов, междометий.

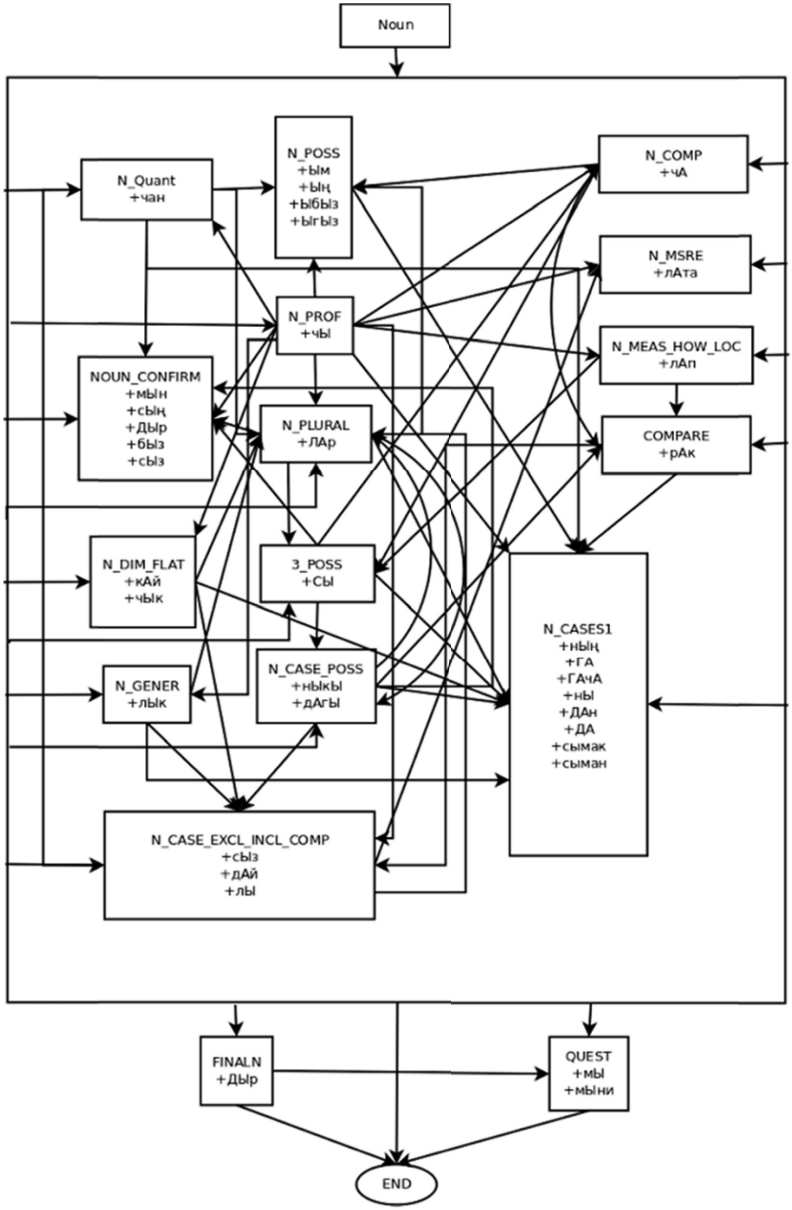


Рис. 2. Морфотактическая схема переходов для именных групп

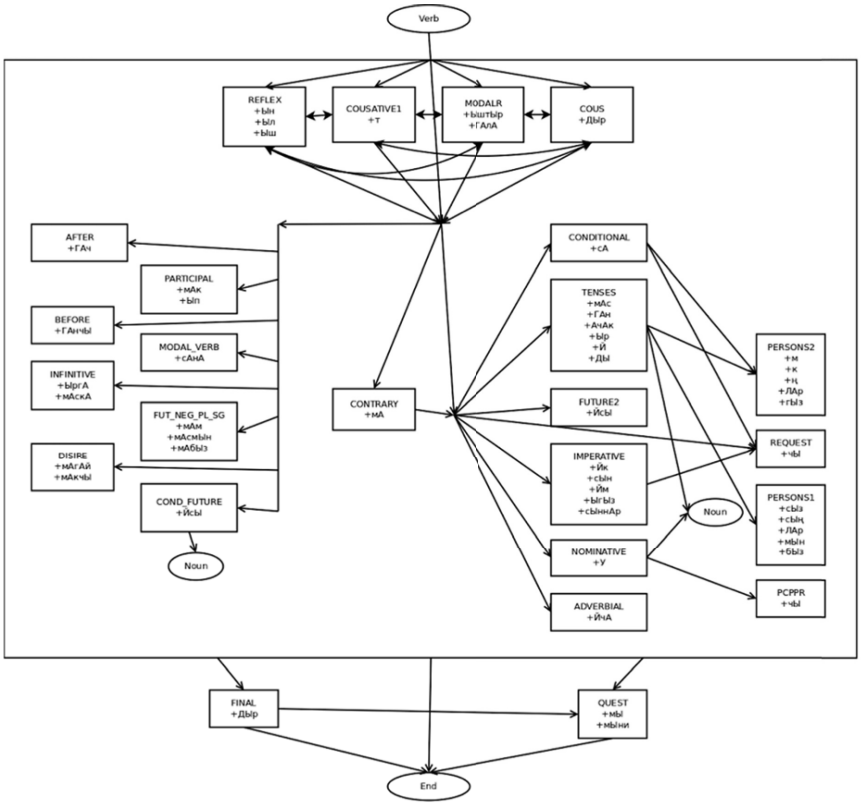


Рис. 3. Морфотактическая схема переходов
для глагольных групп

3. Тестирование и эксперименты

Система морфологического анализа была тестирована на данных национального корпуса татарского языка «Туган тел» [15], который разработан специалистами Института прикладной семиотики Академии наук и Казанского федерального университета. Объем корпуса в настоящее время составляет 63,508,127 лексических единиц. Таблица 3 содержит результаты графематического анализа, выполненного на тех же корпусных данных «Туган тел».

Таблица 3

Лексическая единица	**Теги**	**Уникальные слова**	**%**	**Все использованные слова**	**%**
Слова		375298	40,55	38629659	60,8
Знаки пунктуации	Type2	138	0,01	11134395	17,5
Конец предложения	Type1	4	0,00	5305378	8,3
Не распознано	NR	482338	52,12	4113428	6,4
Число, Латин, Буква, Знаки, др.	Type3	18720	2,02	2872947	4,5
Ошибка	Error	49022	5,30	1452320	2,3
Всего		925520	100	63508127	100

На момент исследования, объем национального корпуса татарского языка «Туган тел», составлял порядка 63 миллионов лексических единиц, включая слова, знаки пунктуации, числа, слова на латинской графике и др. 67,30% объема всего корпуса представлен словами. Морфологический анализатор распознал 90,37% и не распознал 9,62% лексических единиц. Некоторые лексические единицы размечены как “Error”, означающий, что они содержат кириллические, латинские буквы, числа и некоторые другие символы одновременно, поэтому не могли быть распознаны, как татарское слово. Количество таких слов составило порядка 2,29% от всего объема корпуса.

В таблице 4 дано распределение слов по частям речи.

Таблица 4

Части речи	**POS Tag**	**Уникальные слова**	**%**	**Все слова**	**%**
Сущ.	N	137530	55,67	11331955	40,11
Глагол	V	83916	33,97	6007993	21,27
Местоимение	PN	1398	0,57	2411290	8,54
Прилагательное	Adj	12234	4,95	2187637	7,74
Частица	PART	73	0,03	1813949	6,42
Имя собственное	PROP	9893	4,00	1231996	4,36

Части речи	POS Tag	Уникальные слова	%	Все слова	%
Послелог	POST	61	0,02	890237	3,15
Наречие	Adv	569	0,23	787836	2,79
Союз	CNJ	30	0,01	785302	2,78
Число	Num	1162	0,47	426047	1,51
Модальное слово	MOD	66	0,03	311207	1,10
Междометие	INTRJ	116	0,05	63781	0,23
Подражательное слово	IMIT	15	0,01	2141	0,01

Как и ожидалось, большинство (77,65%) слов – это имена существительные (40,11%), остальные – глаголы (21,27%), местоимения (8,54%) и прилагательные (7,74%). Вместе с тем, в вычислениях не приняты во внимание неоднозначные слова, которые составляют 26,86% от всех распознанных слов.

Распределение неоднозначных слов и число возможных вариантов их разбора дано в таблице 5.

Таблица 5

Количество вариантов	Уникальные слова	%	Все слова	%
1 (однозначно)	247063	65,83	28251371	73,13
2	94699	25,23	7640618	19,78
3	14074	3,75	1690613	4,38
4	15292	4,07	889574	2,30
5	603	0,16	52147	0,13
6	3078	0,82	80256	0,21
7	21	0,01	7842	0,02
8	274	0,07	10182	0,03
9	123	0,03	272	0,00
10	42	0,01	6726	0,02
11	1	0,00	1	0,00
12	27	0,01	54	0,00
15	1	0,00	3	0,00

Количество неоднозначных слов составило 26,86% от количества всех распознанных слов. Большинство из них (19,78%) имеют только два варианта разбора.

Таблица 6 иллюстрирует распределение слов в соответствии с длиной цепочки аффиксов, а Таблица 7 описывает их распределение согласно количеству нераспознанных слов, что интересно для оценки системы.

Таблица 6

Число аффиксов	**Уникальные слова**	**%**	**Все слова**	**%**
0	16162	6,5417	13555659	47,98
1	73860	29,8952	9845456	34,85
2	97728	39,5559	3913178	13,85
3	47928	19,3991	862454	3,05
4	9883	4,0002	70470	0,25
5	1438	0,5820	4057	0,01
6	63	0,0255	96	0,00
7	1	0,0004	1	0,00

Таблица 7

Причина		**Уникальные слова**	**%**	**Все слова**	**%**
Не словарное слово	Имя собственное	246	24,6	324544	24,2
	Другие	170	17,0	234276	17,4
Неполнота правил	Русское или европейское происхождение слова	190	19,0	257597	19,2
	Татарское происхождение слова	102	10,2	173778	12,9
	Арабское, персидское происхождение слова	25	2,5	41683	3,1
Ошибочное или не татарское слово		267	26,7	296097	22,0
Всего		1000	100	1342108	100

Как показал анализ 1000 наиболее частотных нераспознанных слов, большинство этих слов (41,60% анализированных слов) не было включено в словарь анализатора, из-за чего они и не были распознаны, хотя и имеют регулярные формы. Большинство из них (24,60% анализированных слов) являются именами собственными (например, Татнефть, Гатауллин; остальные – редко используемые слова и аббревиатура (например, *агросәнәгать*, АНТ (Академия наук Татарстана)) или заимствованные слова (например, брифинг, телесериал). Здесь решением является включение этих слов в словарь, используемый анализатором.

Другая причина, по которой не распознается часть слов корпуса, это неполнота правил анализатора. Имеется ряд исключительных ситуаций (21,50% анализированных слов), связанных с неассимилированными заимствованными словами (например, кит (рыба-кит), кабинет, компьютер). Также в некоторых случаях (10,20% анализированных слов) это связано с несоответствием кириллического алфавита при написании ряда татарских слов. Например, большинство исключительных ситуаций, как уже было показано выше, связано наличием сложных букв, типа ***е*** (*йе* или *йы*), ***ю*** (*йү* или *йу*), ***я*** (*йә* или *йа*) в татарском алфавите, из-за которых возникает ситуация двузначности при выборе твердых или мягких вариантов последующих аффиксов (например, *баю – байуы*, *бию – биүе*). Возможное решение в этом случае – исправить некоторые правила и, впоследствии, исправить словарь в соответствии с ними. Другие слова (26,70% анализированных слов) являются ошибками или не татарскими словами. В этих случаях, очевидно, не требуется внесения изменений в морфологический анализатор.

Заключение

В данной работе представлена система двухуровневого морфологического анализа татарского языка. Наряду со структурой и возможностями анализатора, дана также оценка полноты системы морфологического анализа на основе Национального корпуса татарского языка. Лучший показатель полноты системы был достигнут при добавлении правил для исключительных ситуаций с нарушениями морфологии; он вырос с 85% до 95%. Результаты

тестов делают хороший материал для дальнейшего улучшения морфологического анализатора.

В настоящее время, наряду с грамматическим аннотированием национального корпуса татарского языка, а также и других специализированных корпусов татарского языка, таких как, например, социо-политический корпус татарского языка [26], система морфологического анализа активно используется в системе русско-татарского машинного перевода Яндекс [27]. Этот сервис доступен по адресу: http://tatmorphan.pythonanywhere.com/.

ЛИТЕРАТУРА

1. Kemal Oflazer. Two-level Description of Turkish Morphology. Literary and Linguistic Computing, – Vol. 9, No 2, – 1994.

2. Altintas K., Cicekli I. A morphological analyzer for Crimean Tatar // Proceedings of the 10th Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN'2001). – 2001. – P. 180–189.

3. Çağrı Çöltekin (2014) A Set of Open Source Tools for Turkish Natural Language Processing In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) Ed. by N. Calzolari et al. 1079–1086.

4. G. Kessikbayeva and I. Cicekli, "Rule based morphological analyzer of Kazakh language," in Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM. Baltimore, Maryland: ACL, June 2014, pp. 46–54.

5. C. Tantug, E. Adali, K. Oflazer. Computer Analysis of the Turkmen Language Morphology, 5th International Conference on NLP(FinTAL 2006), Turku, Finland, 186–193, 2006.

6. M. Orhun, C. Tantug, E. Adali. Rule Based Analysis of the Uyghur Nouns, International Journal of Asian Lang. Proc.,19(1), 33-44, 2009.

7. Сулейманов Д.Ш. Об адекватном алфавите для татарского языка: латиница или кириллица // В сб. Трудов Первой международной конференции «Компьютерная обработка тюркских языков». – Астана: ЕНУ им. Л.Н. Гумилева, 2013. – С. 75–84.

8. Сулейманов Д.Ш., Гильмуллин А.А., Гильмуллин Р.А. Двухуровневое описание морфологии татарского языка Тез. Межд. научн. конф., посвященной 200-летию университета: "Языковая семантика и образ мира" (Казань, 7–10 октября 1997г.). Книга 2. – Казань: Изд-во Казан. ун-та, 1997. – С. 65–67.

9. Gökgöz, Ercan, et al. Two-Level Qazan Tatar Morphology. Proceedings of the 1-st International Conference on Foreign Language Teaching and Applied Linguistics, May 5-7 2011 Sarajevo, P. 428–432.

10. Davliyeva, A. R. (2011). An investigation of Kazan Tatar morphology (Doctoral dissertation, San Diego State University).

11. Сулейманов Д.Ш., Гатиатуллин А.Р. Модель татарской аффиксальной морфемы и ее реализация // Серия: Интеллект. Язык. Компьютер. – Вып.4. -Казань: Изд-во "Фэн". – 1996. – 113–127.

12. Сулейманов Д.Ш., Шафигуллин Р.Н. Морфологический корректор татарских текстов ТАТКОР // Татарский язык и новые информационные технологии. Серия: Интеллект. Язык. Компьютер. – Вып.2. – Казань: Изд-во Казан. ун-та. – 1995. – С.86–89.

13. HFST. https://kitwiki.csc.fi/twiki/bin/view/KitWiki/HfstHome

14. Evan L. Antworth. *PC-KIMMO*: A *Two-level* Processor for Morphological Analysis// Volume 16 of Occasional publications in academic computing / Summer Institute of Linguistics, 1990.

15. Tatar Nacional Corpus. http://tugantel.tatar/?lang=en

16. Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2016. Ethnologue: Languages of the World, Nineteenth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

17. V. Berment, "Me´thodes pour informatiser des langues et des groups de langues peu dote´es", Ph.D. Thesis, J. Fourier University, Grenoble I, 2004.

18. S. Krauwer, "The basic language resource kit (BLARK) as the first milestone for the language resources roadmap", In Proc. of International Workshop Speech and Computer SPEECOM, Moscow, Russia, 2003, P. 8–15.

19. Yandex Translate. Online version: https://translate.yandex.com/translator/Russian-Tatar.

20. D. Suleymnov, A. Gatiatullin, R. Gilmullin, "Lexicograficheskaya baza dannykh dlya system mashinnogo perevoda blizkorodstvennykh yazykov", In Proc. of Third International Conference «Informatizatciya obschestva», Astana, Kazakhstan, 2012, P. 585–587.

21. A. F. Khusainov, D. Sh. Suleymanov, "Language Identification System for the Tatar Language", Speech and Computer, Lecture Notes in Computer Science. 2013. Volume 8113. P. 203–210.

22. Karttunen L. Constructing Lexical Transducers. //15th International Conference on Computational Linguistics. Coling 94, I, pages 406–411. August 5–9, 1994. Kyoto, Japan.

23. Xerox, MLTT-95/Application of Finite-State Networks. Online version: www.xrce.xerox.com/research.

24. Beesley R. K. and Karttunen L. 2003. Finite State Morphology. CSLI Publications, Stanford, CA, USA.

25. R. Gilmullin, “Matematicheskoye modelirovaniye v mnogoyazykovykh sistemakh obrabotki dannykh na osnove avtomatov konechnykh sostoyaniy”, Ph.D. Thesis, Kazan, 2009, P. 48–94.

26. Socio-Political Corpus of the Tatar language. http://tugantel.tatar/corpus/op/

27. A.Sokolov, A.Egorov, S.Gubanov, D.Khrystich, M.Shmatova, I.Galinskaya, A.Baytin. Eksperimental’naya versiya tatarsko-russkogo statisticheskogo mashinnogo perevoda. Proceedings of the International Conference “Turkic Language Processing: Turklang-2015”. – Kazan: Academy of Science of the Republic of Tatarstan Press, 2015. P. 67–76.

УДК 81'33

EXPERIENCE OF COMPUTER-ORIENTED DESCRIPTION OF THE TUVAN MORPHOPHONOLOGY WITHIN THE FRAMEWORK OF THE SYSTEM OF AUTOMATIC MORPHOLOGICAL ANALYSIS

A. Dybo, A. Sheymovich
Institute of linguistics of RAS, Moscow, Russia
adybo@mail.ru, asheimovich@yandex.ru

This paper describes brief results of the most recent stage of the work on the development of an automatic morphological analyzer for Tuvan language, which was started in a frame of the Presidium of RAS project “Corpus Linguistics”, and resumed in line with the DHPS RAS program “The working out, creation and developing of electronic parallel linguistic corpora of minority Turkic languages and dialects of Russia”. The primary purpose of the present paper is the ordering of segment rules of the Tuvan wordform generation as well as their interpretation in the parser design.

Keywords: wordform generation, segment rules of allomorph choise, morphological rules, morphophonological rules, phonological rules, graphic rules, automatic morphological analyser.

ОПЫТ КОМПЬЮТЕРНО-ОРИЕНТИРОВАННОГО ОПИСАНИЯ ТУВИНСКОЙ МОРФОНОЛОГИИ В РАМКАХ СИСТЕМЫ АВТОМАТИЧЕСКОГО МОРФОЛОГИЧЕСКОГО АНАЛИЗА[1]

А.В. Дыбо, А.В. Шеймович
Институт языкознания РАН, Москва
adybo@mail.ru, asheimovich@yandex.ru

В работе представлены краткие результаты очередного этапа работы над корпусными технологиями, начатыми в рамках программы Президиума РАН «Корпусная лингвистика» и продолжающимися по проекту программы ОИФН РАН «Разработка, создание и развитие электронных параллельных лингвистических корпусов миноритарных тюркских языков и диалектов РФ». В настоящей статье основное внимание уделено классификации сегментных правил синтеза словоформы тувинского языка и тому, каким образом эти правила интерпретируются в теле разрабатываемого парсера.

[1] Работа выполнялась по гранту РГНФ № 15-04-00370 «Разработка анкет для сбора материалов к интегральному описанию миноритарных тюркских языков и диалектов России».

Ключевые слова: синтез словоформы, сегментные правила выбора алломорфов, морфологические правила, морфонологические правила, фонологические правила, графические правила, автоматический морфологический анализ.

Введение. О наших проектах

Уже несколько лет развиваются исследования по программам РГНФ (РФФИ), Президиума РАН и ОИФН РАН «Корпусная лингвистика», направление «Корпуса языков народов России». В рамках проекта по корпусам миноритарных тюркских языков создан и пополняется новыми материалами корпус хакасского языка, насчитывающий к настоящему времени около 500 тыс. словоформ. В 2014 г. на научно-практическом семинаре UniTurk "Унификация систем грамматической разметки в корпусах тюркских языков" в рамках конференции TEL2014 в Казани была представлена рабочая модель автоматического парсера для корпуса хакасского языка, опирающаяся на компьютерную модель хакасской словоформы (Дыбо, Шеймович 2014); также продолжается грамматическое, словообразовательное и семантическое аннотирование хакасского корпуса. Наработки в области автоматической разметки хакаского корпуса мы планируем использовать при создании парсеров для тувинского (в настоящее время) и шорского (в ближайшем будущем) языков.

Примерно с того же времени (2012 г.) коллектив научного центра при Тувинском госуниверситете в Кызыле работал над созданием тувинского корпуса и предпринимал шаги по направлению к его лингвистическому аннотированию. Разработки А.Б.Хертек и Б.Ч. Ооржак в области создания модели тувинской словоформы были представлены на конференции TurkLang-2015 (Oorzhak, Khertek 2015). С использованием этих разработок на основе инструментария Helsinki Finite-State Toolkit (HFST) американскими тюркологами из университета Блумингтона был построен автоматический анализатор для тувинского языка (Washington, Bayyr-ool 2016) Наша модель в целом близка предложенной американскими коллегами, но отличается в нескольких существенных моментах:

1. Наш анализатор обрабатывает словоформу не слева направо, а справа налево, что по нашему мнению, лучше соответствует реальному устройству тюркской морфологии.

2. В анализаторе используется грамматический тувинско-русский словарь, существующий в виде электронной базы данных с грамматической, морфонологической, словообразовательной и семантической разметками, сформированный на основе ТРС 1968.

3. В отличие от Washington, Bayyr-ool 2016, мы не используем двух отдельных алгоритмов для анализа глагольных и именных словоформ. Используется единый алгоритм, не выстраивающий парадигмы. Каждая грамматическая категория считается либо выраженной, либо не выраженной, т.о. мы избегаем нулевых аффиксов. См. ниже.

4. Пока не анализируются композитные слова, пишущиеся через пробел, и аналитические словоформы, в т.ч. личные формы глагола. Проблема определения аналитических личных форм глагола, так же как и проблема любых аналитических форм должна решаться на этапе синтаксического анализа с использованием информации об относительном порядке синтетических словоформ.

1. Грамматика

По аналогии с компьютерно-ориентированной моделью хакасской словоформы, с опорой на принципы грамматики порядков мы разработали модель тувинской словоформы и собрали инвентарь словоизменительных показателей для тувинского языка (Дыбо, Шеймович 2015). Для этого использован традиционный инвентарь словоизменительных морфем из Грамматики тувинского языка (ГТЯ 1961) и очерка Ш.Ч. Сата к Тувинско-русскому словарю (Сат 1955). Эта работа была сделана с учетом исследований коллег из научного центра при Тувинском госуниверситете (Хертек, Ооржак 2012; Oorzhak, Khertek 2015; Salchak, Bayir-ool 2015). В настоящей статье представлена несколько доработанная версия модели тувинской словоформы (см. Приложения, табл. 3) с упором на описание сегментных преобразований, с которыми приходится сталкиваться парсингу, оказавшееся куда более сложным, чем для хакасского: процессы ассимиляции согласных (например, при выпадении беглого гласного) последовательно отражаются в тувинской орфографии, что представляет определенные проблемы для автоматического морфологического анализатора.

Имея опыт построения порядковой модели хакасской словоформы, можно сказать о главных отличиях ее от тувинской.

Порядковая модель тувинской словоформы устроена проще хакасской в основном благодаря не так далеко зашедшей фузии внутри сложной глагольной словоформы. Во-первых, в тувинском лицо и число при главной предикации в значительной части случаев (при именах и причастиях) выражаются аналитически, с помощью местоименных частиц («личных показателей» по Сат 1955, 685), по форме совпадающих с личными местоимениями. Внутрь синтетической словоформы попадает только кумулятивное выражение 3 лица ед.числа в форме презенса трех глаголов: *тур-* ‘стоять’, *олур* – ‘сидеть’, *чор* – ‘ходить’; кумулятивное выражение лица и числа при претерите на *-ды* и кумулятивное выражение лица, числа и наклонения в аффиксах императива, условного и предельного наклонений. Во-вторых, в хакасском в синтетическую глагольную словоформу включился ряд акциональных аффиксов, восходящих к вспомогательным глаголам (как дуратив на *-чат-* или презенс на *-ча*, восходящие к глаголу **jat-* ‘лежать’), которые в тувинском сохранили свой аналитический статус. Ср.: хак. *ойнапчабыс* ‘мы играем’ vs тув. *ойнап тур бис* ‘мы играем’.

На первый взгляд тувинская модель проще, в ней меньше мест, пока не предвидится дублирования глагольных слотов.

2. Словарь

Для работы нашего парсера необходим словарь в форме электронной базы данных, содержащей частеречные пометы и указания на чередования основ, не описанные фонологическими правилами, а также некоторую другую информацию (семантическую, словообразовательную, этимологическую и пр.). К настоящему времени средствами СУБД Starling конвертирован в словарную базу данных Тувинско-русский словарь под ред. Э.Р.Тенишева (ТРС 1968), включающий более 17 тыс. лексем. Образцом для словарной статьи тув. словарной базы послужила уже апробированная форма статьи базы данных на основе БХРС. Частей речи в ней выделяется столько же, сколько для хакасского языка: Имя (Nomen), Глагол (Verbum), неизменяемое (Invar). О выделении грамматических классов и частей речи в тюркской морфологии применительно к нуждам компьютерной обработки текста неоднократно говорилось в процессе разработки хакасского парсера и

модели тувинской словоформы, поэтому здесь мы ограничимся отсылками к (Шеймович 2012а-б; Дыбо, Шеймович 2014; Дыбо, Шеймович 2015).

3. Разграничение словообразовательной и словоизменительной морфологии

Морфологическая разметка содержит информацию о словоизменительных, но не о словообразовательных признаках лексемы. Деривативы, записанные в словаре, не являются предметом грамматического парсинга (в словарной базе данных имеются соответствующие поля, содержащие расчлененную морфонологическую запись (DERIV) и словообразовательную разметку (DERIVGLOSS) основы). Специфический вопрос, который здесь встает, это трактовка залогов. Вообще говоря, залог в грамматиках тюркских языков часто принято описывать как словоизменительную категорию. Это связано, во-первых, с относительной регулярностью способов его образования, во-вторых, с синтаксическим характером его грамматического значения. Причины, по которым мы оставляем залог за пределами грамматического анализа, относя его к словообразованию, были неоднократно указаны в предшествующих работах, посвященных хакасскому парсеру (Шеймович 2012а-б; Дыбо, Шеймович 2014 и след.). В тюркологии давно показано, что невозможно выделить порядковое место залога в словоформе (Циммер 1987). Нередки случаи, когда наличие формального показателя залога в слове не связано с наличием у него соответствующего значения: например, основы, содержащие словообразовательный каузативный аффикс, не всегда являются каузативами семантически[1].

[1] См. ГТЯ 276: «Глаголы в понудительном залоге, образованные от переходных глаголов, могут быть как переходными, так и непереходными. В первом случае они имеют понудительное значение, во втором – страдательное. Переходность (понудительное значение) или непереходность (страдательное значение) таких глаголов зависит от контекста. Примеры: 1) Ол алгыны меңээ эттеткен ‘Ту шкуру дали выделать мне’ (понудительное значение); Меңээ эттеткен алгы бо-дур ‘Вот это выделанная мной шкура’ (страдательное значение); 2) Хойларны бөрүге өлүртпес ‘Не давать волкам резать овец’ (понудительное значение); Бөрүге өлүрткен хоювус ийи ‘Волком зарезано две наших овцы’».

Помимо залоговых аффиксов в сфере словообразования нами оставлены (вслед за ГТЯ и в соответствии с наличием этих образований в ТРС и ТСТЯ), например, аффикс «желательности» *(Ы) КсА* (ГТЯ 269) (*алыксаар /ал=ЫКсА-/* желат. от *ал-* (см. *алыр*) ‘хотеть брать’: *бо номну алыксап тур мен* ‘я хочу взять эту книгу’; *суксаар /су=КсА-/* ‘жаждать, сильно хотеть пить’) и аффикс «прекратительности» *БАстА* (ГТЯ 409) (*тоовастаар /тоо=БАстА-/* ‘переставать обращать внимание’, прекрат. от *тоор /тоо-/* ‘обращать внимание’: *Бистиң чамдык аңчыларывыс дииң болгаш ас, күзен тоовастаан* ‘У нас некоторые наши охотники белку и горностая, колонка перестали ценить’ (ТСТЯ 251)).

4. Сегментные преобразования в тувинской грамматике

Что касается сегментных преобразований, то обработка явлений сингармонизма уже была формализована для хакасского парсера. Эти наработки с незначительными изменениями (такими, как наличие губных вариантов (отсутствующих в хакасском) для словоизменительных аффиксов с узким гласным) использованы для тувинского анализатора.

Основной сложностью, с которой сталкивается построение тувинского парсера, в отличие от хакасского, является описание тувинских сандхи – морфонологических процессов, происходящих на границах морфем, т.е. изменений, возникающих вследствие взаимодействия между контактными морфемами: между основой и аффиксами, между соседними аффиксами – с целью последующей их переработки в правила для автоматического глоссирования тувинских текстов.

Как уже было сказано выше, обработка сандхи в тувинском парсере гораздо сложнее, чем в хакасском, в частности, поскольку процессы ассимиляции согласных последовательно отражены в тувинской орфографии (в отличие от хакасской).

Ниже перечислены основные правила сегментных преобразований, действующие в тувинском. Мы излагаем правила поуровнево, в том порядке, в котором они следуют в грамматике синтеза словоформы, поскольку анализ осуществляется «через синтез»,

т.е. путем многократных прямых и обратных проходов с проверкой гипотез[1].

0. **Сегментные правила выбора алломорфов (поверхностно-морфологические правила выбора одного из алломорфов морфемы, при заданном морфемном составе словоформы)**:

0.1. Гласная в скобках, стоящая в начале морфа, обозначает, что если предыдущий морф кончается на согласную, будет выбран вариант данного морфа, начинающийся на эту гласную морфонему; а если предыдущий морф кончается на гласную, будет выбран вариант данного морфа, начинающийся на согласную морфонему (т.е. гласная в скобках опускается). Согласная в скобках, стоящая в начале морфа, обозначает, что если предыдущий морф кончается на гласную, будет выбран вариант данного морфа, начинающийся на эту согласную морфонему; а если предыдущий морф кончается на согласную, будет выбран вариант данного морфа, начинающийся на гласную морфонему (т.е. согласная в скобках опускается).

0.2. Алломорфы, стоящие в одной клетке таблицы 3 через запятую (Fut *ЫР, Ар*; ConvPraes *А, ЫЛ*), выбираются согласно информации в грамматическом словаре основ, в поле ALTERNATEN.

0.3. Алломорф ConvPraes *-й* выбирается в случае, если предыдущий морф оканчивается на гласную (*бода-* ‘думать’ – *бода-й* ‘думая’, *тары-* ‘сеять, сажать’ – *тары-й* ‘сея, сажая’).

1. **Правила морфологических чередований**, применение которых требует информации из словарей морфем (из грамматического словаря основ – например, сведения о лексико-грамматическом классе основы или словарные данные этой основы, – и

[1] Для облегчения использования человеком электронного грамматического словаря тувинского языка мы приняли решение выписывать в морфонологической транскрипции, т.е. с использованием условных букв, только звуковую оболочку аффиксов, а не корней (морфонологическая транскрипция словообразовательных аффиксов дается внутри расчлененной морфонологической записи основы в поле словарной базы DERIV). Это решение является чисто формальным. Отсутствие условной морфонологической записи корней компенсируется информацией из полей словарной базы ALTERNAT, ALTERNATEN и FORM; поскольку соответствующая информация является релевантной для небольшого числа основ, такое решение представляется оправданным. Ниже в тексте статьи мы приводим примеры основ в условной морфонологической записи.

из таблицы словоизменительных аффиксов – например, граммема данной морфемы).

1.1. Преобразование конечных (стоящих непосредственно перед морфологической словоизменительной границей[1]) согласных *-к, -К* неодносложной последовательности морфем на границе с аффиксом принадлежности (*Ы, Ым*, *Ың, Ывыс, ЫңАр*) в *Г*: *тавак + Ым > таваГ-Ым* ‘мое блюдо’; *белек + Ым > белеГ-Ым* ‘мой подарок’, *кезек*[2] *+ Ы > кезеГ-Ы* (*> кезээ*) ‘его часть’ (для односложной последовательности правило не работает: *аак + Ы > аак-Ы* ‘его последствие’). Информация о наличии чередования извлекается из сегментного состава морфем, стоящих перед морфологической границей, и сегментного состава и имени аффикса принадлежности.

1.2. Преобразование конечных (стоящих непосредственно перед морфологической словоизменительной границей) *к, п, м, л*[3] некоторых односложных глагольных основ на границе с афф. деепричастия на *Ып* в *Г*: *бол- + Ып > боГ-Ып/бол-Ып (> бооп/болуп)* ‘быв’, *кел- + Ып > кээп* ‘придя’, но *бил- + Ып > билип* ‘узнав’. Информация о наличии и обязательности преобразования извлекается из полей ALTERNATEN или FORM базы грамматического словаря основ и сегментного состава и имени аффикса деепричастия[4].

---

[1] В условной морфонологической расчлененной записи граница между словообразовательными и словоизменительными показателями различается (у нас это «=» для словообразовательных и «-» для словоизменительных). Формулируемое правило учитывает только «-».

[2] В условной записи в поле DERIV *кеС=Ак*, но на вход при порождении поступает лексическая основа.

[3] По данным ГТЯ 38, опционально это правило действует также для односложных глаголов на *-ң* (*доң-Ып > дооп/доңуп* ‘замерзнув’), но по примерам в ТСТЯ все такие основы сохраняют *ң*. Вопрос нуждается в корпусном и диалектологическом исследовании.

[4] На этом этапе технически имеет смысл просто извлекать готовую форму деепричастия из поля словарной базы. Дело в том, что один из глаголов на *-л* (*ал-* ‘брать’), все глаголы на *-к*, для которых релевантно выпадение, и по крайней мере часть глаголов на *-п* получают после выпадения согласного в форме деепричастия на *-Ып* краткую, а не долгую гласную: *ап* ‘взяв', *хып* ‘загоревшись’ (*хывар*), *соп* ‘ударив’ (*согар*). К сожалению, полная информация по этим формам отсутствует и в грамматиках, и в словарях тувинского языка; в части случаев извлекается из примеров в словарных статьях ТСТЯ. Вопрос нуждается в корпусном исследовании.

1.3. Преобразование *р, л* некоторых односложных глагольных основ на границе с афф. деепричастия будущего времени *Ыр*, *Ар* в *Г*: *кел- + Ар > кеГ-Ар/кел-Ар* (*> кээр/келир*) ‘приходить’, *бер- + Ар > беГ-Ар* (*> бээр*) ‘давать’. Информация о наличии и обязательности чередования извлекается из поля ALTERNATEN базы грамматического словаря основ и имени аффикса деепричастия.

1.4. «Беглые» гласные.

Узкая гласная *Ы* закрытого конечного слога последовательности морфов, если этой гласной предшествует одиночный согласный (т.е. в последовательности *(С)VCЫC*), если эту последовательность не разрывает словоизменительная морфологическая граница, может выпадать, если к ней присоединяется аффикс, начинающийся на гласную. Эти процессы зависят от словарной информации о составляющих форму морфемах, а не только от их поверхностной структуры и морфемных границ, и потому не могут считаться морфонологическими. Ср. разное развитие в однотипных морфонемных последовательностях: *бүрЫн-Ы > бүрүн-үбиле* ‘полностью’ – 3Poss от *бүрүн* ‘всё’ – без выпадения и *бурЫн-Ы > мурну* ‘перед, раньше’ – 3Poss от *бурун* ‘прежний’ (о начальной согласной см. ниже), – с выпадением; *дайын* ‘война’ – *Ада-чурттуң Улуг дайын-ы* ‘Великая Отечественная война’ – без выпадения, и *оюн* ‘игра’ – 3Poss *ойн-у* – с выпадением; инфинитивы *ажын-ар* /*аШы=н-* 1/ ‘сердиться, дуться, злиться’ без выпадения – *ашт-ыр* /*аШы=н-* 2/ ‘оправдаться, доказать свою невинность’ с выпадением (и ассимиляцией, см. ниже). Вид основы с выпавшей гласной указан в статье словарной базы в поле FORM.

Примеры.

Имя:

Выпадение беглой гласной из 2-го закрытого слога двусложной именной основы при присоединении афф. принадлежности: *ойЫн-Ы > ойн-Ы > ойну* ‘его игра’ (при формулировании правила для автоматического парсера графема *ю* раскрывается в виде *йу*, см. ниже); *эрин-Ы > эрн-Ы > эрни* ‘его губа’; *оГЫл-Ы > огл-Ы > оглу* ‘его сын’.

Глагол:

Выпадение беглой гласной из залогового аффикса (иногда уже окаменевшего): *хайын-Ыр > хайн-Ыр > хайныр* ‘кипеть’, *дир=Ыл-Ы > дирл-Ы > дирли* ‘ожив’, *ажы=н=Ыш-Ыр > ажыныш-Ыр > ажынчыр* ‘сердиться друг на друга’.

1.5. Уникальные чередования: основа *бөрт* ‘шапка’ в позиции перед вокалическим началом следующего морфа имеет форму *бөрг-*: *бөрт + Ы > бөрг-ү* ‘его шапка’. Основа *аас* ‘рот’ в той же позиции имеет форму *акс-: аас + Ы > аксы* ‘его рот’[1]. Сведения – в словарной базе, в поле FORM.

2. **Морфонологические правила**, применение которых определяется сегментным составом сочетающихся морфем, включающим информацию о морфемных границах.

2.1. Ассимиляции в сочетаниях согласных морфонем, не разделенных словоизменительной границей. Сочетания согласных морфонем, не разделенные словоизменительной морфологической границей, претерпевают следующие ассимиляции:

2.1.1. (Если при присоединении аффиксов выпадает узкая гласная, стоявшая в позиции между согласными *л-н,* то) сочетание согласных *лн > нн*:

Имя: *келин + Ы > келн-Ы > кенн-Ы > кенни* ‘его невестка’;

Глагол: *лVн > нн*: *кыл=ын- + Ыр > кылн-Ыр > кынн-Ыр > кынныр* ‘сделаться’;

2.1.2. (Если при присоединении аффиксов выпадает узкий гласный, стоявший в позиции между согласными $С_{сильн}Vн$, $С_{сильн}$-*л*, то) *CVн, CVл > Ст*[2]:

Имя: *иШин + Ы > иШн-Ы > ишт-Ы > ишти* ‘его живот’, *эКин + Ы > эКн-Ы > экт-Ы > экти* ‘его плечо’;

Глагол: *саКын-* ‘вспоминать’ + *Ып> сактып*; *оТун-* 'просыпаться' + *Ыр > оттур, тыП=ыл-* ‘находиться’ + *Ыр > тыптыр*

2.1.3. (Если при присоединении аффиксов выпадает узкий гласный, стоявший в позиции между согласными $С_{сильн}Vш$, или

---

[1] Эта основа должна была бы получить условную запись аГЫс, но по обычным правилам выпадения и ассимиляции мы бы получили *агз-ы, ср. процессы в глаголе дагзыр /даас-/ ‘обязывать кого-либо’.

[2] Таким образом, очевидно, что действие правила выпадения узкой гласной в тувинском предшествует действию правила интервокального озвончения.

лVш, то) *CVш > Сч: диК=иш-* ‘помочь строить’ + *ЫР >дикчир*[1], *кыП=ыс+ЫI > кыпсы* ‘зажигая’. (Если при присоединении аффиксов выпадает узкий гласный, стоявший в позиции между согласными *рVш,* то) *рVш > рж: таПар=ыш-* ‘встречаться’ + *ЫР > таваржыр.*

С сандхи типа 2.1. связан еще один тип фонетических изменений, дистантная регрессивная ассимиляция согласной в начале слова, вызванная последней согласной в пределах одного закрытого слога:

2.2.1. В именных и глагольных основах вышеописанного типа с начальной *б-*: *бил=ин* ‘признаваться’ + *ЫР > билн-ЫР > биннЫР* (контактная регрессивная ассимиляция по назальности) *> миннир*; *бурЫн-ЫI > мурну* ‘перед, раньше’ – 3 Poss от *бурун* ‘прежний’. Это чередование также отражается в словаре в поле ALTERNATEN.

2.2.2. Чередование согласных в начале односложных глагольных основ, оканчивающихся на сильный согласный и с фарингализованным гласным при присоединении к ним аффиксов на гласную: *хыъп* ‘гори!’ + Fut *Ар > кыъвар* ‘гореть’, *төък* 'вылей!' + Fut *Ар > дөъгер* ‘выливать’. Это чередование лишь частично отражено в тувинской орфографии (см. Пальмбах 1956, 111-112), а именно отражено для заднеязычных и не отражено для зубных согласных. Оно также сводится к дистантной ассимиляции начального согласного согласному, закрывающему слог, по силе/глухости/придыхательности в условиях наличия фарингализованной гласной. Мы записываем его в поле ALTERNATEN в случае отражения в орфографии.

2.3. Ассимиляция согласных, разделенных словоизменительной морфемной границей (см. табл. 1).

Аффиксы, начинающиеся с определенной морфонемы, получают реализацию этой морфонемы в зависимости от качества конеч-

---

[1] В словообразовательной морфонологии (т.е. в соседстве с морфемными границами типа =) сочетаемость согласных морфонем отличается, и, в частности, имеются пары слов с разным поведением одного и того же аффикса, привязанным к разнице значений (*садыг=жы* ‘торговец’, ‘купец’ // *садык=чы* ‘продавец’; ср. *садыг* ‘торговля’, ‘продажа’, *сат-* ‘торговать’, ‘продавать’; *тараа=жы* ‘хлебороб’ // *тараа=чы* ‘тот, кто едет за хлебом’; ср. *тараа* ‘хлеб’), см. ГТЯ 149. Частично это объясняется бытованием монголизмов на *-чи*, см. ГТЯ 169.

ной морфонемы предшествующего морфа[1] (инвентарь словоизменительных морфем в морфонологической условной записи см. в таблице 3; условная запись прочих морфем извлекается из полей лексической базы ALTERNAT, ALTERNATEN, или DERIV; либо, если эти поля в словарной статье не заполнены, то совпадает с начальной формой слова).

Аффиксы с начальной морфонемой *Б-* получают *в* после гласной, *б* после слабой неносовой согласной, *п* после сильной, *м* после носовой;

Аффиксы с начальной морфонемой *К-* получают *г* после слабой согласной или гласной, *к* после сильной согласной.

Аффиксы с начальной морфонемой *Г-* получают 0 после гласной (выпадение с последующим стяжением гласных вокруг морфемной границы в долгую *АА,* **с последующим сингармоническим преобразованием в зависимости от рядности гласной основы)**: *хову + ГА > ховаа* ‘к степи’, *кижи + ГА > кижээ* ‘человеку’*; сана- + ГАн > санаан* ‘сосчитал’, *номчу- + ГАн > номчаан* ‘прочитал’, *бижи- + ГАн > бижээн* ‘написал’, *чүлү-+ ГАш > чүлээш* ‘побрив’, *сөгле- + ГАй > сөглээй* ‘можешь сказать’, *ойна- + ГАлА > ойнаала* ‘как только поиграл’*, номчу- + ГЫже > номчааже* ‘пока не прочитает’); *г* после слабой согласной, *к* после сильной согласной.

Аффиксы с начальной морфонемой *Т-* получают *д* после слабой согласной или гласной (*уруг-да* ‘в ребенке’), *т* после сильной согласной (*тавак-та* ‘в блюде’).

Аффиксы с начальной морфонемой *Н-* получают *д* после слабой неносовой согласной или гласной, *н* после носовой согласной, *т* после сильной согласной.

Аффиксы с начальной морфонемой *С-* получают *з* после слабой согласной или гласной, *с* после сильной согласной.

Аффиксы с начальной морфонемой *Л-* получают *л* после гласной или слабой неносовой согласной (но не *л*), *т* после сильной, *н* после носовой; *д* – после *л*.

Аффиксы с начальной морфонемой *Ч-* получают *ч* после сильной согласной, а также после сонорных согласных *л, м, н, ң*; *ж* – после гласных и слабой *г,* а также после сонорных *й, р.*

---

[1] Эти процессы связаны с наличием фонологических запретов на сочетаемость согласных, см. табл. 4 в приложении.

Таблица 1

Начальные морфонемы афф. Конечные морфонемы основы	**Б**	**К**	**Т**	**Н**	**С**	**Л**	**Ч**
V	в	г	д	н	з	л	ж
т, Т	п	к	т	т	с	т	ч
п, П	п	к	т	т	с	т	ч
м	м	г	д	н	з	н	ч
н	м	г	д	н	з	н	ч
ң	м	г	д	н	з	н	ч
л	б	г	д	д	з	д	ч
р	б	г	д	н	з	л	ж
й	б	г	д	д	з	л	ж
с, С	п	к	т	т	с	т	ч
к, К	п	к	т	т	с	т	ч
г, Г	б	г	д	д	з	л	ж
ч, Ч	п	к	т	т	с	т	ч
ш	п	к	т	т	с	т	ч

3. **Фонологические правила**, включающие исключительно фонологические условия (не зависящие от морфологического членения словоформы). Это фактически правила перехода с морфонемного на фонемный уровень записи, при котором осуществляется пересчет «морфонем» предыдущего уровня в «фонемы». Для парсера, работающего с текстом в орфографической форме, это пересчет с записи, содержащей условные буквы, в просто буквенную запись.

3.0. Все границы морфем ликвидируются. Если с обеих сторон от морфемной границы стояли гласные (уже без скобок – скобки были убраны на этапе 0), то эти гласные стягиваются в одну долгую, **являющуюся соотносительной парой по долготе гласной, предшествовавшей** морфемной границе. На этом этапе, например, происходит стяжение конечного гласного глагольной осно-

вы с первым гласным афф. деепричастия будущего времени *ЫР*, *Ар*: *ойна- + Ар > ойнаар* ‘играть’, *сөгле- + Ар >сөглээр* ‘сказать’, *чыры- + ЫР > чырыыр* ‘светить’ (в словарной базе такие стяженные глагольные формы представлены в поле FIELD1, т.к. деепричастие будущего времени используется в тувинском языке в качестве неопределенной формы глагола (инфинитива)).

3.1. Все *Г, г*, имеющиеся в условной записи словоформы до сих пор и оказавшиеся в интервокальной позиции, выпадают, а окружающие их гласные стягиваются в одну долгую, **являющуюся соотносительной парой по долготе гласной, предшествовавшей выпавшей *Г*:** *оГЫл > оол, аГЫс > аас, таваГ-ЫМ > таваам* ‘мое блюдо’*; уруг-Ы > уруу ‘его ребенок’; белеГ-ЫМ > белээм* ‘мой подарок’; *кес=еГ-Ы > кезээ* ‘его часть’; *даг-Ы > даа* ‘его гора’; *суг-Ы > суу* ‘его вода’; *саг-ЫП > саап* ‘подоив’, *каг-Ар > каар* ‘оставлять’, *чуг-ЫР > чуур* ‘мыть’, *кеГ-Ар/кел-Ар > кээр/ келир* ‘приходить’*, беГ-Ар > бээр* ‘давать’.

3.2. Ослабление (графическое озвончение) глухой согласной при попадании ее в интервокальную позицию. Для парсинга существенно это озвончение на границе основы и аффикса, начинающегося на гласную: *ат* ‘имя’ – *ад-ы* ‘его имя’, *час* ‘весна’ – *чазын* ‘весной’, *аак-Ы* ‘его последствие’ > *аагы*: *соок аагы* ‘последствия мороза’; но в принципе это правило действует независимо от прохождения морфемных границ: *эКЫН* ‘плечо’ > *эгин, иШЫН* ‘живот’ > *ижин*.

3.3. Рядный и губной сингармонизм (см. подробное описание ГТЯ 46-49). Большинство словоизменительных аффиксов имеют заднерядный, и переднерядный, огубленный и неогубленный варианты, которые выбираются в соответствии с характеристикой последней гласной словоизменительной основы по рядности и огубленности. Аффиксы, содержащие гласную морфонему А, выбирают алломорф с *а* после заднерядной гласной; с *е* после переднерядной гласной. Аффиксы, содержащие гласную морфонему Ы, выбирают алломорф *ы* после заднерядной неогубленной гласной; *и* после переднерядной неогубленной гласной, *у* после заднерядной огубленной гласной, *ү* после переднерядной неогубленной гласной. Это правило в принципе действует независимо от границ и состава морфем в словоформе, так что его можно переформулировать так:

Таблица 2

$\mathbf{V_n}$ / $\mathbf{V_{n-1}}$	**А**	**Ы**
V_{back}	а	ы
V_{front}	е	и
$V_{front\ lab}$	е	ү
$V_{back\ lab}$	а	у

3.4. Все сохранившиеся на этом этапе условные буквы (здесь в морфонологической транскрипции применены как условные кириллические буквы верхнего регистра) преобразуются в соответствующие буквы нижнего регистра, таким образом возвращая орфографическую форму тувинской словоформы (до обработки графическими правилами).

4. **«Графические сандхи»**, обусловленные способом употребления кириллической графики (запись сочетаний звуков вида *й+V* с помощью кириллических йотированных букв), например:

4.1. *йу > ю, йа > я, йе > е* (ГТЯ 43).

Примеры:

хой ‘овца’ – Poss.1Pl *хоювус* ‘наша овца’ < *хой-ЫвЫс*;

ой- ‘прорубить, пробивать’, Prosp *ойгаш* < *ой-КАш*, инфинитив *ояр*< *ой-Ар*, ConvPast *оюп* < *ой-Ып* (ТСТЯ 481);

өй- ‘валять, катать’, Prosp *өйгеш* < *өй-КАш* и инфинитив *өер* < *өй-Ар,* ConvPast *өйүп* < *өй-Ып* (ТСТЯ 490).

4.2. Сочетание букв *ее*, возникшее при порождении словоформы, на поверхностном уровне заменяется на ээ: *кел+Ыр > кеер >* Fut *кээр* ‘приходить’ (см. правило 1.3).

Заключение

Можно видеть, что сегментные процессы, обусловливающие синхронный облик тувинской словоформы, представляют довольно сложную, многоуровневую систему, которая может быть проинтерпретирована как конгломерат разновременных фонетических процессов, постепенно вытеснявшихся на более глубинные уровни языка, а также вымывавшихся аналогическими процессами. Ср. рассуждения о «частичном восстановлении исторического

облика» некоторых тувинских слов в определенных морфологических условиях в ГТЯ 118. Автоматический морфологический анализатор может обойти часть этих явлений с помощью «лексических исключений», поместив соответствующую информацию в словарные статьи лексической базы, но значительное их число все-таки должно обрабатываться алгоритмом. Отметим, что существующие словари и грамматики тувинского языка содержат значительные лакуны в описании сегментного поведения ряда морфем и типов морфем, которые могут быть закрыты только с помощью корпусного исследования.

ЛИТЕРАТУРА

1. Aelita Salchak, Aziyana Bayir-ool. The main results of the project on creation an electronic corpus of Tuvan language // Сборник трудов конференции «TurkLang-2015». Казань, 2015. С. 259–268.

2. Oorzhak B., Khertek A. Development of semantyc markup the corpus of Tuvan language // Сборник трудов конференции «TurkLang-2015». Казань, 2015. С. 351–362.

3. Gleason 1955 – Gleason, H. Introduction to descriptive linguistics. New York, 1955: Holt, Rinehart and Winston.

4. ГТЯ – Исхаков Ф.Г., Пальмбах А.А. Грамматика тувинского языка. М.,1961. 473 с.

5. Дыбо А.В., Шеймович А.В. Автоматический морфологический анализ для корпусов тюркских языков // Филология и культура. 2014. № 2 (36). С. 20–26;

6. Дыбо А.В., Шеймович А.В. Автоматический морфологический анализ для корпусов хакасского и древнетюркского языков // Научное обозрение Саяно-Алтая. Серия: Филология. № 8, 2014. С. 9–30.

7. Дыбо А.В., Шеймович А.В, Крылов С.А. Некоторые возможности семантической и этимологической разметки для корпусов тюркских языков (расстановка семантических тэгов в электронном хакасско-русском словаре // Сборник трудов международной конференции TurkLang 2015. Казань, 2015. С. 304–327. 488 с.

8. Дыбо А.В., Шеймович А.В. Порядковая модель тувинской словоформы // Материалы региональной конференции «Языки народов Сибири и сопредельных регионов». Новосибирск, ИФ СО РАН, 6–9 октября 2015 г.

9. Мальцева В. 2004 – Мальцева В.С. Структура глагольной словоформы в сагайском диалекте хакасского языка (говор с. Казановка). Дипломная работа. М., 2004.

10. Пальмбах 1956 – Пальмбах А.А. Система согласных тувинского языка и ее отражение в письменности. // Ученые записки Тув. научно-исслед. института языка, литературы и истории. IV. Кызыл, 1956.

11. ТРС 1968 – Тувинско-русский словарь / Под ред. Э.Р. Тенишева. М., 1968.

12. Сат 1955 – Сат Ш.Ч. Тувинский язык (краткий очерк) // Тувинско-русский словарь / под ред. А.А. Пальмбаха. М., 1955. С. 613–721.

13. ТСТЯ – Толковый словарь тувинского языка. Т. II. Новосибирск 2011.

14. Хертек А.Б., Ооржак Б.Ч. О морфологической разметке электронного корпуса текстов тувинского языка // Филологические науки. Вопросы теории и практики. Тамбов. 2012 г. № 7 (18). Ч. II. С. 214–218.

15. Циммер К. О некоторых ограничениях на каузативизацию в турецком языке // Новое в зарубежной лингвистике. Т. XIX. 1987. С. 283–298.

16. Шамина Л.А. Аналитические грамматические формы и конструкции в функции сказуемого в тувинском языке. Новосибирск, 2010 г., 240 с.

17. Шеймович 2012а – Шеймович А.В. Некоторые особенности автоматического анализа морфологии хакасского языка (на материале корпуса) // д. на Международном научном форуме «Н.Ф.Катанов и современность». Абакан, 2012.

18. Шеймович А.В. О принципах построения автоматического морфологического анализатора для корпуса хакасского языка // д. на Отчетном собрании Российского комитета тюркологов. Москва, 2013.

19. Washington, Bayyr-ool 2016 – Washington J. N, Bayyr-ool A. et al. The development of a finite-state morphological analyser for Tuvan // Родной язык, № 1(4), М., 2016.

Таблица 3. **Модель изменяемой тувинской глагольной**

0	1	2	3	4	5
S	**Distr**	**Perf**	**Neg**	**Tense (Praes, Past, Future) Mood**	**Num(Pl)**
	Distr *КЫлА*	Perf *(Ы)ББЫт*	Neg *БА*	Praes *ыр*	Pl *ЛАр*
			Neg.Conv *БАйн*	Past *ТЫ*	
			Neg.Fut *БАс*	PastRes *ГАн*	
				PastRtr *ЧЫк*	
				Cont *ББЫшААн*	
				Fut *Ыр, Ар*	
				Cunc *ГАлАк*	
				ConvLim *ГАлА*	
				ConvPast *(Ы)п*	
				ConvPraes *А, Ы, й*	
				Prosp *ГАш*	
				ConvLim2 *ГЫже*	
				Abstr *КЫ*	
				Opt. *ГАй*	

ПРИЛОЖЕНИЯ

словоформы и набор тувинских словоизменительных аффиксов

6	7		8	9
Poss	**Case**		**Person (1, 2)**	**Ptcl**
	Simple declencion	**Possessive declencion**		
1pos.sg *(Ы)м*	Gen *НЫң*	Gen *НЫң*	1sg *м*	Interr *(Ы)л*
2pos.sg *(Ы)ң*	Dat *ГА*	Dat *нГА*	2sg *ң*	Add *-даа*
3pos. *(з)Ы*	Acc *НЫ*	Acc *Н*	1pl *ББІс*	Indir *-ТЫр*
1pl *(Ы)ББІс*	Loc *ТА*	Loc *нТА*	2pl *ңАр*	Emph *-Ла*
2 pl *(Ы)ңАр*	Abl *ТАн*	Abl *нТАн*	3pl *ЛАр*	
Gen.3Pos *НЫЫ*	Lat1 *Че*	Lat1 *нЧе*	Praes.3sg *у*	
	Lat2 *ТЫвА*	Lat2 *нТЫвА*	Imp.1sg *ЫЙн, Айн*	
	Lat3 *КЫдЫ*	Lat *КЫдЫ*	Imp.3sg *СЫн*	
			Imp.1.Dual *ААл(Ы), ЫЫІл(Ы)*	
			Imp.1.pl *ААлЫңАр, ЫЫІлЫңАр,*	
			Imp.2.pl *(Ы)ңАр*	
			Imp.3.pl *СЫн.(нАр)*	
			1.sg.Lim *м.че*	
			2.sg.Lim *ң.че*	
			1.pl.Lim *вис.че*	
			2.pl.Lim *ңер.же*	
			1.sg.Cond *СЫ.м.зА*	
			2.sg.Cond *СЫ.ң.зА*	
			3sg,pl Cond *зА*	
			1.pl.Cond *СЫ.вЫс.сА*	
			2.pl.Cond *СЫ.ңАр.зА*	

Условные обозначения

Abstr – абстрактное имя действия (герундий)

Add – аддитивная частица (*и, ни, же, ведь…*)

Case – падеж

Simple declension – набор падежных аффиксов простого склонения

Possessive declension – набор падежных аффиксов притяжательного склонения (после показателя принадлежности)

Список падежей

Nom – номинатив, или нулевой падеж, отсутствует в таблице, т.к. не имеет поверхностного выражения

Gen – генетив (родительный)

Dat – датив

Acc – аккузатив (винительный)

Loc – локатив (местный)

Abl – аблатив (исходный)

Lat – латив (направительный)

Instr – инструментальный (творительный)

Cond – условное наклонение

Conv – деепричастие

ConvLim – деепричастие предела в прошлом

Cunc – кункатив, еще не совершившееся действие

Distr – дистрибутив, обозначает множественность субъекта или объекта действия;

Emph – эмфатическая частица (*-mo*)

Fut – будущее время

Imp – императив (повелительное наклонение)

Indir – индиректив, косвенная эвиденциальность (неочевидность либо заглазность) действия

Interr – вопросительность

Lim – предельность («Предельное наклонение»)

Mood – наклонение

Neg – отрицание

Neg.Conv – отрицательная форма деепричастия

Neg.Fut – отрицательная форма буд.вр.

Num (Sg, Pl, Dual) – число (ед., мн., двойств.)

Opt – желательное наклонение

Past – прошедшее время

Perf – перфектив (завершенность действия)
Person, prs – лицо
Poss, pos – принадлежность
Praes = Pres – настоящее время
Prosp – проспектив (состояние, предшествующее действию)
Ptcl – частица (пишущаяся слитно со словоформой, а также вставная)

S – основа. Основа включает корень со словообразовательными показателями и присутствует в словаре в качестве заголовка словарной статьи. Регулярное наличие показателя в словаре в составе заглавного слова служило критерием невключения того или иного показателя (например, аффиксков деятеля) в разряд словоизменительных.

Tense – время

Кумулятивно выраженные граммемы разделяются точками.

Значение условных символов в системе морфонем:

Согласные морфонемы
Б: б/п/м/в
К: *г/к*
Г: *г/к/0*
Т: т/д
Н:д/т/н
С: с/з
Л: л/т/д/н
Ч:ч/ж/ш

Гласные морфонемы
А: е, а
Ы: и, ы, у, ү

Таблица 4. **Фрагмент системы допустимых сочетаний согласных в тувинском языке**

	т	д	л	н	с	з	ч	ш	ж
с	кес-тик	–	–	–	бас-сын	–	кас-чыр-	–	–
т	эът-тиг	–	–	–	четсе	–	бутчак	–	–
к	өөк-тээр	–	–	–	сукса-	–	көрүк-чү	акша	–
ш	чеш-тин-	–	–	–	ыш-сыг	–	дашчы	–	–
р	эрте	көрдү	баар-лыг	силер-ниң	–	барза	кадар-чы	–	хааржак

л	–	мал-дың	–	–	–	эл-зиит-	болчур	–	–
м	–	амдан вкус	–	эмне-	–	хүлүм-зүр-	кымчы	–	–
н	–	ындыг	–	хүннер	–	хүнзе-	кинчи	–	–
ң	–	аңдар-	–	деңне-	–	даңзы	аңчы	–	–
й	ойта-яр	шай-дан	чай-лаг	ойнаар	–	дай-зын	–	–	хоорай-же, далай-жы
г	–	дагда	суг-лук	дагны	–	аарыг-зыыр	садыг-чы	–	дагже

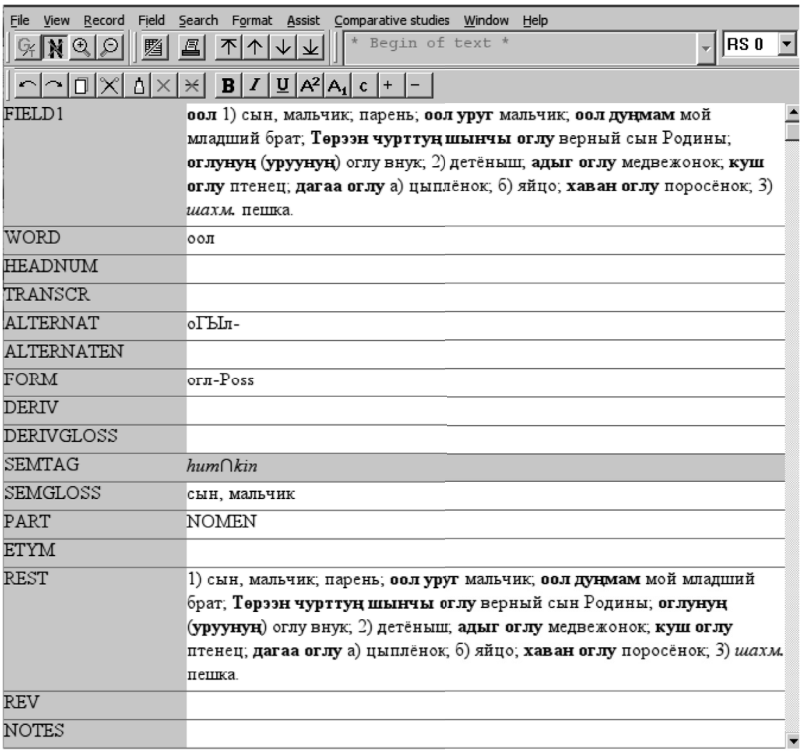


Рис. 1. Образец именной статьи тувинско-русской электронной базы данных

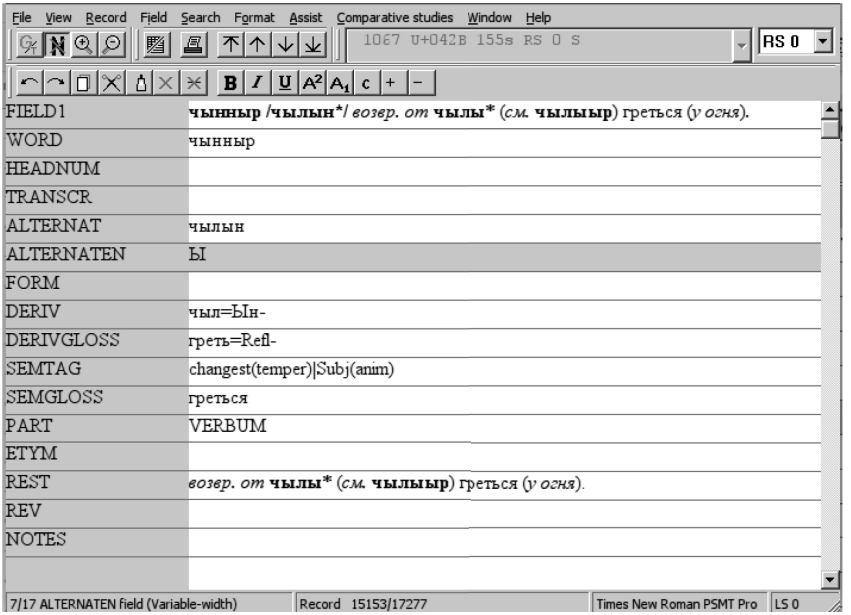


Рис. 2. Образец глагольной статьи тувинско-русской электронной базы данных

УДК 81’33

MORPHOLOGICAL ANALYZER OF TURKIC WORD FORMS BASED ON THE STRUCTURAL-FUNCTIONAL MODEL OF THE TURKIC MORPHEME

A. Gatiatullin, A. Bashirov
Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic, Kazan, Russia
ayrat.gatiatullin@gmail.com, a.basheerov@gmail.com

The article describes methods and technologies used in the computer applications for morphological analysis and synthesis of Turkic word forms. The modules are used in a multifunctional multilingual web service called ModMorph II, which has been developed by the Institute of Applied Semiotics in the Tatarstan Academy of Sciences. The morphological analysis and synthesis modules are based upon the structural and functional model of the Turkic morpheme.

Keywords: Morphological analyzer, multifunctional multilingual service, model of the Turkic morpheme.

МОРФОЛОГИЧЕСКИЙ АНАЛИЗАТОР ТЮРКСКИХ СЛОВОФОРМ НА БАЗЕ СТРУКТУРНО-ФУНКЦИОНАЛЬНОЙ МОДЕЛИ ТЮРКСКОЙ МОРФЕМЫ

А. Р. Гатиатуллин, А.М. Баширов
Институт прикладной семиотики АН РТ
ayrat.gatiatullin@gmail.com, a.basheerov@gmail.com

В статье описываются методы и технологии, использованные в программах морфологического анализа и синтеза тюркских словоформ. Представленные модули используются в многофункциональном многоязычном Интернет сервисе ModMorph II, разрабатываемом в институте прикладной семиотики Академии наук Республики Татарстан. Модули морфологического анализа и синтеза разработаны на основе структурно-функциональной модели тюркской морфемы.

Ключевые слова: морфологический анализатор, многофункциональный многоязычный сервис, модель тюркской морфемы.

Введение

Задача морфологического анализа является одной из первоочередных при компьютерной обработке языка и входит в качестве основного компонента в целый ряд языковых программ. Морфологические анализаторы для тюркских языков разрабатываются начиная с 60-х годов 20-го века [1, 2, 3]. Основной особенностью первых разработок является, то что, они все являются языкозависимыми, поскольку морфотактические правила языка были встроены в саму программу. Соответственно для использования этой программы с другим тюркским языком необходимо переписывать сам программный код. С течением времени технологии менялись, создавались анализаторы, универсальные для разных тюркских языков, увеличивались размеры словарей, росла скорость обработки, процент анализируемых слов. Несмотря на то, что с начала создания первых программ морфологического анализа для тюркских языков прошло уже около 50 лет, работы в данном направлении не прекращаются и в настоящее время. Подтверждением этого являются следующие работы по морфологическим анализаторам: универсального [17], для татарского [9-12], башкирского [13,14], казахского [15], чувашского [16], турецкого [7], хакасского [8] и др. языков.

В настоящее время программа морфологического анализа может включать следующие основные компоненты:

– стемматизация,

– лемматизация,

– приписывание граммем.

Стемматизация (или стемминг от англ. stemming) – это процесс нахождения основы (псевдоосновы) слова для заданного исходного слова. При этом, получаемая псевдооснова не обязана совпадать с грамматической основой рассматриваемой словоформы; достаточно, чтобы словоформы, соответствующие одной парадигме, получали в результате работы алгоритма одну и ту же псевдооснову. Фактически – это максимальная общая часть всех слов парадигмы.

Примеры:

Русский язык. Для словоформ *зеленый, зелень, зеленеть, зеленеющий* в результате стемматизации будет получена псевдооснова зелен.

Татарский язык. Для словоформ *китаплар* и *китабым* псевдооснова будет *кита*.

Лемматизация – процесс приведения словоформы к лемме. Лемма – нормальная (словарная) форма слова.

Пример:

Лемматизация наиболее актуальна для языков типа русского, так как для тюркских языков довольно часто псевдооснова совпадает с леммой. В русском языке лемма имени существительного – это форма слова в единственном числе (если оно есть у существительного) и именительном падеже. Словоформе *столов* соответствует лемма СТОЛ. Лемма имени прилагательного – это форма слова в единственном числе мужского рода. Словоформе *зеленых* соответствует лемма ЗЕЛЕНЫЙ.

В тюркских языках понятие леммы для глаголов может быть неоднозначным, так как в одних словарях может быть использована безаффиксальная форма в повелительном наклонении, а в других словарях часто используется форма имени действия, которая не является кратчайшей формой, а образуется путем присоединения аффиксальной морфемы -У.

Приписывание словоформе множества наборов граммем – приписывание словоформе грамматических характеристик (грамматических признаков). Граммема (грамматическая характеристика) – это элементарный морфологический описатель, относящий словоформу к какому-то определенному морфологическому классу.

Пример.

Словоформе *стол* с леммой СТОЛ будет приписан следующий набор граммем: (*мр, ед, им, неод*) и (*мр, ед, вн, неод*). Такой способ представления подходит для флективных языков, к которым относится русский язык.

Для тюркских языков такая форма представления граммем не подходит по тем причинам, что в тюркских языках, как правило, одна морфема соответствует одной грамммеме и одна и та же морфема множественности может встретиться в словоформе несколько раз.

Например:

урманнардагыларга – урман+ ЛАр+ДАгЫ+ЛАр+ГА.

По используемым технологиям морфологического анализа, используемым для языков агглютинативного типа, можно выделить два основных подхода:

- парадигматический [4,5];
- автоматный [6,7].

В парадигматическом подходе используются два словаря: словарь основ и словарь парадигм. Принцип работы морфологических анализаторов, базирующихся на этом подходе состоит в том, что каждой лемме в словаре присваивается индекс типа парадигмы, который отсылает к списку образцов парадигм. Парадигматический подход чаще всего используется для анализа флективных языков, в число которых входит и русский язык. Во флективных языках размеры парадигм невелики, но зато велико количество этих парадигм. В базе данных анализатора хранится полная парадигма для каждого типа основ. Такой парадигматический подход был использован и в анализаторе UniParser Архангельского [17].

База данных UniParser содержит следующие файлы:

– список лексем, для которых указываются основы, словоизменительные классы и любая другая лексическая информация, т. е. та информация, которая должна быть приписана каждой словоформе данной лексемы;

– список словоизменительных показателей, объединённых в парадигмы разных словоизменительных типов.

Однако тюркские языки обладают целым набором структурных особенностей, из-за которых парадигматический подход для работы с ними не всегда эффективен.

Это такие свойства тюркских языков, которые отличают их от индоевропейских:

– автоматная правосторонняя морфология,

– агглютинация,

– отсутствие жесткой границы между парадигматическими классами,

– потенциально неограниченный объем парадигмы,

– нежесткое распределение лексики по грамматическим классам и частям речи.

Так в работе [17] описано, что для работы с агглютинативными языками в UniParser используется комбинирование подходов в

виде разбиения парадигмы на несколько подпарадигм. Например, если в языке к основе существительного могут последовательно присоединяться аффиксы множественности, падежа и посессивности, то формат UniParser даёт возможность описать отдельно эти три парадигмы и поставить ссылки с одной на другую, указав тем самым, что после аффикса множественности в словоформе должен следовать посессивный, а после любого посессивного – падежный.

Языки, входящие в тюркскую группу, структурно достаточно близки между собой, и реализация программ морфологического анализа и синтеза для одного из тюркских языков позволит разработать технологии, которые без существенных изменений могут быть использованы для других языков тюркской группы.

Наиболее подходящими для агглютинативных языков, к которым относятся все тюркские языки, являются морфологические анализаторы на основе автоматного подхода. Анализаторы на базе автоматного подхода представляют собой конечные преобразователи (finite state transducer – FST) – взвешенный конечный преобразователь (англ. *weighted finite state transducer*, сокр. англ. *WFST*) – вариант конечного автомата, которые кроме распознавания входной последовательности, также формируют выходную последовательность.

К конечным преобразователям подобного типа относятся PC-KIMMO – морфологический анализатор, в котором фонологические преобразования описываются с помощью FST, Helsinki Finite State Transducer (HFST), Juiser и др.

Суть всех этих анализаторов в том, они разработаны с использованием грамматики порядков и в своей работе используют правила следования морфлогических единиц в словоформе. Разница между ними заключается в том, правила следования каких единиц используются:

1. правила следования морфем и последующего выбора необходимого алломорфа;

2. правила следования алломорфов.

Так, в работах [8-12] используются правила первого вида, они по сравнению со вторым типом имеют более компактную базу данных правил следования, но выполняются в два этапа.

Также различаются направления анализа словоформы, слева направо или справа налево. Направление анализа показывает с какого конца словоформы идет отсечение языковых единиц – алломорфов. Так в алгоритме анализа, разработанного Ф. Крыловым [8], анализ словоформы идет справа налево. Сначала программа ищет в словаре основ целую словоформу. Если ее там не оказывается, парсер ищет с правого конца словоформы словоизменительный формант и, если обнаруживается последовательность символов, похожая на аффикс из базы, она отрезается и проходит проверку на возможность следования непосредственно за основой, а левая часть снова сравнивается со словарем основ. При положительном результате парсер предлагает для такой словоформы вариант анализа. При отрицательном результате программа снова обращается к правому концу словоформы и ищет следующий формант, сравнивая его с базой аффиксов. Так продолжается до тех пор, пока оставшаяся слева часть словоформы не совпадет со словом из словаря основ.

Второй вариант анализа, когда в словаре основ сначала находятся все основы, которые могут полностью входить в искомую словоформу. Самый левый символ, найденной основы из базы основ, должен совпадать с самым левым символом искомой словоформы.

Например:

Искомая словоформа на татарском языке – *каргаган*

Найденные основы:

Основа 1 – *кар.* Правая часть – гаган

Основа 2 – *карга*. Правая часть – ган.

После этого проверяется, могут-ли правые части представлять собой цепочку алломорфов в соответствии с правилами следования, представленными в базе данных.

Этот второй подход использован в программе морфологического анализа и синтеза, разработанной на базе структурно-функциональной модели тюркской морфемы (Рис.1.). Данная программа является модулем морфологического анализа, встроенного в многофункциональный, многоязычный Интернет сервис. Программа работает с несколькими тюркскими языками: татарским, казахским, кыргызским, узбекским, турецким и крымско-татарским.

1. Модуль морфологического анализа в составе ModMorph II

Модуль морфологического анализа в составе ModMorph II является одним из сервисов Многофункционального, многоязычного Интернет сервиса, реализованного на базе структурно-функциональной модели тюркской морфемы. Многофункциональный многоязычный сервис доступен по адресу tmm.turklang.tatar. Интерфейс сервиса морфологического анализа представлен на рис.1.

Представленный морфологический анализатор полностью использует базу данных многофункционального многоязычного Интернет сервиса, соответственно структура данных и система обозначений совпадают с представленной в структурно-функциональной модели тюркской морфемы. В настоящее время модуль морфологического анализа может анализировать тексты на татарском, казахском, кыргызском, крымскотатарском, узбекском и турецком языках, поскольку в базе данных сервиса заполнена информация на базе структурно-функциональной модели тюркской модели для перечисленных языков.

2. Алгоритм работы морфологического анализатора

Первый этап – поиск возможных основ.

Слово, поступившее на вход морфологического анализатора разбивается на части, которые гипотетически могут соответствовать основам в словаре.

Например, для словоформы *кешеләргә ‘людям’* будут образованы следующие возможные варианты основ: ‘к’, ‘ке’, ‘кеш’, ‘кеше’, ‘кешел’, ‘кешелә’, ‘кешеләр’ .

В конфигурации программы можно задавать максимально возможную длину основы, чтобы уменьшить число сравниваемых вариантов. После этого производится поиск гипотетических основ в словаре.

Для найденных основ определяются их морфонологические типы и производится анализ правых частей, полученных в результате отсечения основ, найденных в словаре.

Так в предыдущем примере найдена основа *кеше ‘человек’* с морфонологическим типом N02 и получена правая часть *ләргә*.

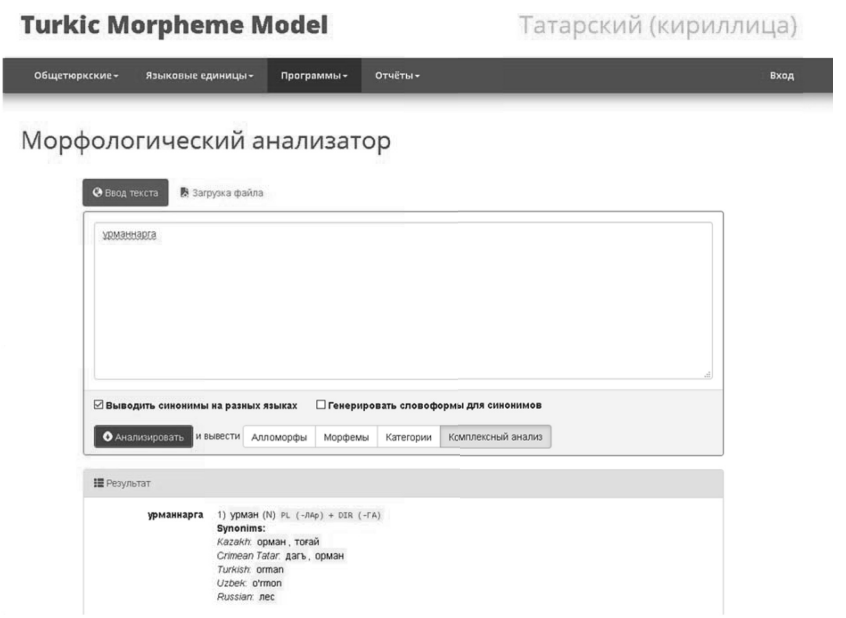


Рис. 1. Интерфейс многоязычного морфологического анализатора

Из оставшейся правой части словоформы образуется последовательность: ‘л’, ‘лә’, ‘ләр’, ‘ләрг’ и ‘ләргә’, для каждого из них ищется совпадающий с ним алломорф во фрагменте правил следования алломорфов морфонологического типа N02. В анализируемом примере найден алломорф ‘ләр’, который отсекается от правой части и остается правая часть ‘гә’. Дальше по тому же алгоритму анализируется правая часть ‘гә’ и так далее до тех пор, пока не окончится остаточная часть слова.

Из полученной информации формируется результат анализа. Формат результата зависит от параметров, заданных при запуске анализатора. Можно задать следующие форматы результатов анализа:

1. Только алломорфы

бүләкләргә

1) бүләк (N) ләр + гә

2. Только морфемы:

урмандагыларга

1) урман (N) -ДАгЫ + -ЛАр + -ГА

3. Только морфологические категории:
урмандагыларга
1) урман (N) ATTR_LOC + PL + DIR
4. Комплексный вывод (морфемы и категории):
урмандагыларга
1) урман (N) ATTR_LOC (-ДАгЫ) + PL (-ЛАр) + DIR (-ГА)

В примере использованы обозначения морфологических категорий, которые выражаются выделенными из словоформы морфемами. Например, PL – обозначение морфологической категории множественности, в татарском языке эта категория выражается с помощью аффикса – *ЛАр*.

DIR – обозначение морфологической категории, называемой директивом, в татарской грамматике эта морфологическая категория называется направительным падежом. Для выражения этой морфологической категории в тексте в татарском языке используется аффикс – *ГА*.

Такая система обозначений морфологических категорий отличается от системы обозначений морфологических категорий русского языка, поскольку для тюркских языков свойственно то, что для обозначения одной морфологической категории используется, как правило, одна аффиксальная морфема.

Из описанных действий алгоритма анализа видно, что он достаточно простой и представляет собой несколько операторов на языке программирования запросов к базе данных. По аналогии с сервисом реализована локальная версия морфологического анализатора тюркских словоформ. Внешний вид локальной версии представлен на рис.2.

Программа морфологического анализа работает и в режиме синтеза. На рис. 2 показаны результаты анализа текста на крымскотатарском языке и синтеза синонимичных словоформ на татарском, казахском, кыргызском, узбекском и турецком языках.

Например:
‹къуйрукълы›
1) N(къуйрукъ) + ATTR_MUN (-лЫ)
Synonims:
Kyrgyz:
куйруктүү: N(куйрук) + ATTR_MUN (-ЛУУ),
дүмдүү: N(дүм) + ATTR_MUN (-ЛУУ)

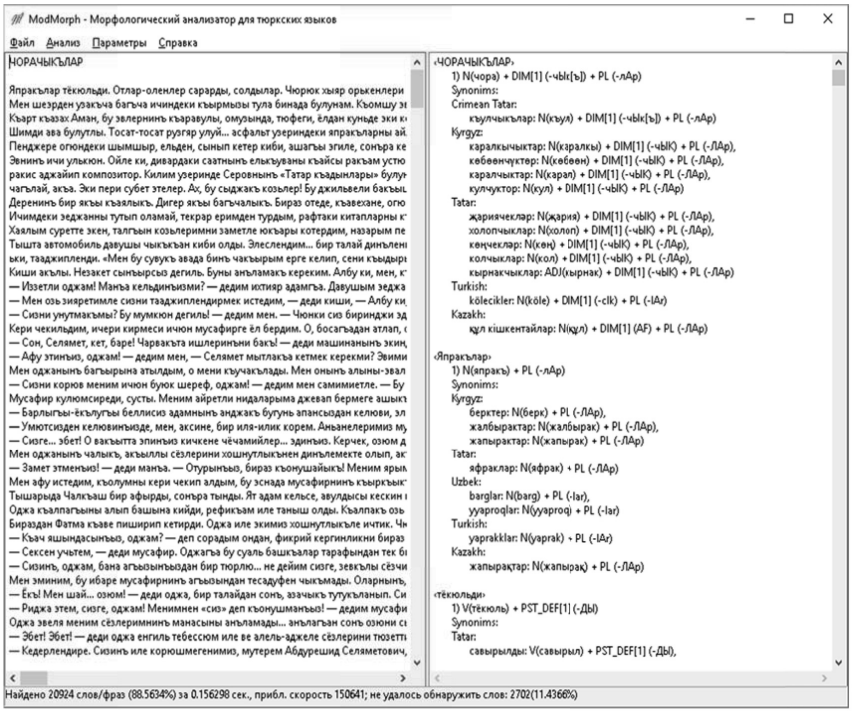


Рис. 2. Локальная версия морфоанализатора с текстом на крымскотатарском языке

Uzbek:

dumli: N(dum) + ATTR_MUN (-li),
ketli: N(ket) + ATTR_MUN (-li)

Turkish:

kuyruklu: N(kuyruk) + ATTR_MUN (-lI)

Kazakh:

құйрықты: N(құйрық) + ATTR_MUN (-ЛЫ)

Рассмотрим информацию, представленную в структурно-функциональной модели тюркской морфемы, которая используется в программе морфологического анализа тюркских словоформ.

3. Морфологические категории

Структурно-функциональная модель тюркской морфемы содержит, как части, которые являются общими для всех тюркских языков, так и части, специфические для каждого языка. Так, названия морфологических категорий и тэгов для их обозначения являются общими для всех тюркских языков, поэтому они находятся в общетюркской части модели (Рис. 3).

Информация о морфологических тэгах используется морфологическим анализатором при выводе результатов анализа. Эти обозначения могут быть использованы для аннотирования электронных корпусов текстов на тюркских языках.

Для всех морфологических категорий в общей части БД хранятся общетипологические названия категорий и тэгов, а в файлах с описанием национальных языковых единиц хранятся названия морфологических категорий на национальных языках.

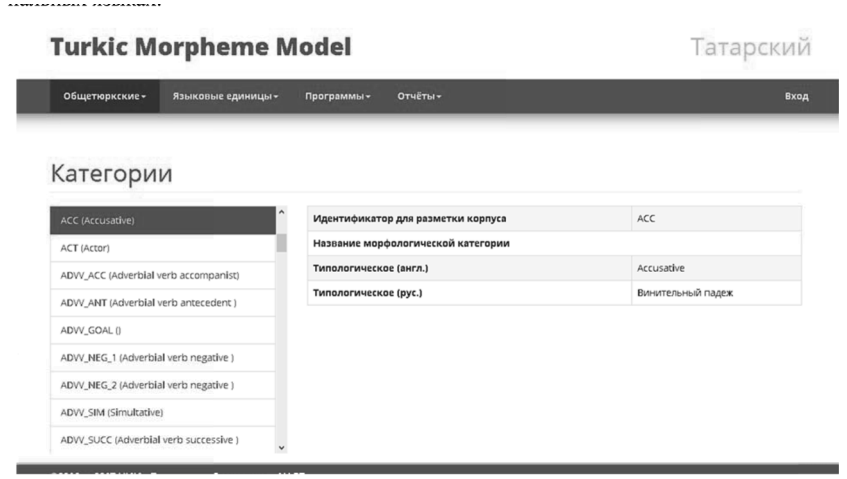


Рис. 3. Морфологические категории

4. Определение возможных основ

В структурно-функциональной модели тюркской морфемы для каждого из тюркского языков описываются две взаимосвязанные подмодели:

– модель аффиксальной морфемы;
– модель корневой морфемы.

В программах морфологического анализа со словарем, используется словарь основ. В данной программе таким словарем основ является подмодель корневой морфемы. Для обеспечения работы морфологического анализатора необходима лишь часть параметров, которые представлены в подмодели корневой морфемы. Эта часть параметров подмодели корневой морфемы, необходимая для работы морфологического анализатора представлена в таблице 1.

Таблица 1

ID	Лемма	Псевдооснова	Часть речи	Морфонологический тип
24709	сапла	сапл	V	V01
24710	саплам	саплам	N	N07
24711	сапламла	сапламл	V	V01
24712	саплан	саплан	V	V05
24713	саплаяк	саплая	N	N11

Рассмотрим поля в таблице 1. Каждая корневая морфема имеет свой ID, который фактически является номером концепта, с помощью этого ID обеспечивается связь с другими корневыми морфемами на разных тюркских языках, которые также связаны с этим концептом. Морфологический анализатор выдает ID корневой морфемы, как выходной параметр, что позволяет использовать результаты морфологического анализа в других программных модулях.

Каждая корневая морфема в базе данных модели имеет две формы: лемму и псевдооснову. Лемма – это словарная форма корневой морфемы. Программа морфологического анализа в качестве результата выдает лемму. А псевдооснова – это максимальная неизменяемая часть корневой морфемы, которая необходима для работы самой программы морфологического анализа. Это сделано с той целью, чтобы в процессе морфологического анализа и синтеза не проверять правила чередования типа б-п, г-к и разные исключения в языке типа *халык ‘народ’ – халкым ‘мой народ’*, *курык ‘бойся’ – курку ‘боязнь’*. При таком способе представле-

ния для леммы *халык* в базе хранится псевдооснова *хал* и уже к ней присоединяются -ык или -к в зависимости от присоединяемых алломорфов. Для каждого типа исключения в базе данных также определен свой морфонологический тип (Таблица 1). Благодаря этому нет необходимости хранить отдельную базу с правилами чередований и базу с исключениями.

Отдельно следует остановиться на частях речи. В русском языке часть речи определяет типы аффиксов, присоединяемых к основам этой части речи. В тюркских же языках, классические части речи, которые приписываются основам в лексических словарях не определяют однозначно наборы присоединяемых аффиксов. Для классификации тюркских основ в зависимости от присоединяемых к ним аффиксальных морфем вводится понятие грамматического класса.

Так, авторами Дыбо, Шеймович в работе [8] в хакасском языке выделяется три основных грамматических класса: имена, глаголы и неизменяемые (частицы, послелоги, союзы и т.п.).

В морфологическом анализаторе, построенном на базе структурно-функциональной модели тюркской морфемы, выделяются следующие грамматические классы:

– имя (N),
– глагол (V),
– признак (A),
– числительное (D),
– неизменяемая часть речи (S).

Такое разделение на морфологические типы обусловлено только правилами морфотактики и не затрагивает синтаксических и семантических особенностей этих морфологических типов. На выходе морфологический анализатор выдает, как грамматический класс основы, так и его классическую часть речи, указанную в лексикографическом словаре. Такое дублирование сделано, чтобы было удобно филологам, которые еще не привыкли к новой системе обозначений. Хотя в работе [8] отмечается, что дифференциация между грамматическими классами выражена слабо, особенно между разрядами имени: слово может трактоваться как существительное, прилагательное или наречие в зависимости от его синтаксической функции:

• существительное может выступать в роли определения: ***таш*** *йортлар* ‘каменные дома’;

• прилагательное может выполнять в предложении любую функцию: *олыга урын бир* ‘старшему место уступи’.

Также из группы имен выделяется подтип имен собственных (PN). Имена собственные задаются отдельными словарями: имен, фамилий, географических названий и др.

В поле ‘Части речи‘ определяются классические части речи, которые присвоены этим основам в словаре. Правила присоединения аффиксальных алломорфов задаются в поле ‘Морфонологический тип’.

Интерфейс таблицы для заполнения морфонологических типов татарского языка представлен на рис. 4, а казахского языка – на рис.5. В левой части страницы представлен список самих морфонологических классов, а в правой части – таблица с указанием алломорфа и дополнительных символов.

На рисунке 4 представлено описание морфонологического типа N11 с леммой *балык ‘рыба’* и с псевдоосновой *балы*. Для каждой морфемы указывается присоединяемый алломорф и в случае наличия промежуточный символ.

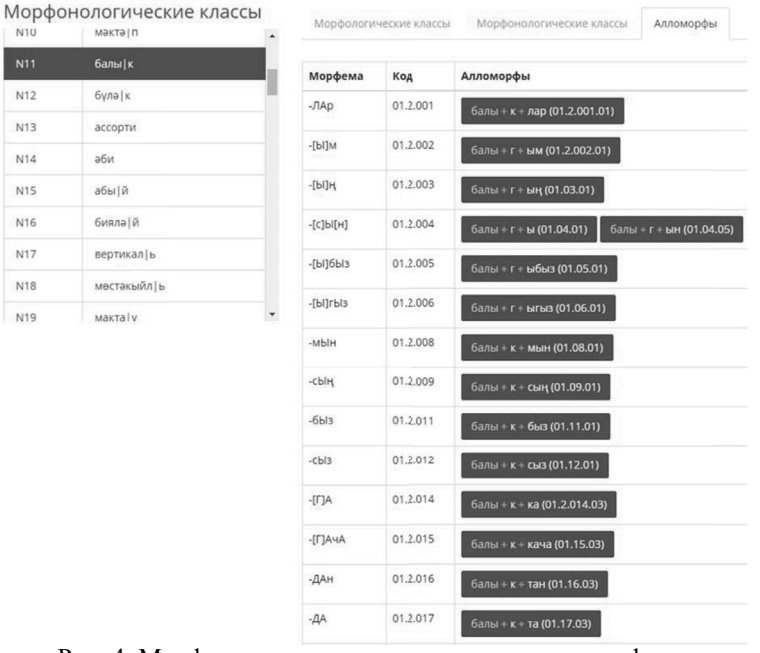


Рис. 4. Морфонологические классы корневых морфем

На рисунке 5 приведен пример морфонологического типа, леммы которых имеют мягкий знак на конце. На рисунке видно, что при присоединении некоторых алломорфов промежуточный символ в виде мягкого знака присутствует, а других отсутствует.

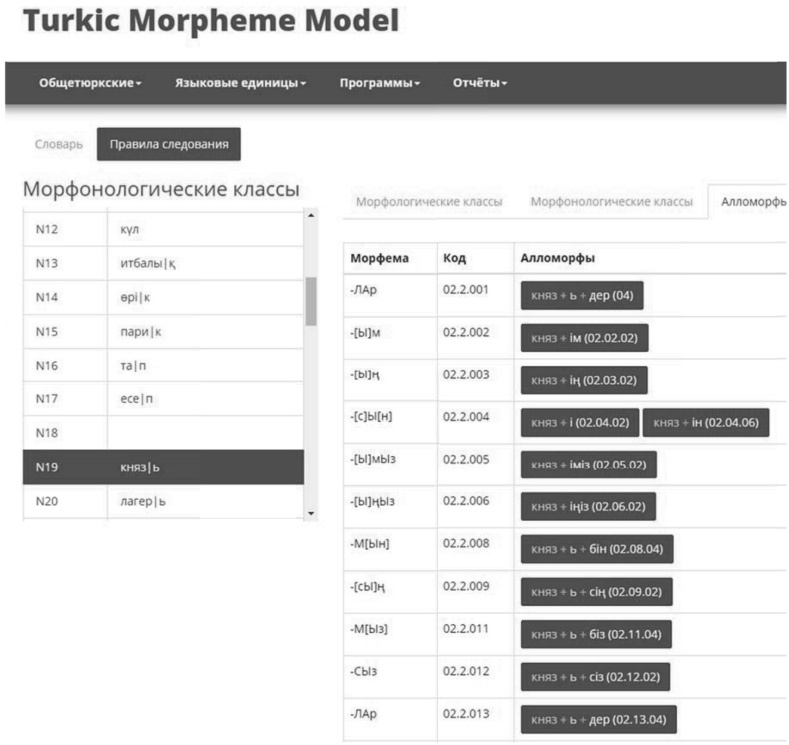


Рис. 5. Морфонологические классы основ казахского языка

Одна из причин выбора такого способа представления, в том что не всегда только на основе правил в тюркских языках можно однозначно определить какой алломорф использовать, поэтому эту информацию приходится хранить в словаре основ. Например, в татарском языке у инфинитива –[Ы]РГА не всегда с помощью морфотактических правил можно однозначно определить какой алломорф использовать -ырга или -арга.

Также, например, нерегулярными являются формы залогов глаголов с аффиксами -Ыл и -Ын. Информацию о возможности присоединения которых тоже следует хранить в словаре основ.

Например:

туклан ‘питаться’ – нет формы тукла, а у туплан ‘собираться’ – есть основа тупла.

5. Проверка цепочек алломорфов

В программе морфологического анализа после того, как выделены возможные основы и отсечены правые части словоформ с указанием их морфонологического типа, необходимо проверить правильность и структуру оставшейся после отсечения основы правой части словоформы.

Для проверки правой части необходимо использовать правила образования цепочек алломорфов. Эта информация представлена в разделе ‘Правила следования‘ подмодели аффиксальной морфемы (Рис.6.).

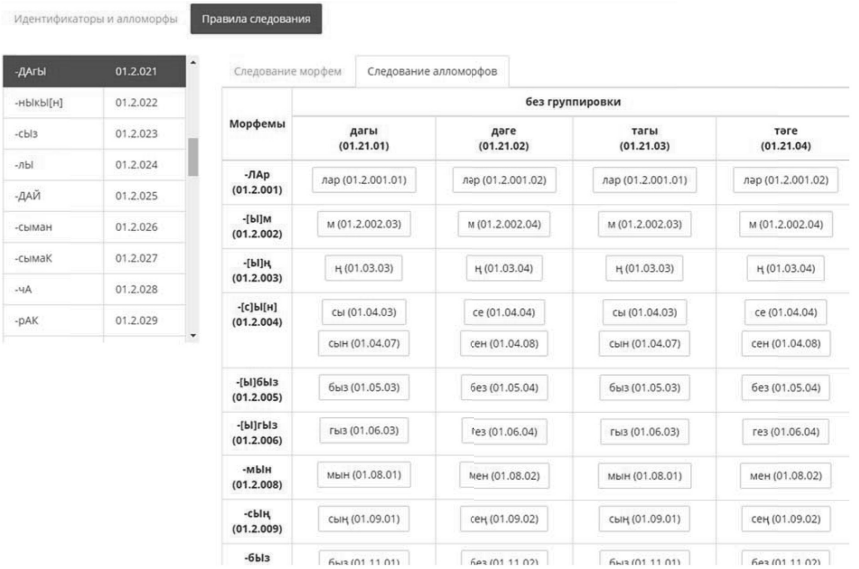


Рис. 6. Правила следования алломорфов

В левой части окна представлен список всех аффиксальных морфем текущего языка, информация о которых имеется на данный момент в базе данных (Рис.6). При выборе одной морфемы из списка в правой части окна открывается таблица с правилами следования для алломорфов этой морфемы. В верхней части таблицы приводится список всех алломорфов выбранной морфемы, а в колонке под этим алломорфом задается список тех алломорфов, которые в словоформе могут следовать справа от данного алломорфа. В примере на рис.6 видно, что за алломорфом -дагы в словоформе могут следовать алломорфы: -лар, -м, -ң, -сы, -сын, -быз и др., обрзуя цепочки -дагылар, -дагым, -дагың, -дагысы, -дагысын, дагыбыз и др.

6. Комбинированный вариант анализа

Морфологический анализатор реализован как один из сервисов многофункционального многоязычного Интернет сервиса, так и в локальной версии. Обе версии позволяют работать как с отдельными словоформами, так и с большими текстами.

Первый вариант Интернет сервиса был реализован только на автоматном подходе, однако из-за потерь скорости при пересылке данных по сети, скорость работы анализатора с обращением к базам данных, находящимися на сервере, была неудовлетворительной. В итоге было принято решение об использовании комбинированного подхода. Комбинированный подход предполагал использование для небольшого фрагмента из наиболее частотных словоформ готовых результатов анализа. Наиболее частотные словоформы были закэшированы в базе данных. В систему были внесены следующие изменения: сначала производится поиск словоформ в базе часто встречающихся форм слов словаря, и только в случае, если форма не найдена, производится анализ этого слова. Таким образом, в базу данных были добавлены таблицы часто встречающихся словоформ и морфологических структур соответствующих словоформ.

Наиболее распространенные формы слов были выбраны на основе автоматического анализа небольшого корпуса текстов. В результате подобных изменений скорость работы программы выросла в 10–12 раз. Если использовать только распознавание часто

встречающихся словоформ, то скорость увеличивается до 18 раз, но при этом количество проанализированных слов снижается.

7. Морфологический анализатор без словаря основ

В текстах на тюркских языках встречается много слов, которых нет в базовых лексических словарях. Это различные имена собственные или заимствования из других языков. Строить для подобных словоформ отдельные словари нецелесообразно, поскольку такие слова могут встретиться лишь однажды в одном конкретном тексте. При решении некоторых задач требуется определять лишь категориальную принадлежность основы словоформы, а также набор его аффиксов. Для этих целей был реализован вариант морфологического анализа тюркской словоформы без использования словаря основ.

В результате работы морфологического анализатора без использования словаря основ, количество варинатов анализа будет намного больше, чем при работе со словарем. При выдаче результатов анализа программа автоматически производит ранжирование выдаваемых варианов анализа. Результатами с наибольшим рангом будут цепочки аффиксальных морфем с наибольшим чилом аффиксов.

Например, получая на вход словоформу «*Иванныкыларга*», программа морфологического анализа без использовния словаря основ выдаст следующие варианты анализа:

N (*Иван*) – нЫкЫ – Лар – ГА ранг 3
N (*Иванныкы*) – Лар – ГА ранг 2
N (*Иванныкылар*) – ГА ранг 1
V(*Иванныкыла*) – ЫРГА ранг 1

Для реализации анализатора использована уже существующая база данных без словаря основ.

В модуле морфологического анализа без поиска основ реализован алгоритм обратного поиска. Рассмотрим тот же пример: словоформа «кешеләргә». Сначала производится создание списка возможных вариантов аффиксов с конца слова: «ә», «гә», «ргә», «әргә» и т.д. Количество вариантов ограничено максимально возможной длиной аффикса, задающейся в конфигурации. Поиск возможных вариантов аффиксов ведется в таблице аффиксов-алломорфов. Затем для каждого найденного аффикса проводит-

ся проверка оставшейся части слова с учетом дальнейших возможных аффиксов (на основе таблицы взаимосвязей аффиксов). В результате обязательно должна остаться какая-то часть слова, которая не была распознана (поскольку не бывает слов без основ). После того, как остается основа, исходя из найденных сочетаний аффиксов мы можем определить, какая часть речи и какое склонение может соответствовать каждому возможному случаю распознавания. Исходя из возможных склонений слов, мы можем восстановить возможные основы слов. Например, для словоформы «китабыма» (рус. «для моей книги») в качестве основы система предложит как вариант «китаб», так и вариант «китап». Второй вариант является правильным, однако система не может этого знать, поскольку не использует словарь основ.

В качестве дополнительного элемента уточнения результата мы также выполняем сравнение окончания основы с возможными окончаниями основ данного склонения. Например, для словоформы «этапта», основой которой является слово «этап», проверяется, что склонение, соответствующее аффиксу «-та», применимо лишь к словам, завершающимся на глухой согласный звук, в противном случае аффикс должен быть «-да». Таким образом, для записи «этапда» основа «этап» будет невозможной.

Для повышения эффективности работы данного модуля может потребоваться дальнейшее уточнение возможных окончаний основ склонений, а также кэширование в оперативной памяти наиболее распространенных сочетаний аффиксов с применением оптимизированных для данной системы индексов.

8. Расширенный морфологический анализатор с аналитическими формами

В тюркских языках есть несколько видов аналитических форм, таких как послеложные формы или аналитические формы глаголов, образованные с помощью вспомогательных форм глагола. Эти формы тоже являются морфологическими формами, поэтому морфологический анализатор был доработан до возможности анализа аналитических форм тюркских языков.

Например:

Анализ послеложной формы, образованной с помощью послелога кадәр.

‹урманга кадәр›

1) PP-AF(урман) + DIR_LIM (DIR (-[Г]А); POST‹кадәр›)

Для работы расширенного морфологического анализатора используется информация, представленная в разделе модели ‘Аналитические формы’. В этом разделе описываются правила извлечения аналитических конструкций, основанные на наборе грамматических тэгов и корневых морфем с учетом порядка расположения компонентов и возможности/невозможности вставки внешних элементов между компонентами конструкции. В частности, база для представления двухкомпонентных глагольных аналитических форм имеет следующую структуру: первый компонент имеет произвольную (для большей части форм) глагольную основу с обязательным аффиксом (набором аффиксов), описываемым специальной формулой, и грамматически является относительно инвариантной единицей), второй компонент, как правило, может иметь все словоизменительные аффиксы, допустимые для глагольных единиц.

В таблице 1 представлена структура раздела ‘Аналитические формы’ струтурно-функциональной модели тюркской морфемы.

Таблица 2. **Структура параметра аналитические формы**

Слово	**Тип**	**Ор.**	**Основа слева**	**Аффикс слева**	**Основа справа**	**Only**	**Категория**	**Код**
сәбәпле	PP-AF	R	V	VN[1]	ALL	1	MOTIV	507
ал	V-AF	R	ALL	PRES[1]	ALL	1	POT	253
ал	V-AF	R	ALL	ADVV_ACC[1]	ALL	1	MOMENT	528
алып	PP-AF	R	ALL	ABL	ALL	1	EGRES	512

Таблица 2 дает общее представление о структуре и характере разметки аналитических форм, которые образуются при помощи недостаточного глагола *иде* в функции вспомогательного.

Таблица 1. **Примеры разметки АК**

Структура	Пример	Перевод	Тэг АФ
V+PRES_3SG+Phase_V(башла)	Җырлый башлау	начать петь	INCHOAT_1
V+ADVV +Phase_V(бет)	Агып бетү	вытечь	COMPLET_2
OBL +Vf(кил)	Җырлыйсы килү	хотеть петь	DESID_2
PRES_3SG+Vf(бел)	Ясый белү	уметь делать	CAPAC_1

ЛИТЕРАТУРА

1. Халитова Н. А., Закирова Р. А., Гимадудтинова Р.У. Морфологический анализ при машинном переводе с татарского языка на русский / Вероятностные методы и кибернетика I, Сборник работ НИИММ им. Н. Г. Чеботарева при Казанском университете, Учен. зап. Казан. ун-та, 122, № 4, Изд-во Казанского ун-та, Казань, 1962, 98–105.

2. Махмудов, Масуд Ахмед оглы. Разработка системы формального морфологического анализа тюркской словоформы: на материале азербайджанского языка: диссертация ... кандидата филологических наук : 10.02.06. – Баку, 1982.

3. Проблемы моделирования тюркской морфологии : (Аспект порождения кирг. имен. словоформы) / Т. Садыков; АН КиргССР, Ин-т яз. и лит. – Фрунзе : Илим, 1987. – 120 с.

4. Тузов В.А. Морфологический анализатор русского языка //Вестник СПбГУ, сер. 1. 1996. Вып. 1 (N15). С. 41–45.

5. Сегалович И., Маслов М. Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог’99». Т. 2. С. 547–552. Казань, 1998.

6. Antworth, E.L. PC-KIMMO: a two-level processor for morphological analysis. Occasional Publications in Academic Computing No. 16. Dallas: Summer Institute of Linguistics, 1990, 273 p.

7. Kemal Oflazer. Two-level Description of Turkish Morphology. Literary and Linguistic Computing, – Vol. 9, No 2, – 1994.

8. Дыбо А.В., Шеймович А.В. Автоматический морфологический анализ для корпусов тюркских языков // Филология и культура. – 2014. – №2.

9. Сулейманов Д.Ш., Гильмуллин А.А., Гильмуллин Р.А. База морфотактических правил для татарского глагола как основа двухуровневого морфологического анализатора // Сборник трудов Международного семинара «Диалог», Казань, 1998. – С. 597–609.

10. Сулейманов Д.Ш., Гатиатуллин А.Р. К разработке Лемматора татарских словоформ // В сб. Трудов Международного семинара Диалог-99 по компьютерной лингвистике и ее приложениям в двух томах (Таруса, 31 мая – 4 июня 1999 г.). Т. 2. – Таруса, 1999. – С. 306–314.

11. Сулейманов Д.Ш. Реализация татарской морфологии (на англ. яз.) // In Iss.: Computational linguistics and intelligent text processing: third international conference; proceedings / CICLing 2002, Mexico, February 17–23, 2002. Alexandr Gelbukh (ed.). – Berlin: Springer, 2002 (Lecture notes in computer science; Vol. 2276). – P. 327–329.

12. Сулейманов Д.Ш., Гатиатуллин А.Р. Структурно-функциональная компьютерная модель татарских морфем. – Казань, Изд-во “Фэн”, 2003. – 220 с.

13. Сиразитдинов З. А. Алгоритмическая грамматика словоизменения башкирского языка // [Электронный ресурс]. URL: http://mfbl.ru/bashdb/algram/algram.htm (дата обращения: 19.09.2015).

14. Орехов Б. В., Слободян Е. А. Проблемы автоматической морфологии агглютинативных языков и парсер башкирского языка // Информационные технологии и письменное наследие: материалы международной научной конференции (Уфа, 28–31 октября 2010 г.) / отв. ред. В. А. Баранов. Уфа; Ижевск: Вагант, 2010. С. 167–171.

15. Шарипбаев А.А., Бекманова Г.Т., Ергеш Б.Ж., Бурибаева А.К., Карабалаева М.Х. Интеллектуальный морфологический анализатор, основанный на семантических сетях // Материалы международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем» (OSTIS-2012). Минск, БГУИР, 16–18 февраля 2012 г. С. 397–400.

16. Желтов П.В. Морфологический анализатор чувашского языка. Материалы Международной конференции студентов и аспирантов по фундаментальным наукам «Ломоносов 2002», М., 2002.

17. Архангельский Т.А. Принципы построения морфологического парсера для разноструктурных языков // диссертация ... кандидата филологических наук : 10.02.21. – Москва, 2012.

УДК 81’33

THE TASK OF THE MORPHOLOGICAL ANALYZER (MA) OF THE CHUVASH LANGUAGE

P. Zheltov
Federal state budget educational institution of higher education «Chuvash State University named after I.N. Ulyanov», Cheboksary
chnk@mail.ru

The task of the morphological analyzer (MA) of the Chuvash language is the establishment of the morphemic composition of words, as well as the morphological features used in the tasks of the syntactic and semantic analyzers. The knowledge base of morphological analysis of the Chuvash language consists of a list of enumerated types (describing the morphological features), dictionaries (describing the meaning of morphemes), reference books (describing the relationship of morphemes and attributes), the structure of the working data model (for storing texts, words and word parsing options) and meta-rules (Describing the sequence of parsing the words of different parts of speech). The developed model is language-independent and can be applied for Russian and for the Chuvash language. The practical result of the program is given. In all cases, in the presence of morphemes in the dictionary, the program produced an accurate analysis with the correct definition of all attributes.

Keywords: morphological analysis, computer linguistics, morphological analyzer, Chuvash language, text analysis.

РАЗРАБОТКА МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА ЧУВАШСКОГО ЯЗЫКА

П.В. Желтов
Чебоксарский госуниверситет им. И.Н. Ульянова
chnk@mail.ru

Задачей морфологического анализатора (МА) чувашского языка является установление морфемного состава слов, а также морфологических признаков, используемых в задачах синтаксического и семантического анализаторов. База знаний морфологического анализа чувашского языка состоит из списка перечислимых типов (описывающих морфологические признаки), словарей (описывающих значение морфем), справочников (описывающих взаимосвязь морфем и признаков), структуры рабочей модели данных (для хранения текстов, слов и вариантов разбора слов) и метаправил (описывающих последовательность разбора слов разных частей речи). Разработанная модель является языконезависимой и может быть применена для русского и для чувашского языка. Приведен практический результат работы програм-

мы. Во всех случаях при наличии морфем в словаре программа производила точный анализ с правильным определением всех атрибутов.

Ключевые слова: морфологический анализ, компьютерная лингвистика, морфологический анализатор, чувашский язык, анализ текста.

Введение

Задачей морфологического анализатора (МА) чувашского языка является установление морфемного состава слов, а также морфологических признаков, используемых в задачах синтаксического и семантического анализаторов.

База знаний морфологического анализа чувашского языка состоит из списка перечислимых типов (описывающих морфологические признаки), словарей (описывающих значение морфем), справочников (описывающих взаимосвязь морфем и признаков), структуры рабочей модели данных (для хранения текстов, слов и вариантов разбора слов) и метаправил (описывающих последовательность разбора слов разных частей речи). В чувашском языке принято выделять следующие морфологические характеристики (табл. 1):

Таблица 1

Часть речи	Существительное, прилагательное, числительное, местоимение, глагол, наречие, частица, союз, предлог, междометие, подражательное слово
Одушевленность	Одушевленное, неодушевленное
Число	Единственное, множественное
Падеж	Основной, притяжательный, дательно-винительный, местный, исходный, творительный, лишительный, причинно – целевой
Лицо	первое лицо, второе лицо, третье лицо
Форма	Уподобительная, усиления, предельная, разделительная, образа действия, уподобления, сравнения, выделительная, достаточности, направительная, частотности, обладания, ласкательная, выделения, вероятностная, порядка, принадлежности, сравнения
Аспект	Утвердительный, отрицательный, возможности.
Время	Настоящее, будущее, прошедшее

Распределение морфологических характеристик по частям речи приведено в табл. 2:

Таблица 2

Существительное	Лицо, число, падеж, форма (1-17), одушевленность, время
Прилагательное	Лицо, число, падеж, форма (5, 12, 10, 3, 8, 17, 18), время
Числительное	Репрезентация, форма (4, 5, 10, 12, 10, 14, 2, 17), падеж
Местоимение	Репрезентация, форма (8, 14, 1, 2), падеж
Глагол	Репрезентация, вид, аспект, залог, время, падеж(В), форма(9)
Причастие	Вид, аспект, залог, время, форма (7, 1, 9, 11), падеж
Деепричастие	Вид, аспект, залог, время
Союз	Репрезентация, форма
Частица	Значение, форма
Послелог	Значение, форма

1. Структура морфологического анализатора

Проект библиотеки морфологического анализатора состоит из 7 классов (рис.1).

• *MorfParser.cs* – Главный класс МА. Имеет два открытых метода для работы.

• *AffixInfoProvider.cs* – Класс предоставляет набор полей и функций для работы с аффиксами.

• *ContextRules.cs* – Класс отвечает за корректное определение правила восстановления согласно контексту.

• *FeaturesDeterminers.cs* – Класс предоставляет набор функций для определения атрибутов (характеристик) слов.

• *Helper.cs* – Класс содержит набор вспомогательных статических функций.

• *ImprovedAffixHandler* – Класс принимает слово для разбиения его на составляющие и производит поиск в словаре.

• *MorfConstants* – Класс представляет собой набор констант, введенных и использованных в проекте.

Цель работы: проектирование, разработка динамически подключаемой библиотеки (dll), предоставляющей набор методов и функций для морфологического анализа слов чувашского языка

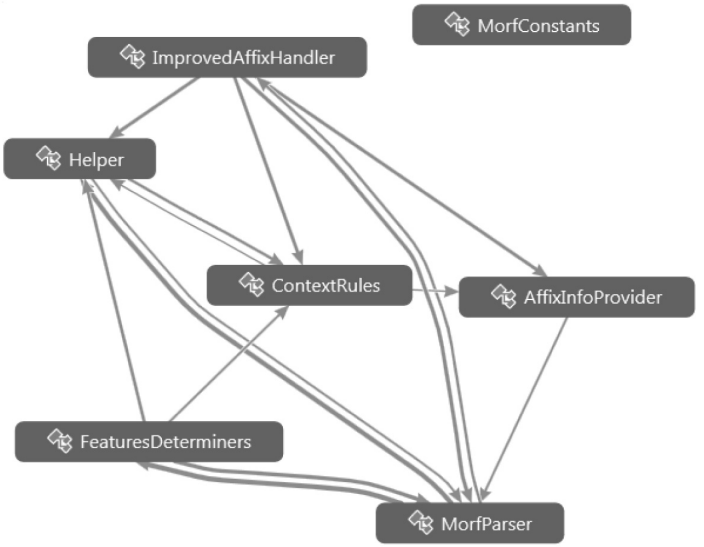


Рис. 1.

для внедрения в системы машинного перевода, лингвистических процессоров. Требования к библиотеке:

1. Определение части речи слова и других морфологических характеристик.
2. Выделение корня и аффиксов.
3. Распознавание контекстов.
4. Анализ слов – исключений.
5. Возможность вывода результатов в файл.
6. Режим логгирования для удобной отладки.

Алгоритм 1 (словесно – пошаговое описание алгоритма решения задачи)

Шаг 1. Начало работы.

Шаг 2. Выполнение анализа предметной области, на основании которого будет разработано техническое задание и составлены критерии определения готовности работы.

Шаг 3. Выполнение анализа инструментов и средств решения задачи, выбор программы и языков создания библиотеки.

Шаг 4. Разработка алгоритмов решения задачи.

Шаг 5. Реализация функционала из технического задания.

Шаг 6. Тестирование всех функций библиотеки.

Шаг 7. Завершение работы.

Структура разработки представляет собой минимальный набор средств для работы приложения. Для использования библиотеки МА требуется платформа . *NET Microsoft*.

2. Словарь и база аффиксов

Структуру чувашских слов можно представить в виде суммы корней и аффиксов. Приставки и окончания, в отличие от русского языка, в чувашском языке отсутствуют, что упрощает разработку МА. Таким образом, для разработки требовался словарь основ (корней) и база аффиксов. Исходный словарь представляет собой текстовый файл, в котором слова представлены следующим образом: слово, часть речи, информация об источнике (рис. 3). В нем собрано более тридцати одной тысячи слов чувашского языка.

```
аббат,noun,c
аббатлăх,noun,z47
аббревиатура,noun,c
абзац,noun,c
абитуриент,noun,c
абонемент,noun,c
абонент,noun,c
абориген,noun,c
аборт,noun,c
абортла,verb,c
абразив,noun,c
абразивлă,adj,c
абрикос,noun,c
абсолютизм,noun,c
абсолютизмла,adv,c
абсолютлă,adj,c
абсорбциле,verb,t28
абстрактлă,adj,c
абстрактлăх,noun,c
абстракци,noun,c
абстракциле,verb,c
абстракцилен,verb,c
абстракциллĕ,adj,c
абстракционизм,noun,c
абстракционизмла,adv,c
абстракционист,noun,c
```

Рис. 3. Фрагмент словаря основ

Аффиксы играют решающую роль и несут основную словообразовательную нагрузку. Их в чувашском языке около 170. Исходная база аффиксов имела схожую структуру со словарем (рис. 4). В целях оптимизации поиска, анализа конструкция базы аффиксов сформирована согласно следующим принципам.

```
и о
ин о
ине о
инче о
ипе о
исĕр о
исем/Р о
ишĕн о
ăсем/Р о
ăн о
ăна о
ăра о
ăран о
ăпа о
ăпалан о
ăсăр о
ăшăн о
ăри о
```

Рис. 4. Фрагмент исходной базы аффиксов

Представление порядка следования аффиксов и их сочетаемости между собой является актуальной задачей в системах машинного перевода и лингвопроцессорах. Аффиксы играют решающую роль и несут основную словообразовательную и словоизменительную нагрузку в естественном языке. Во многих естественных языках агглютинативного строя часто встречаются довольно длинные аналитические конструкции, содержащие до 7 аффиксов, переводимые на русский с помощью целого предложения. Порядок следования и сочетаемости между собой аффиксов в агглютинативных языках необычайно сложен, часты повторы одних и тех же аффиксов и велико число ветвлений. Рассмотрим пример. Слово «витресемпех», означающее «с ведрами же», раскладывается на следующие составляющие: «витре+сем+пе+х». Извлеченные аффиксы сохранят свой порядок и с другими словами («ачасемпех», «ĕçсемпех»). Из этого можно сделать вывод о возможности представления базы аффиксов в виде совокупности уровней. Где на каждом уровне будут храниться аффиксы, которые могут склеиваться с аффиксами, у которых уровни ниже. На основе ранее извлеченных аффиксов БА бы выглядела следующим образом: на первом уровне «сем», на втором «пе», на третьем «х». Вдобавок, каждый аффикс, помимо признака последовательности, хранит в себе характеристику типа. Согласно этому принципу, каждому

аффиксу свойственна своя палитра частей речи, к которым он может присоединиться. Аффиксы из рассмотренного примера могут соединяться со следующими частями речи:

«Сем» – аффикс множественного числа. Существительные, прилагательные, числительные, местоимения.

«Пе» – аффикс дательного падежа. Существительные, прилагательные, числительные, местоимения.

«Х» – аффикс категории усиления. Существительные, прилагательные, числительные, местоимения, глаголы.

Список подходящих частей речи довольно обширный. Для компактности части речи разделены на условные группы. После анализа выявлены определены типы и которые применены в конечной версии базы аффиксов.

OnlyGlagol – тип, под которым объединены аффиксы глагольного типа, т.е. склеивающиеся только с глаголами. К таким относятся: «ма, ме, маç, меç».

NotGlagol – тип, объединяющий неглагольный класс аффиксов. По большей части это падежные аффиксы: «па, пе, ра, ре, та, те».

Any – общий тип аффиксов, способные соединяться и с глагольными и именными частями речи. Например, аффикс «а» относится и к дательному падежу (çурт -> çурта), так и к деепричастиям (чуп -> чупа).

Формирование базы аффиксов завершается определением одного из трех типов для каждого уровня аффиксов. Некоторые уровни, содержащие пересекающиеся типы, разбиваются до наличия лишь однородных аффиксов на один уровень. БА на данный момент разделен на 37 уровней (рис. 5).

```
ни, Any
хи, NotGlagol
ман, мен, OnlyGlagol
акан, екен, аканни, екенни, манни, менни Any
ăм, ĕм, мĕш, NotGlagol
у, ӳ, и, NotGlagol
ллă, ллĕ, лă, лĕ, лли, ли, NotGlagol
мелли, малли, Any
ри, ти, NotGlagol
масть, маст, мест, мас, мес, OnlyGlagol
м, Any
```

Рис. 5. Фрагмент базы аффиксов по уровням

В целях более удобного внесения изменений, словарь и база аффиксов и другие вспомогательные документы хранятся в виде

текстовых файлов. Для редактирования, как специалисту так и пользователю, в качестве программного обеспечения достаточно стандартного блокнота. Тем не менее, при необходимости БА легко может быть конвертирована в таблицы базы данных в силу своей структуры.

3. Ядро морфологического анализа

Разработку морфологического анализатора можно разделить на 2 этапа. Во-первых, слово в исходной форме ищется в словаре основ. Грамматические характеристики в данном случае определяются по умолчанию в зависимости от части речи. Во втором этапе производится непосредственный анализ слова, разбиение его на пары «корень-аффиксы» и выявление характеристик. Оба этапа возвращают произвольное количество частей речи в зависимости от найденных совпадений. При отсутствии совпадений слово возвращается с «неопределенными» характеристиками (рис. 6). Рассмотрим поподробнее каждый из этих этапов.

Первая, и наиболее простая часть – это прямой поиск входного слова в словаре основ. Эту задачу решает функция *SearchInDictionaries.* Алгоритм метода приводится ниже.

Рис. 6. Алгоритм метода прямого поиска входного слова

Алгоритм 2 (алгоритм поиска совпадений в словаре) (см. рис. 7).

Шаг 1. Отделить от входного слова аффиксы-частицы, такие как «-и», «-çке» и другие: «ăшă-çке» (тепло же), «пысăк-и?» (большой ли). Записать аффиксы в строковой переменной. Если слово «двойное», образованное с помощью повторения (кил-кил), убрать повторяющуюся часть слова.

Шаг 2. По началу слова определить диапазон поиска. Так как словарь отсортированный, непосредственно проход по всему словарю не требуется.

Шаг 3. В выявленном диапазоне производить поиск слова. Вне зависимости от найденных совпадений продолжать поиск до конца.

Шаг 4. Извлечь часть речи найденного слова. На основе него определить остальные грамматические характеристики.

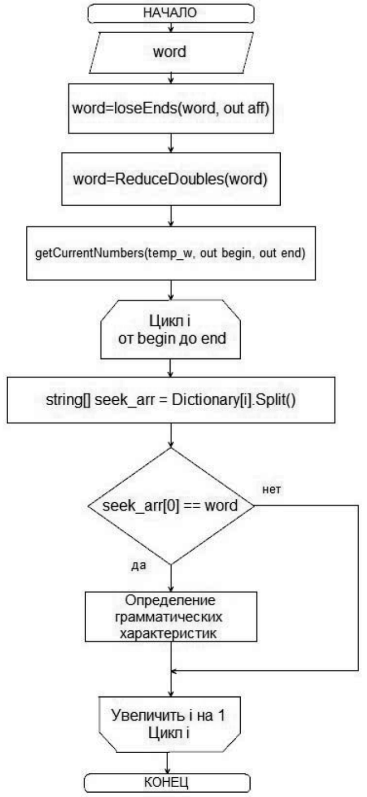


Рис. 7. Алгоритм поиска совпадений в словаре

Определение границ поиска непосредственно основано на рассмотрении всех возможных вариаций пар символов, с которых может начинаться слово (рис. 8). Составив список пар, отсортировав их в алфавитном порядке и сопоставив их со словарем, получается некая карта, где каждой паре слева справа соответствует номер строки, с которой начинаются подобные слова. Однако эти данные не определяются заранее. Так как в МА предусмотрена возможность динамического подключения разных словарей, выявление границ поиска определено в функцию. Таким образом, функция автоматически запускается при подключении нового словаря, обеспечивая анализатор актуальными данными.

```
"аб", "ав", "аг", "ад", "аз", "аи", "ай", "ак", "ал", "ам", "ан", "ао", "ап", "ар",
"ăв", "ăй", "ăл", "ăм", "ăн", "ăп", "ăр", "ăс", "ăç", "ăт", "ăф", "ăх", "ăш",
"ба", "бе", "би", "бл", "бо", "бр", "бу", "бы", "бю",
"ва", "вă", "ве", "вё", "вě", "вз", "ви", "вк", "вл", "во", "вр", "вс", "вт", "ву",
"га", "га", "ге", "гё", "ги", "гл", "гн", "го", "гр", "гу", "гя",
"да", "дв", "де", "дж", "дз", "ди", "дн", "до", "др", "ду", "ды", "дю",
"ев", "ег", "ед", "ей", "ек", "ел", "ем", "ен", "еп", "ер", "ес", "ет", "еф", "ех",
"ёк", "ёл", "ём", "ён", "ёп", "ёр", "ёс", "ёç", "ёт", "ёф", "ёх", "ёш",
"жа", "жг", "же", "жи", "жм", "жн", "жо", "жр", "жу", "жю",
"за", "зв", "зд", "зе", "зн", "зо", "зу", "зэ",
"иб", "ив", "иг", "ид", "ие", "иж", "из", "ий", "ик", "ил", "им", "ин", "ио", "ип",
"йа", "йă", "йё", "йо", "йў", "йы",
"ка", "кă", "кв", "ке", "кё", "ки", "кл", "кн", "ко", "кр", "кс", "ку", "кў", "кх",
"ла", "лă", "ле", "лё", "лě", "ли", "ло", "лу", "лў", "ль", "лю", "ля",
"ма", "мă", "ме", "мё", "ми", "мн", "мо", "мр", "му", "мў", "мы", "мэ", "мя",
"на", "нă", "не", "нё", "ни", "но", "нр", "ну", "нў", "нэ", "ня",
"оа", "об", "ов", "ог", "од", "оз", "ой", "ок", "ол", "ом", "он", "оп", "ор", "ос",
"па", "пă", "пе", "пё", "пи", "пл", "пн", "по", "пр", "пс", "пт", "пу", "пў", "пы",
"ра", "ре", "ри", "ро", "рт", "ру", "ры", "рэ", "рю", "ря",
"са", "сă", "сб", "св", "сд", "се", "сё", "сě", "си", "ск", "сл", "см", "сн", "со",
"ça", "çă", "çв", "çе", "çё", "çи", "çт", "çу", "çў", "çы",
"та", "тă", "тв", "те", "тё", "тě", "ти", "тк", "то", "тп", "тр", "тс", "ту", "тў",
"уа", "уб", "ув", "уг", "уд", "уе", "уз", "уй", "ук", "ул", "ум", "ун", "уп", "ур",
"ўк", "ўл", "ўн", "ўп", "ўр", "ўс", "ўт", "ўх",
```

Рис. 8. Фрагмент базы начальных букв для алгоритма определения границ поиска

Стоит отметить, установление морфологических характеристик для местоимений несколько отличается от других частей речи. Дело в том, что местоимения в начальной форме могут быть и во втором или третьем лице (*эпĕ, эсĕ, вăл*), быть во множественном числе (*вĕсем, эпир*). Для решения этой задачи грамматические характеристики исключительных местоимений занесены в отдельный массив (рис 9). При обнаружении слова в данном

массиве грамматические характеристики для него извлекаются из этого набора, иначе из общих массивов (рис. 10).

```
"эпё," + MorfConstants.ED_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE1,
"эп," + MorfConstants.ED_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE1,
"эсё," + MorfConstants.ED_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE2,
"эс," + MorfConstants.ED_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE2,
"вӑл," + MorfConstants.ED_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE3,
"эпир," + MorfConstants.MN_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE1,
"эсир," + MorfConstants.MN_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE2,
"вӗсем," + MorfConstants.MN_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE3,
"хам," + MorfConstants.ED_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE1,
"ху," + MorfConstants.ED_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE2,
"хӑй," + MorfConstants.ED_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE3,
"хамӑр," + MorfConstants.MN_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE1,
"хӑвӑр," + MorfConstants.MN_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE2,
"хӑв," + MorfConstants.ED_CHISLO + "," + MorfConstants.OSN_P + "," + MorfConstants.FACE2,
"пирӗн," + MorfConstants.MN_CHISLO + "," + MorfConstants.ROD_P + "," + MorfConstants.FACE1,
"сирӗн," + MorfConstants.MN_CHISLO + "," + MorfConstants.ROD_P + "," + MorfConstants.FACE2,
"вӗсен," + MorfConstants.MN_CHISLO + "," + MorfConstants.ROD_P + "," + MorfConstants.FACE3,
"ман," + MorfConstants.ED_CHISLO + "," + MorfConstants.ROD_P + "," + MorfConstants.FACE1,
"сан," + MorfConstants.ED_CHISLO + "," + MorfConstants.ROD_P + "," + MorfConstants.FACE2,
"ун," + MorfConstants.ED_CHISLO + "," + MorfConstants.ROD_P + "," + MorfConstants.FACE3
```

Рис. 9. Фрагмент массива грамматических характеристик исключительных местоимений

```
"noun," + MorfConstants.NOUN + "," + MorfConstants.ED_CHISLO + "," + MorfConstants.NULL + "," + MorfConstants.OSN_P + ","
"verb," + MorfConstants.VERB + "," + MorfConstants.ED_CHISLO + "," + MorfConstants.NAST_V + "," + MorfConstants.NULL + ","
"adj," + MorfConstants.ADJECTIVE + "," + MorfConstants.ED_CHISLO + "," + MorfConstants.NULL + "," + MorfConstants.UNKNOWN
"adv," + MorfConstants.ADVERB + "," + MorfConstants.ED_CHISLO + "," + MorfConstants.NULL + "," + MorfConstants.NULL + ","
"pron," + MorfConstants.PRONOUN + "," + MorfConstants.ED_CHISLO + "," + MorfConstants.NULL + "," + MorfConstants.OSN_P + "
```

Рис. 10. Фрагмент общего массива характеристик частей речи

Второй этап – непосредственный анализ. Он реализован в функции *DetermineOnYourOwn*. Исходное слово подвергается пошаговому разбиению и исследование его составляющих на предмет соответствия к какой-либо части речи. Алгоритм морфологического анализа приведен ниже (рис.11).

Алгоритм 3 (алгоритм морфологического анализа)

Шаг 1. Убрать «окончания» входного слова. Удалить повторяющуюся часть, если таковая имеется. Конвертировать некоторые варианты слов.

Шаг 2. Запустить функцию определения части речи. Все выявленные результаты заносить в динамический список.

Шаг 3. В зависимости от количества определенных частей речи, корня и аффиксов из исходного слова, запускается цикл

определения остальных морфологических характеристик для каждого элемента списка.

Шаг 4. Полученные данные заносятся в переменные.

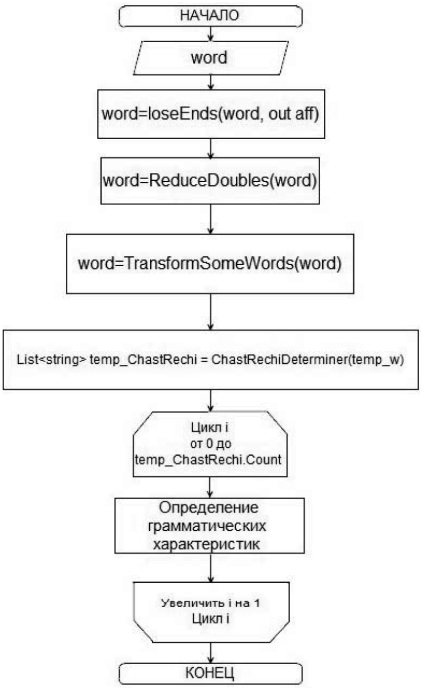


Рис. 11. Алгоритм морфологического анализа

Определение части речи происходит в классе *ImprovedAffixHandler*. Класс предоставляет набор функций для решения следующих задач:

1. Разбить исходное слово на составляющие корень и аффиксы.
2. Производить поиск по выявленным кандидатам.

Выполнением первой задачи занимается метод *MainWordDivider*. Цель метода – предоставить функции поиска кандидатов: корень и аффиксы (рис. 12).

Алгоритм 4 (алгоритм разбивки слова на составляющие)

Шаг 1. На вход принимаются исходное слово и юнит. При первом входе $i = 1$, юнит равен пустой строке. Если i меньше длины слова, то выделить от слова i символов, записать в строку.

Шаг 2. При первом входе проверить существуют ли такой аффикс. При дальнейших заходах также проверить совместимость отделенного аффикса и юнита аффиксов. Если аффиксы существуют и совместимы – переход к шагу 3, иначе к шагу 4.

Шаг 3. К юниту аффиксов прибавляется новый аффикс. Предполагаемые корень и юнит отправляются на вход функции поиска. Функция *MainWordDivider* вызывает саму себя с «новыми» урезанным словом и юнитом на входе.

Шаг 4. *i* увеличивается на 1. Переход к шагу 1.

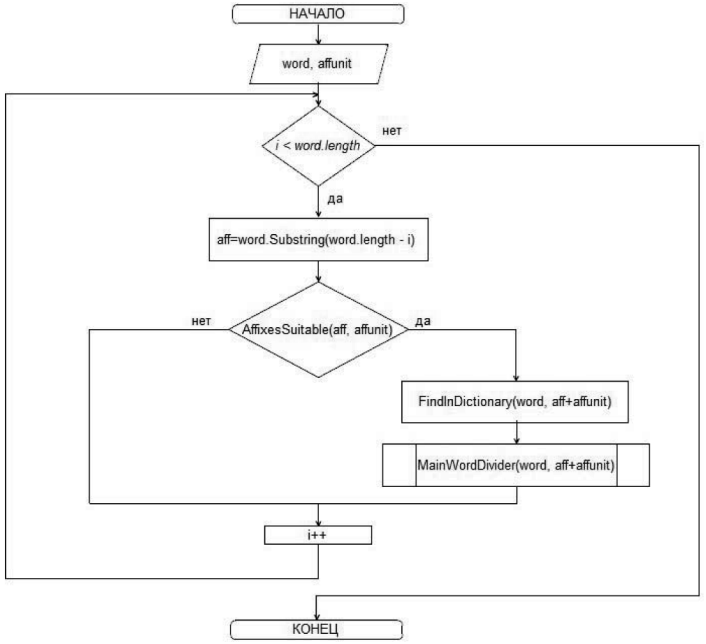


Рис. 12. Алгоритм поиска кандидатов: корень и аффиксы

Проверка совместимости аффиксов полностью полагается на структуру БА. Благодаря разделению на уровни и типы контроль «подходимости» аффиксов сводится к простому «просмотру» БА. Аффикс и юнит совместимы, если соблюдаются следующие условия:

1. Их типы совпадают или один из типов относится к *Any*.
2. Уровень аффикса меньше уровня юнита.

После прохождения проверки аффикс присоединяется к юниту и становится его частью. В следующей итерации отделяется новый аффикс и анализируется на сочетание с новым юнитом. Пример работы алгоритма с итерациями на слове «*килмерĕ*» («не пришел») представлен ниже (рис. 13). Условные обозначения: i – количество извлекаемых символов с правого края слова, *aff* – отделенный аффикс длиной *i* символов, *cmp* – проверка существования и совместимости аффиксов, *found* – найденный результат.

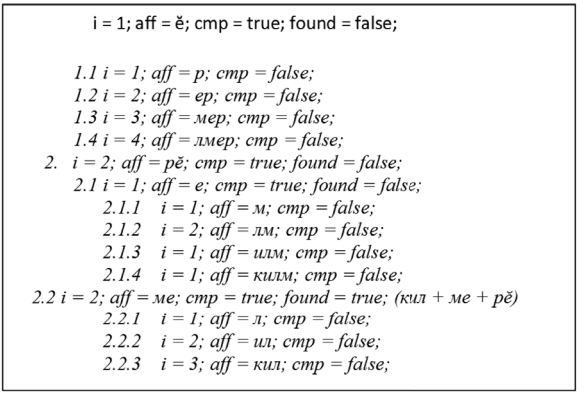


Рис. 13. Пример работы алгоритма с итерациями

Вторая задача решается функцией *FindInDictionary*. Задача – подвергнуть анализу входные корень и аффиксы; среди большого количества вариантов выделить корректные и сохранить в динамических списках (рис. 14).

Алгоритм 5 (алгоритм поиска, выявление совпадений, определение морфологических характеристик)

Шаг 1. На основе входного юнита определить паттерн: набор частей речи, «пригодных» для массива входных аффиксов.

Шаг 2. Определить правила контекста. При отсутствии подходящих правил установить значения по умолчанию.

Шаг 3. В указанном диапазоне производить поиск. Если часть речи выбранного кандидата подходит под паттерн, переход к шагу 4. Если нет – поиск нового номинанта.

Шаг 4. Все элементы массива символов восстановления «примеряются» к корню-кандидату. При положительном фиттинге переход к шагу 5. Иначе возврат к шагу 3.

Шаг 5. Конечная проверка совместимости выявленных корней и аффиксов. Прошедшие анализ на сочетаемость объекты заносятся в списки. Переход к шагу 3 – поиск новых кандидатов до конца диапазона.

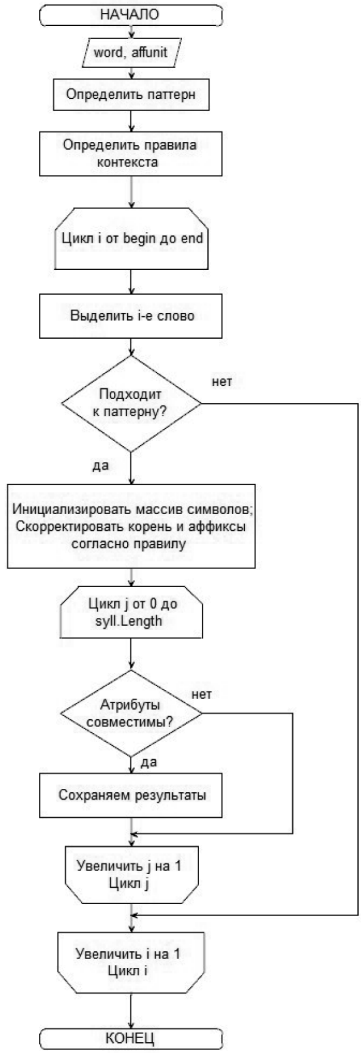


Рис. 14. Алгоритм определения корректности корня и аффиксов и сохранение их в динамических списках

4. Анализ контекстов

Для чувашского языка свойственны фонологические изменения, которые вызывает морфема при ее присоединении к основам и аффиксам. К ним можно отнести выпадение гласных (*лаша-> лаш(а) +и +не*), удвоение согласных (*пуртӑ -> пурттти*) и другие.

В чувашском языке, как и в других агглютинативных языках, в отличие от флективных, основы не «искажаются». Исключение составляют ограниченный перечень таких фонологических явлений, как

1) элизия (выпадение гласных):

а) в словообразовании: *ват çын* вместо *ватӑ çын* (старый человек); *пурнӑç* вместо *пурӑнӑç* (от глагола *пурӑн* – жить);

б) в словоизменении: *вула* (читать), *вул-ӑп* (буду читать);

2) выпадение согласных: в некоторых формах глаголов на -р выпадает р, например, *пыр – пытӑм* (пришел). Всего таких глаголов 10.

3) удвоение согласных в существительных, стоящих в дательно-винительном падеже на *ӑ, ӗ: тулӑ* (пшеница), *тулл-а* (пшенице, пшеницу);сӗлӗ (овёс), *сӗлл-е* (овсу).

4) чередование конечных согласных *-у, – ӳ с – ӑв, -ӗв: çыру* (письмо), *çырӑва* (письму); *вӗренӳ* (учение, учеба), *вӗренӗве* (учебе, учебу, учению, учение).

5) орфографические изменения, происходящие в словах заимствованных из русского или в чувашских словах из-за особенностей чувашского языка при передаче его слов принятой системой письма:

1. Если слово оканчивается на *о*, на которую не падает ударение: то в падежах кроме основного: оно выпадает: вместо нее появляется *ӑ* (*агенство, агенствӑна*).

2. Если слово оканчивается на *е*, на которую не падает ударение, то в падежах кроме основного оно выпадает, вместо нее появляется *ӗ* (*бензохранилище, бензохранилищӗн*).

3. Если слово оканчивается на *ея*, то в падежах кроме основного, *ея* выпадает, вместо нее появляется *йӑ* (*аллея, аллейӑн*).

4. Если слово оканчивается на *ий*, то в падежах кроме основного: *й* выпадает (*опий, опишӗн*).

5. Если слово оканчивается на *ия*, то в падежах кроме основного: *я* выпадает (*стихия, стихисĕр*).

6. Если слово оканчивается на *сть*, то в падежах кроме основного *сть* меняется на *ç* (*область, облаçе*).

7. Если слово находится во множественном числе (аффикс *-сем*), то в притяжательном, дательно-винительном, местном, лишительном падежах м меняется на н (например: *ялсем, ялсен*, *ялсене*, *ялсенче*, *ялсенчен*).

Задачей МА является способность корректно обрабатывать каждый из рассмотренных случаев. Для решения предложена система правил, результатом анализа которой будет список символов, «пригодных» для восстановления.

Структура правил разбита на две части. Первая – раздел входных параметров. Входными являются левый и правый контексты. Они представляют собой преобразованные основу и аффиксы соответственно. Вторая часть – раздел выходных параметров: символы восстановления, инструкции по изменению левого контекста, а также часть речи, для которой это правило применимо.

Принятые условные обозначения: *B* – твердая гласная, *F* – мягкая гласная, 2*C* – 2 согласные рядом («тт», «нн»). Произвольное состояние контекста обозначается пустой строкой.

Инструкции по изменению левого контекста: -*N* – удалить *N* символов, *S* – добавить к левому контексту символ S, пустая строка – изменения не требуются. Каждая рубрика отделяется от другой символом «;».

Рассмотрим примеры:

1. В чувашском языке есть 10 глаголов, у которых выпадает –р при присоединении некоторых аффиксов. Правило для восстановления –р будет выглядеть следующим образом (*кĕ, пы, шă, тă, па, я, пе, йĕ, ху, кӳ;;;р*; глагол).

2. Существительным, оканчивающимся на *–у (çыру, пăру, ăмăрту*), свойственно преобразование *–у* на *–ăв* в некоторых падежах. Следующее правило способно восстановить выпадающую –у и убрать лишнее *–ăв*: (*ăв*;;-2;*у*;сущ-е).

3. В словах «*пуртă*», «*атă*», «*усă*» согласные перед «*ă*» удваивается в разных вариациях. Правило записывается таким образом:

(2С;В;-1;*ă*;сущ-е)

Разработано 17 правил обработки контекстов (рис.15).

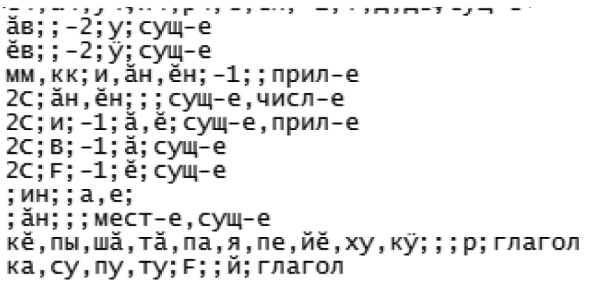


Рис. 15. Фрагмент базы правил обработки контекстов

Для обработки правил разработан класс *ContextRules*. Класс содержит набор методов и функций для обработки входных контекстов, анализа файла правил. Центральной функцией является *FindTemplate*. Задача функции на основе входных кандидатов (корня и аффиксов) определить подходящее контекстное правило, иначе – определить набор выходных данных по умолчанию. Здесь приведен алгоритм анализа контекстов. Ниже текстовый алгоритм представлен графически (рис. 16).

Алгоритм 6 (алгоритм анализа контекстных правил)

Шаг 1. На вход функции подаются предполагаемые корень и аффиксы. Для анализа контекстов от корня извлекаются крайние 2 символа. При длине корня-кандидата менее 2 символов слово не обрезается. От юнита аффиксов выбирается первый аффикс вне зависимости от его длины.

Шаг 2. Извлеченные символы и аффикс преобразуются в контекстный вид согласно разработанным функциям.

Шаг 3. Выполняется поиск в файле правил в секции входных параметров. Полученные левый и правый контексты сравниваются с соответствующими аргументами. Если найдено совпадение – переход к шагу 4, иначе переходим к следующей строки правил.

Шаг 4. Из строки правил извлекаются выходные данные: набор символов восстановления, инструкции по модифицированию корня, юнита аффиксов.

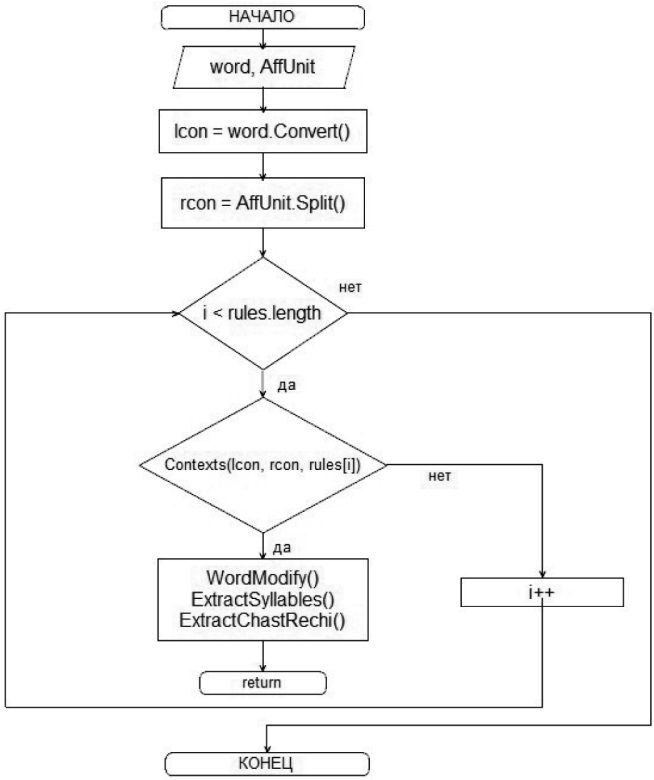


Рис. 16. Алгоритм анализа контекстов

5. Обработка аффиксов

Для модифицирования, склеивания, анализа аффиксов разработан класс ***AffixInfoProvider***. Представляет собой набор методов и полей, предназначенных для обработки аффиксов, определения паттернов.

AddAffixToUnit. Функция сложения отдельного аффикса с юнитом. Вызывается после проверки сочетаемости. Главная задача функции объединять разрозненные аффиксы в юниты, отделять друг от друга знаком разделения. Для этой цели используется символ «|». Юнит из нескольких аффиксов на выходе функции выглядит следующим образом: «сем|пе|х|ч|ё».

AffixesSuitable. Функция анализирует аффиксы по следующим критериям:

1. Существует ли такой аффикс.
2. Если да, определяет уровень и тип аффикса.
3. Для юнита извлекает старые данные о типе и уровне.
4. Уровень аффикса должен быть меньше уровня юнита, типы должны быть совместимы.

LevelOfSingleAffix, ***LevelOfAffUnit***. Вспомогательные функции AffixesSuitable. Выполняют непосредственное определение уровня и типа аффиксов.

Алгоритм 7 (алгоритм выявления типа и уровня аффикса)

Шаг 1. Проверяемому аффиксу устанавливается значение уровня по умолчанию.

Шаг 2. Исследование самого верхнего уровня БА. Если совпадение найдено, переход к шагу 3. Иначе – переход на уровень ниже.

Шаг 3. Присваиваем новое значение уровня. Извлекаем тип аффикса.

Алгоритм аналогичен и для юнита аффиксов с некоторыми оговорками:

1. Уровень юнита определяется уровнем крайнего левого аффикса. В юните «сем|пе|х|ч|ĕ» уровень «-сем» отражает уровень всего юнита.

2. Тип юнита принадлежит *NotGlagol* или *OnlyGlagol*, если есть хотя бы один аффикс такого же типа, иначе Any. В «сем|пе|х|ч|ĕ» тип *NotGlagol*, потому что «пе» того же типа, хотя аффикс «х» имеет тип Any.

TypeOfSearch. Предназначение функции – определить тип поиска на основе юнита, определить паттерн. Задача решается рассмотрением каждого аффикса в юните. Обработкой разных случаев занимается блок *switch*/*case*. В case-блоках перечисляются схожие аффиксы, относящиеся к одной группе. При невхождении ни в одну из групп, тип поиска присваивается типу юнита. Результат возвращается как элемент одноименного перечисления. В перечислении кроме основных типов заданы: *Prichastie*, *DeePrichastie*, *Prilagatelnoe*, *NotNarechieGlagol*, *Chislitelnoe*, *deenoun*. Соответствие между типами и аффиксами представлено в табл. 3.

Таблица 3

Prichastie	«ни», «акан», «екен», «аканни», «екенни», «малли», «мелли», «малла», «мелле», «манни», «менни», «нă», «нĕ», «ас», «ес», «асси», «есси», «асшăн», «есшĕн», «нăçем», «нĕçем»
DeePrichastie	«са», «се», «сан», «сассăн», «сессĕн», «масăр», «месĕр»
Prilagatelnoe	«лă», «лĕ», «ллă», «ллĕ», «ли», «лли»
NotNarechieGlagol	«на», «не»
Chislitelnoe	«мăш», «мĕш»
deenoun	«а», «е», «сен»

TypeOfRecovery. Имеет схожую структуру с *TypeOfSearch*. Функция отвечает за определение типа восстановления на основе аффикса. При помощи блока switch/case аффиксы подразделяются на 3 группы: *With*, *Without*, *Both* (табл. 4).

Таблица 4

With	«и», «у»
WithOut	«м», «а», «е», «ти», «ри», «рен», «аттăм», «еттĕм», «аттăн», «еттĕн», «ттăм», «ттĕм», «ттăн», «ттĕн», «атчĕç», «етчĕç», «тчĕç», «сассăн», «тăп», «тĕп», «тăн», «тĕн», «атăп», «етĕп», «атăн», «етĕн», «тпăр», «тпĕр», «атпăр», «етпĕр», «атт», «етт», «тан», «тен», «та», «те», «ла», «ле», «сăр», «сĕр», «сăмăр», «сĕмĕр», «ам», «ем», «ăм», «ĕм», «тăм», «тĕм», «на», «не», «ра», «ре», «сен», «хи», «ччен», «чен», «серен», «сем», «н», «алла», «елле», «па», «пе», «сам», «ран», «сан», «сене», «ма», «ме», «масăр», «месĕр»

В блоке *WithOut* собраны аффиксы, перед которыми восстановление символа не требуется. Так как подобные аффиксы не имеют визуальных признаков, обширный список вручную пополнялся на этапе тестирования. Если аффикс не принадлежит ни к одной категории к нему прикрепляется тип Both: режим и с восстановлением и без восстановления.

6. Определение морфологических характеристик

Для распознавания второстепенных морфологических характеристик, таких как падеж, число, время и т.д. создан класс *FeaturesDeterminers*. После прохода через ядро МА выявлены часть речи, корень и аффиксы входного слова. Цель класса – обеспечить корректное установление характеристик, пользуясь полученными данными. Класс условно разделен на две части:

1. Раздел переменных. Для каждой грамматической характеристики заданы переменные строкового типа. В свою очередь, эти переменные разделены соответственно для каждой части речи (рис. 17).

```
#region VERB
//FACE
string[] first_verb = { "ăп", "ĕп", "п", "ап", "еп", "ăм", "ĕм" };
string[] second_verb = { "ăн", "ĕн", "ăр", "ĕр", "ан", "ен", "на", "не" };
string[] third_verb = { "ать", "ет", "аç", "еç", "ё", "ччĕр", "ччăр", "ç", "нă", "нĕ", "ть", "т", "ччĕ", "тăр", "тĕр" };

//CHISLO
string[] pl = { "ăр", "ĕр", "аç", "еç", "ç", "ччĕр" };
//VREMYA
string[] nast_verb = { "ат", "ет", "ать", "аç", "еç", "ть", "п", "ап", "еп", "мас", "мес", "ан", "ен" };
string[] prosh_verb = { "р", "ч", "ăм", "ĕм", "атт", "ett", "сатт", "сетт", "сачч", "сечч", "атч", "етч", "ччĕ", "чĕ" };
string[] bud_verb = { "ĕ" };

string[] ambi_vremya_verb = { "ăп", "ĕп", "ăн", "ĕн", "ăр", "ĕр" };
string[] unknown_vremya = { "ма", "ме", "иччен" };
//NEGATIVE
string[] negative = { "мас", "мес", "маç", "меç", "м", "сăр", "сĕр" };
//INFINITIV
string[] infinitiv = { "ма", "ме", "машкăн", "мешкĕн" };
```

Рис. 17. Фрагмент переменных класса FeaturesDeterminer

2. Раздел функций. Аналогично переменным, функции подразделены согласно каждой части речи. Каждая функция отвечает за установление одной грамматической характеристики (рис. 18).

На основе анализа чувашской грамматики разработана библиотека *MorfParsingLibrary*.dll. Библиотека предоставляет набор методов для морфологического анализа слов чувашского языка. Продукт может быть использован в системах машинного перевода, быть составной частью лингвистического процессора. Библиотека написана на языке *C#* в среде *Visual Studio* 2013.

Для работы необходимо подключить через *References* библиотеку, создать и проинициализировать объект класса *MorfParser*. Единственный открытый конструктор класса принимает в качестве входного параметра булеву переменную. При истинном зна-

```
//noun
internal string DeterminePadezhOfSusch(string word, string[] aff_massiv)
{
    if (aff_massiv.Contains("ён"))
    {
        if (aff_massiv.Contains("че")) return MorfConstants.MEST_P;
        if (aff_massiv.Contains("чи")) return MorfConstants.MEST_P;
        if (aff_massiv.Contains("чен")) return MorfConstants.ROD_P;
        return MorfConstants.ROD_P;
    }

    if (aff_massiv.Any(el => rod.Contains(el))) return MorfConstants.ROD_P;
    if (aff_massiv.Any(el => dat.Contains(el))) return MorfConstants.DAT_P;
    if (aff_massiv.Any(el => mest.Contains(el))) return MorfConstants.MEST_P;
    if (aff_massiv.Any(el => isch.Contains(el))) return MorfConstants.ISCH_P;
    if (aff_massiv.Any(el => tvor.Contains(el))) return MorfConstants.TVOR_P;
    if (aff_massiv.Any(el => lish.Contains(el))) return MorfConstants.LISH_P;
    if (aff_massiv.Any(el => prich_celevoy.Contains(el))) return MorfConstants.PR_CEL_P;

    return MorfConstants.OSN_P;
}
```

Рис. 18. Функция определения падежа существительных

чении включается режим логгирования. Система логгирования заносит в текстовый файл значимые события работы для более легкого мониторинга работы библиотеки. Логгирование основано на методах класса Trace. Для лучшей производительности рекомендуется лично настроить выбор событий, результаты которых интересуют в качестве лога. Файл создается в директории библиотеки (рис.19).

```
На вход поступило слово: мана

Анализируем: ман |а
Тип поиска: [deenoun]
No Context Rule
Совпадение найдено.  Слово совместимо с аффиксом? - True
Добавляем: [местоимение: ман а]
Совпадение найдено.  Слово совместимо с аффиксом? - True
Добавляем: [деепричастие: ман а]
Переход к следующей паре.
==================================================
Анализируем: ма |на
Тип поиска: [NotNarechieGlagol]
No Context Rule
Переход к следующей паре.
==================================================
Анализируем: м |а|на
Тип поиска: [deenoun]
No Context Rule
Переход к следующей паре.
==================================================
```

Рис. 19. Фрагмент файла лога в директории библиотеки

В файл лога записываются:

1. Поступившее слово в исходном виде.

2. Предполагаемые корень и аффиксы для каждой итерации.
3. Тип поиска.
4. Информация о контекстном правиле.
5. Информация о совместимости пар-кандидатов.

Также в классе *MorfParser* для МА доступны два открытых метода. Первый метод *DealWithManualText* принимает на вход текстовую строку. Функция разбивает строку на отдельные слова и каждое из них пропускает через ядро МА. Морфологическая структура каждого слова записывается в свойствах класса *MorfParser* и добавляется в список. На выходе функция возвращает динамический список объектов *MorfParser*. Вторая функция имеет схожую структуру, однако на вход принимает путь к файлу. В этом случае все определенные результаты записываются в текстовый файл (рис. 20), создаваемый в той же директории, что и библиотека. Функцию вывода атрибутов слов в файл можно использовать независимо, вызвав ее вручную. Достаточно передать массив проанализированных объектов и указать путь сохранения файла.

```
мана местоимение единственное unknown дательный ман а 1е неотрицание null
мана деепричастие unknown unknown unknown ман а unknown неотрицание null
ачасем имя существительное множественное unknown основной ача сем 1е неотрицание nul
итлемеççĕ глагол множественное настоящее unknown итле м|еç|ç|ĕ 3е отрицание неинфини
витрепе имя существительное единственное unknown творительный витре пе 1е неотрицани
килтермеллисемпеленехчĕ-и причастие множественное unknown творительный килтер мелл|и
велосипедлисемпеленехчĕ-и прилагательное множественное прошлое творительный велосипе
эп местоимение единственное null основной эп  1е неотрицание null
шутларăм глагол единственное прошлое unknown шутла р|ăм 1е неотрицание неинфинитив
та союз null null null та  null неотрицание null
та частица null null null та  null неотрицание null
киле имя существительное единственное unknown дательный кил е 1е неотрицание null
киле деепричастие unknown unknown unknown кил е unknown неотрицание null
каяс причастие unknown будущее unknown кай ас unknown неотрицание null
мар частица null null null мар  null неотрицание null
пуль частица null null null пуль  null неотрицание null
терĕм глагол единственное прошлое unknown те р|ĕм 1е неотрицание неинфинитив
```

Рис. 20. Фрагмент текстового файла с результатами

По умолчанию функция записывает все атрибуты проанализированных слов. По желанию разработчика список выводимых характеристик может меняться. Далее приводится описание содержимого класса *MorfParser*. *ImprovedAffixHandler iah;* //Объект класса, отвечающего за главный анализ и поиск в словаре.

1. *FeaturesDeterminers fd*; //Объект класса, отвечающего за определение характеристик(атрибутов) входных слов
2. *string dictionary_path;* //путь к словарю
3. *string affixes_path;* //путь к аффиксам

4. *string rules_path; //*путь к правилам

5. *string exceptions_path;* //путь к исключениям

6. *string affinfo_path;* //путь к описаниям аффиксов

7. *string InputText_Path;* //путь ко входному файлу

8. *string OutputText_Path;* //путь выходного файла

9. *internal static string[] Dictionary, Affixes, Rules, Exceptions, AffInfo; //*Массивы для хранения содержимого Словаря, Аффиксов, Правил, Исключений, Описаний аффиксов

10. *public string Word { get; private set; }* //Входное слово

11. *public List<string> ChastRechi { get; private set; }* //Часть речи

12. *public List<string> Root { get; private set; }* //Корень

13. *public List<string> Affix { get; private set; }* //Аффиксы

14. *public List<string> Vremya { get; private set; }* //Время

15. *public List<string> Padezh { get; private set; }* //Падеж

16. *public List<string> PluralOrNot { get; private set; }* //Число

17. *public List<string> Face { get; private set; }* //Лицо

18. *public List<string> Negativ { get; private set; }* //Негатив

19. *public List<string> Infinitiv { get; private set; }* //Инфинитив

20. *public List<string> AffixInfo { get; private set; }* //Инфа об аффиксах

21. *List<KeyValuePair<string, string>> RootAffixesCollection; //* Коллекция для хранения пар (корень-аффиксы)

22. *public int COUNT { get; private set; }* //Количество частей речи входного слова

Конструкторы:

23. *public MorfParser(bool enablelogging){}* //Конструктор класса. Инициализирует начальные значения своих полей. Содержимое вспомогательных файлов записывает в соответствующие массивы. По словарю устанавливает указатели(метки) для улучшения поиска по нему. В зависимости от входного булевого параметра вкл/выкл логгирование

24. *private MorfParser(MorfParser other){}* //Конструктор копирования. Используется при передаче объектов MorfParser в список.

25. *private MorfParser(string word){}* //Конструктор для неопределенных(нераспознанных) слов. Возвращает объект *MorfParser* со значением *Unknown* на всех полях для переданного через вх. параметр слова

Функции:

26. *public void InitializeComponents()* //Выполняет начальную инициализацию полей/свойств

27. *public void DealWithTextFromFile(string InputPath)* // Анализирует вх. текст, расположенный заданному пути. Записывает вых. данные в файл по пути *OutputTextPath*

28. *public List<MorfParser> DealWithManualText(string InputText)* //Анализирует вх. текст напрямую. Возвращает список объектов *MorfParser*.

29. *internal void AddtoRootAffixCollection(string root, string affix)* //Добавляет элементы(корень и аффиксы) в коллекцию *RootAffixCollection*

30. *private List<string> ChastRechiDeterminer(string word)* // Определяет части речи вх. слова. Определяются границы поиска в словаре. Далее в этих границах производится поиск и анализ с помощью функции *MainWordDivider* класса *ImprovedAffixHandler*. Результаты заносятся в список и возвращаются.

31. *private string PluralOrNotDeterminer(string word, string chastrechi, string affixes)* //Определяет число вх. слова на основе его части речи и аффиксов.

32. *private string VremyaDeterminer(string word, string chastrechi, string affixes)* //Определяет время

33. *private string PadezhDeterminer(string word, string chastrechi, string affixes)* //Определяет падеж

34. *private string FaceDeterminer(string word, string chastrechi, string affixes)* //Определяет лицо

35. *private string NegativDeterminer(string word, string chastrechi, string affixes)* //Определяет негатив

36. *private string InfinitivDeterminer(string word, string chastrechi, string affixes)* //Определяет инфинитив

37. *private string AffixInfoDeterminer(string word, string chastrechi, string affixes)* //Определяет инфу об аффиксах

38. *private void DetermineOnYourOwn(string word)* //Анализирует и определяет все атрибуты вх. слова. Заносит результаты в соответствующие списки.

39. *private void SearchInDictionaries(string word)* //Ищет исх. слово в словаре. При обнаружении присваивает значения.

40. *private void PutInFile(string attributes[])* //Записывает полученные данные в файл

41. *private string[] OnlyDistinct(string[] word_array)* // Исключает повторяющиеся слова из анализа. Оставляет лишь уникальные.

Заключение

В предлагаемой работе разработана библиотека морфологического анализа чувашского языка. Библиотека создана на платформе *NET.Framework* в среде *Visual Studio* 2013 на языке C#. Приведен оригинальный алгоритм рекурсивного метода разбиения исходного слова на составляющие. Разработанная библиотека выполняет следующие задачи:

– определение части речи слова;
– извлечение корня и аффиксов;
– анализ контекстов, восстановление символов;
– определение морфологических характеристик;
– логгирование работы для удобной отладки;

Результаты работы библиотеки успешно могут применены на входе синтаксического анализатора и являюся составной частью лингвистического процессора.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 15-04-00532.

ЛИТЕРАТУРА

1. Желтов П.В. Создание национального корпуса чувашского языка: проблемы и перспективы // Современные проблемы науки и образования. – 2015. – № 1–1. С. 338.

2. Желтов П.В. Сравнительные исследования морфем чувашского языка. – Чебоксары: Изд-во Чуваш. ун-та, 2013. – 166 с.

3. Желтов П.В. Лингвистические процессоры в системах искусственного интеллекта: Монография. – Чебоксары: Изд-во Чуваш. ун-та, 2007. –100 с.

4. Желтов П.В. Лингвистические процессоры. Формальные модели и методы: теория и практика: Монография. – Чебоксары: Изд-во Чуваш. ун-та, 2006. – 208 с.

5. Желтов П.В., Губанов А.Р., Желтов В.П. Морфологический стандарт национального корпуса чувашского языка // Современные проблемы науки и образования. – 2015. № 2. С. 180.

6. Желтов П.В. Формальные методы и модели в сравнительно-сопоставительном языкознании: Монография. – Чебоксары: Изд-во Чуваш. ун-та, 2006. – 252 с.

7. Желтов П.В. Исследования исторического развития чувашского языка. – Чебоксары : Изд-во Чуваш. ун-та, 2013. – 166 с.

8. Желтов П.В. Компьютерное моделирование многоагентных систем. – Чебоксары: Изд-во Чуваш. ун-та, 2008. – 112 с.

9. Желтов П.В. Модели и методы обработки символьной информации на сетях Петри. – Чебоксары: Издательство Чуваш. ун-та, 2012. – 108 с.

10. Желтов П.В. Моделирование многоагентных систем сетями Петри: Монография. – Чебоксары: Изд-во Чуваш. ун-та, 2008. – 108 с.

УДК 81’33

MORPHOLOGICAL ANALYZER OF THE KYRGYZ LANGUAGE

N. Israilova, P. Bakasova
KSTU after I. Razzakov, Bishkek, Kyrgyzstan
inela.kstu@gmail.com, bakasovap@mail.ru

In this article we propose one an implementation of the morphological analyzer of the Kyrgyz language in the context of the translator. Morphological model, features and rules of the morphology of the Kyrgyz language as well as the algorithm of the morphological analysis of the Kyrgyz language are presented.

Keywords: morphological analyzer, machine translation, algorithm of morphological analysis, algorithm of morphological synthesis.

МОРФОЛОГИЧЕСКИЙ АНАЛИЗАТОР КЫРГЫЗСКОГО ЯЗЫКА

Н.А. Исраилова, П.С. Бакасова,
КГТУ им. И.Раззакова, Кыргызстан
inela.kstu@gmail.com, bakasovap@mail.ru

В данной статье рассмотрен вариант реализации морфологического анализатора кыргызского языка в контексте применения в трансляторе. Приведены морфологическая модель, особенности и закономерности морфологии кыргызского языка, а также алгоритм морфологического анализа кыргызского языка.

Ключевые слова: морфологический анализатор, машинный перевод, алгоритм морфологического анализа, алгоритм морфологического синтеза.

1. Морфологическая модель кыргызского языка

Кыргызский язык характеризуется строгой последовательностью в присоединении аффиксов к корню (или основе), например, для образования имени существительного, к корню сначала присоединяется словообразующий суффикс, затем формообразующие аффиксы: числа, принадлежности, падежей и сказуемости. Пропуск одного из этих аффиксов не меняет порядка их присоединения.

2. Морфологическая структура слов кыргызского языка

Рис. 1. Морфологическая структура слов кыргызского языка

На основании анализа грамматики кыргызского языка можно выделить следующие основные правила при словоизменении имен существительных:

– Форма множественного числа указывает на расчлененную (не собирательную) множественность и выражается при помощи аффикса ***-лар*** с фонетическими вариантами. Учитывается гласная последнего слога основы (или корня), а также, если слог закрытый – звонкость, глухость конечной согласной.

– Единственным нарушением закона сингармонизма следует считать присоединение аффикса ***-лар, -дар, -тар*** после основы, содержащей в последнем слоге гласную *у* или *уу*.

– В структуре слова аффиксы принадлежности находятся либо сразу после корня (основы), либо после корня (основы) и аффикса множественного числа, но предшествуют падежным аффиксам.

– Падежные аффиксы могут следовать сразу за корнем, а также за основой, содержащей кроме корня словообразующие аффиксы, аффиксы множественного числа, аффиксы принадлежности.

– При склонении имен существительных, оканчивающихся на сочетания согласных *ск, нк, фт, нг, нд, мн,* после основы добавляются гласные *-ы, и* или *у*, после чего эти слова склоняются как слова с основой, оканчивающейся на гласную.

– Имена существительные в кыргызском языке могут спрягаться.

– При морфологическом способе к словообразующей основе присоединяются словообразовательные аффиксы. Аффиксы, присоединяющиеся к основе имен существительных.

– В кыргызском языке окончания имен существительных делятся на четыре вида. Описанные ниже окончания непосредственно будут использоваться в разрабатываемом алгоритме определения основы слова.

– Обозначим через E*i* – следующие множества окончаний, для *i = 1, 2, 3, 4*.

Таблица 1. **Обозначения окончаний**

Ei	**Обозначения**
E1	окончания множественного числа
E2	окончания принадлежности
E3	падежные окончания
E4	окончания сказуемости

Порядок присоединения аффиксов:

Рис. 2. Порядок присоединения аффиксов

C – словообразующий суффикс *(-чу, -лык, -дык, -чылык, -лаш, -ма* и т.д.)

В таблице 2 описаны определения морфемного состава (*Ei*, где *i = 1, 2, 3, 4, 5*):

Таблица 2. **Виды окончаний**

№	Виды окончаний	Окончания
1	Оконания множественного числа – *E1*	*-лар, -дар, -тар, -лер, -дер, -тер, -лор, -дор, -тор, – лөр, -төр, -дөр*
2	Окончания принадлежности – *E2*	*-м, -ым, -им, -ум, -үм, -ң, -ың, -иң, -уң, -үң, -ңыз, -ңиз, -ңуз, -ңүз, -иңиз, -уңуз, -үңүз, -ыңыз, -быз, -биз, -буз, -бүз, -ыбыз, -ибиз, -убуз, -үбүз, -ңар, -ыңар, -ңыздар, -ыңыздар, -сы, -ы*
3	Падежные окончания – *E3*	*-нын, -дын, -тын, -га, -ка, -ны, -ды, -ты, -да, -та, -дан, -тан*
4	Окончания сказуемости – *E4*	*-мын, -сың, -сыз, -быз, -сыңар, -сыздар*

Для удобства реализации была исследована систематизация окончаний. Порядок правил имеет следующий вид (таблица 3):

Таблица 3. **Порядок правил**

Наименование	Обозначение
O	корень слова (основа)
A1	E1 + E2, окончание множественного числа + окончание принадлежности
A2	E1 + E3, окончание множественного числа + падежное окончание
A3	E1 + E4, окончание множественного числа + окончание сказуемости
A4	E2 + E3, окончание принадлежности + падежное окончание
A5	E2 + E4, окончание принадлежности + окончание сказуемости
A6	E3 + E4, падежное окончание + окончание сказуемости

Тогда, существительное во множественном числе – *O + E1,*
Пример: *окучуу* **+ *лар***

Таблица 4. **Примеры**

Правило	Пример
О + А1	окучуу + ***лар + ыбыз***
О + А1 + Е3	окучуу + ***лар + ыбыз + дан***
О + А1 + А6	окучуу + ***лар + ыбыз + дан + сыздар***
О + А2	окучуу + ***лар + дан***
О + А2 + Е4	оюнчу + ***лар +дан + быз***
О + А3	окучуу + ***лар + быз***
О + А4	окучуу + ***ңуз + дан***
О + А4 + Е4	окучуу + ***ңуз + дан + мын***
О + А5	окучуу + ***ңуз + мун***
О + А6	окучуу + ***дан + мын***

3. Особенности и закономерности морфологии кыргызского языка

Структурно-типологическая характеристика кыргызского языка связана с его принадлежностью к агглютинативным языкам. Для описания языков агглютинативного типа применяется набор признаков, учитывающих не только морфологические, но и синтаксические и фонетические особенности.

Морфологические признаки агглютинации:

1. Корень слова – в именительном падеже выступает в чистом виде, таким образом, является центром всей парадигмы склонения;
2. Между морфемами четко сохраняется граница;
3. Строгая последовательность присоединения аффиксов.

Фонетические признаки агглютинации:

1. Наличие сингармонизма;
2. Фиксированное ударение, которое способствует сохранению фонетической целостности слова.

Синтаксические признаки агглютинации:

1. Твердый порядок слов в предложении;
2. Определение находится перед определяемым словом;
3. Дополняющее слово находится перед дополняемым словом;
4. Сказуемое в конце предложения.

При склонении существительных во множественном числе падежное окончание ставится после окончания множественного числа: *шаарларда – в городах, балдарда – у детей, сабактарда – на уроках.*

В предложении существительное в родительном падеже является определением для другого существительного. При этом определяемое существительное обязательно принимает окончание -сы {-си, -су, -су), если оно оканчивается на гласную, или -ы (-и, -уг -у), если оканчивается на согласную. Например, *окуучунун дептери – тетрадь ученика, мектептин короосу – двор школы.*

Существительное в родительном падеже, служащее определением, часто переводится на русский язык как прилагательное. При этом окончание родительного падежа может отсутствовать. Например, *колхоз малы – колхозный скот (вм.: колхоздун малы – скот колхоза).*

В кыргызском языке, в отличие от русского, прилагательные и порядковые числительные не согласуются с существительными, а просто примыкают к нему: *жаны үй – новый дом, жаңы үйдө – в новом доме.*

Кроме прошедшего определенного времени (на -ды/-ты) в кыргызском языке имеется прошедшее обычное время, которое обозначает события, происходившие давно и не приуроченные к определенной дате.

В кыргызском языке имена существительные, обозначающие профессию или занятие человека, могут спрягаться, т.е. изменяться, как глаголы, по лицам, принимая личные окончания глаголов будущего времени 1 и 2 л, ед. и мн. числа, за исключением 3 л, ед. и мн. числа, где личные окончания отсутствуют. В 3 л. мн. числа к существительному может присоединяться окончание -лар. Например, *мен окуучумун – я ученик, сен окуучусуң – ты ученик, сиз окуучусуз – Вы ученик, ал окуучу ~ он ученик, биз окуучубуз – мы ученики, силер окуучусуңар – вы ученики, сиздер окуучусуңар – Вы ученики, алар окуучу (-лар) – они ученики.*

В отличие от русского в кыргызском языке при обозначении принадлежности предмета употребляются не только притяжательные местоимения (менин – мой, сенин – твой и т.д.), но и особые притяжательные окончания, которые присоединяются к существительному.

В кыргызском языке прилагательные не согласуются с существительными ни в числе, ни в падеже, а грамматическое понятие рода в кыргызском языке отсутствует.

В кыргызском языке количественные и порядковые числительные склоняются по тем же правилам, что и существительные. Порядковые числительные склоняются лишь в том случае, если они не служат определением к существительному, а употребляются самостоятельно.

4. Структура морфологической базы данных кыргызского языка

Морфологическая база данных должна содержать всю информацию, необходимую для работы процедур морфологического анализа и синтеза.

Если в выбранной морфологической модели принят словарь словоформ, то база данных должна содержать все словоформы учитываемых лексем с указанием их грамматических характеристик и принадлежности определенной лексеме.

Если же в морфологической модели принят словарь основ, то база данных, помимо основ учитываемых лексем, должна содержать словарь списков флексий, соответствующих каждому парадигматическому классу.

С каждой флексией должен быть связан набор значений ГП, приписываемый основе с данной флексией. Если в морфологической модели учитываются какие-либо типичные особенности словоизменения (например, чередование букв в основе), то информация о них также должна храниться в базе данных.

Морфологическая БД помимо лексем с регулярным словоизменением должна содержать лексемы с отсутствующими формами, с супплетивными, неизменяемые существительные. Кроме того, БД обязательно должна содержать омонимичные лексемы (с полной и частичной омонимией).

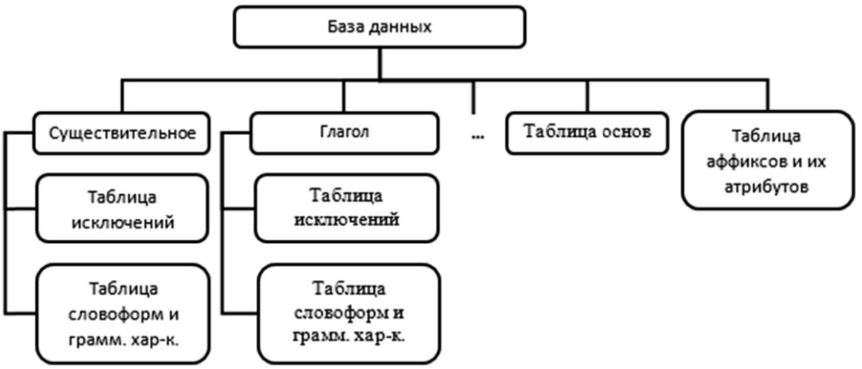


Рис. 3. Структура морфологической базы данных кыргызского языка

5. Алгоритм морфологического анализа кыргызского языка

На входе в МА все лексемы проверяются на наличие в БД основ. Найденные слова подаются на выход с аттрибутами слова (для сущ.: И.п., ед.ч.). А остальные проверяются с помощью алгоритма стемминга.

Алгоритм стемминга направлен на нахождение основы слова для заданного исходного слова. Данный алгоритм работает последовательно, применяя ряд правил отсечения окончаний и суффиксов для определения основы слова.

В реализации алгоритма все слова считываются в обратном порядке (справа налево).

В реализации морфологического анализа были составлены таблицы окончаний и суффиксов и их атрибуты (число, род, падеж и т.д.).

На выходе МА к каждой лексеме присваивается множество атрибутов. Полученные данные анализа служат входными данными для построения семантического дерева.

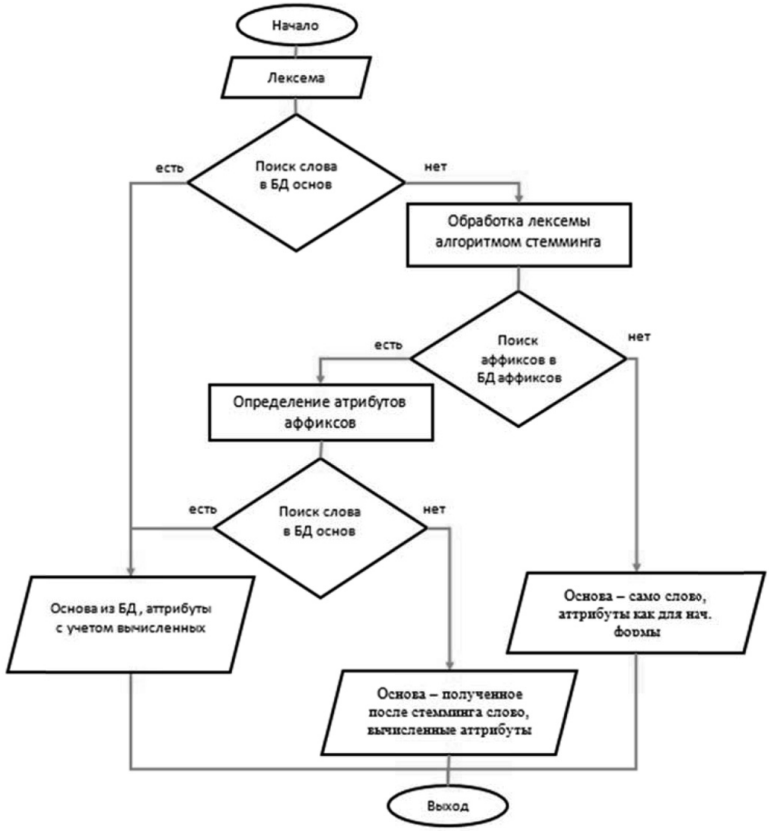


Рис. 4. Алгоритм морфологического анализа слов

6. Алгоритм морфологического синтеза кыргызского языка

Модель морфологического синтеза разработана на основе грамматики кыргызского языка с учетом семантических правил.

Входными данными морфологического синтеза являются выходные данные морфолгического анализатора (корень слова, полученный в результате морфологического анализа и его атрибуты) и заданные атрибуты аффиксов.

Необходимые морфологическому синтезу входные атрибуты указываются в алгоритме словообразования.

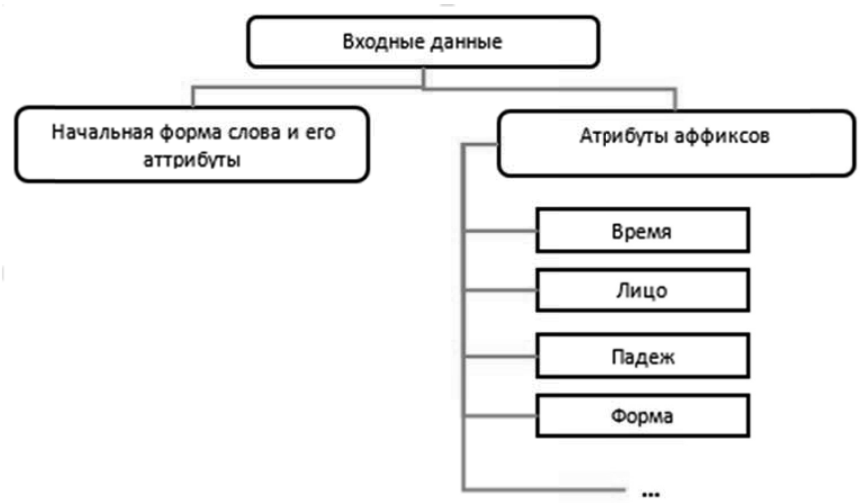


Рис. 5. Входные данные морфологического синтеза

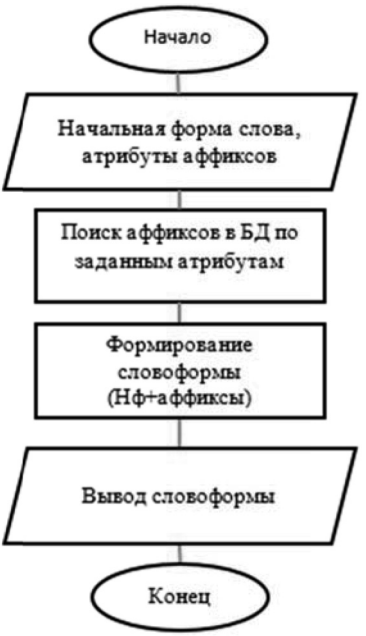


Рис. 6. Алгоритм морфологического синтеза

На выход морфологического синтеза подается словоформа с морфологической структурой. Результат данного алгоритма может использоваться в различных лингвистических системах.

Морфологический синтез на основе перевода обеспечивает образование разных форм входного слова по правилам кыргызского языка и реализовывается в данной системе на основе следующего алгоритма (на примере имени существительного):

1. На вход модуля синтеза подаётся слово в начальной форме (И.п., ед.ч.)

2. Для формировани разных форм имени существительного в кыргызском языке необходимо определить на какую букву оканчивается исходное слово.

3. Если слово оканчивается на гласную букву, то формы слов образуются путем добавления соответствующих окончаний:

– слово оканчивается на ***–а, -я, -ы,*** падежные окончания ***–нын, -га, -ны, -да, -дан;***

– слово оканчивается на ***–и, -е,*** падежные окончания ***–нин, -ге, -ни, -де, -ден;***

– слово оканчивается на ***–о, -у,*** падежные окончания ***–нун, -го, -ну, -до, -дон;***

– слово оканчивается на ***–ө, -ү,*** падежные окончания ***–нүн, -гө, -нү, -дө, -дөн.***

4. Если слово оканчивается на звонкую согласную, то проверяется гласная в последнем слоге и формы слов образуются путем добавления соответствующих окончаний:

– гласная в последнем слоге ***–а, -я, -ы***, падежные окончания ***–дын, -га, -ды, -да, -дан;***

– гласная в последнем слоге ***–и, -е***, падежные окончания ***–дин, -ге, -ди, -де, -ден;***

– гласная в последнем слоге ***–о, -у***, падежные окончания ***–дун, -го, -ду, -до, -дон;***

– гласная в последнем слоге ***–ө, -ү***, падежные окончания ***–дүн, -гө, -дү, -дө, -дөн.***

5. Если слово оканчивается на глухую согласную, то проверяется гласная в последнем слоге и формы слов образуются путем добавления соответствующих окончаний:

– гласная в последнем слоге ***–а, -я, -ы,*** падежные окончания ***–тын, -ка, -ты, -та, -тан;***

– гласная в последнем слоге ***–и, -е,*** падежные окончания ***–тин, -ке, -ти, -те, -тен;***

– гласная в последнем слоге ***–о, -у,*** падежные окончания ***–тун, -ко, -ту, -ко, -дон;***

– гласная в последнем слоге ***–ө, -ү,*** падежные окончания ***–түн, -кө, -тү, -төө, -төн.***

6. В Слово1 запишем существительное во множественном числе.

В кыргызском языке множественное число существительных образуется при помощи окончаний: а***) -лар (-лер -лор, -лор)***, если слово оканчивается на гласную, а также Й, Р: ***алма – алмалар, терезе – терезелер, тоо – тоолорг, кочө – кечөлөр, агай – агайлар, кар – карлар;***

б) ***-дар (-дер, -дор, -дор)***, если слово оканчивается на звонкую согласную: ***кыз – кыздар, эл – элдер, колхоз – колхоздор, мугалим – мугалимдер, көл – көлдөр;***

в) ***-тар (-тер, тор, тер)***, если слово оканчивается на глухую согласную: ***мектеп – мектептер, тоок – тооктор, кат – каттар, өрдак – өрдөктөр.*** Исключение: ***бала – балдар.***

7. Затем производится склонение существительного во множественном числе.

8. Если слово оканчивается на звонкую согласную, то проверяется гласная в последнем слоге и формы слов образуются путем добавления соответствующих окончаний:

– гласная в последнем слоге ***–а, -я, -ы,*** падежные окончания ***–дын, -га, -ды, -да, -дан;***

– гласная в последнем слоге ***–и, -е,*** падежные окончания ***–дин, -ге, -ди, -де, -ден;***

– гласная в последнем слоге ***–о, -у***, падежные окончания ***–дун, -го, -ду, -до, -дон;***

– гласная в последнем слоге ***–ө, -ү***, падежные окончания ***–дүн, -гө, -дү, -дө, -дөн.***

9. Если слово оканчивается на глухую согласную, то проверяется гласная в последнем слоге и формы слов образуются путем добавления соответствующих окончаний:

– гласная в последнем слоге ***–а, -я, -ы***, падежные окончания ***–тын, -ка, -ты, -та, -тан;***

– гласная в последнем слоге ***–и, -е***, падежные окончания ***–тин, -ке, -ти, -те, -тен;***

– гласная в последнем слоге ***–о, -у***, падежные окончания ***–тун, -ко, -ту, -ко, -дон;***

– гласная в последнем слоге ***–ө, -ү,*** падежные окончания ***–түн, -кө, -тү, -тө, -төн.***

10. Образованные словоформы записываются и сохраняются в базе слов.

7. Результаты тестирования

Для тестирования алгоритмов было разработано приложение на C#, которое определяет основу слова кыргызского языка и генерирует словоформы. Пример работы приложения показан на рисунке 7 (для краткости приведена только часть словоформ).

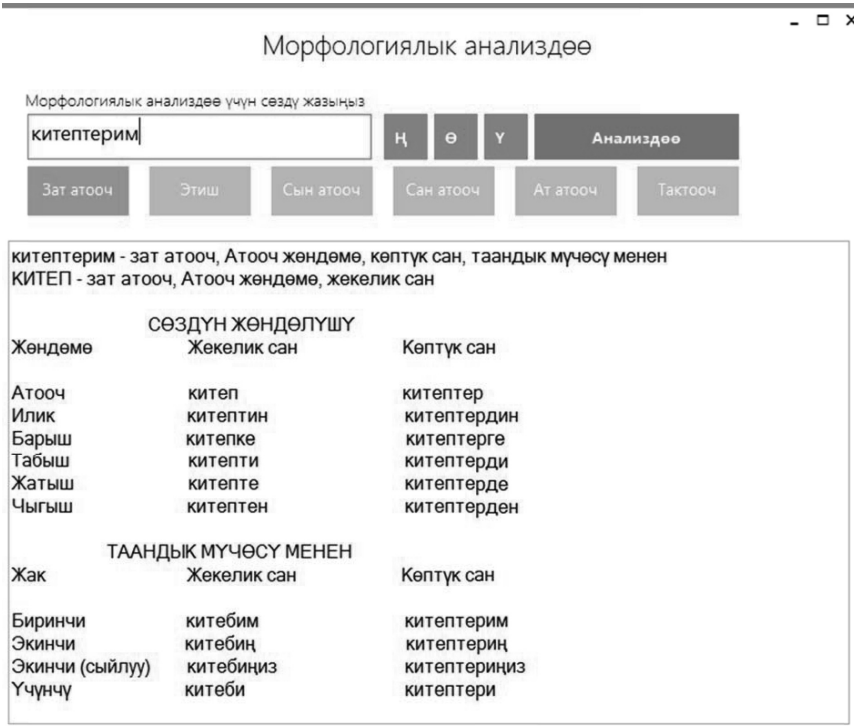


Рис. 7. Пример работы приложения

Для проверки правильности определения основы слов были протестированы существительные и глаголы с разными вариациями суффиксов и окончаний. В таблице 5 представлены примеры результатов тестирования слов.

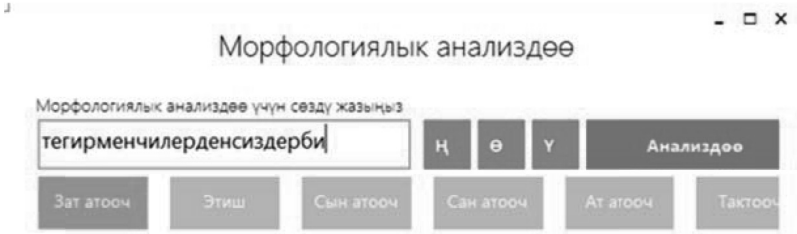
Таблица 5. **Результаты обработки**

Название словоформы	Аффиксы	Количество аффиксов	Часть речи
Тегирмен	*чи + лер + ди + ки + лер +ден + сиз + дер + би*	9	Сущ.
Табышмак	*тар + ыңыз + дар + дан + бы*	5	Сущ.
Кам	*сыз + дан + дыр + ыл + ба + ган + дык + тан*	8	Сущ.
Жаз	*дыр + а + сыңар +бы*	4	Глаг.
Оку	*ган + сыңар + бы*	3	Глаг.
Бекит	*кен + биз*	2	Глаг.

Была проведена проверка правильности преобразования словоформ. За исходные данные были взяты результаты тестирования алгоритма определения основы слова (таблица 5).

Пример синтеза существительных показан на рисунке 8.

Пример синтеза глаголов показан на рисунке 9.



тегирменчилерденсиздерби - зат атооч, Чыгыш жөндөмө, көптүк сан, суроо түрү, баяндоочтук категориясында
ТЕГИРМЕНЧИ - зат атооч, Атооч жөндөмө, жекелик сан

СӨЗДҮН ЖӨНДӨЛҮШҮ

Жөндөмө	Жекелик сан	Көптүк сан
Атооч	тегирменчи	тегирменчилер
Илик	тегирменчинин	тегирменчилердин
Барыш	тегирменчиге	тегирменчиоерге
Табыш	тегирменчини	тегирменчилерди
Жатыш	тегирменчиде	тегирменчилерде
Чыгыш	тегирменчиден	тегирменчилерден

ТААНДЫК МҮЧӨСҮ МЕНЕН

Жак	Жекелик сан	Көптүк сан
Биринчи	тегирменчим	тегирменчибиз
Экинчи	тегирменчиң	тегирменчиңер
Экинчи (сыйлуу)	тегирменчиңиз	тегирменчиңиздер
Үчүнчү	тегирменчиси	тегирменчиси

ТААНДЫК МҮЧӨСҮ МЕНЕН (Көптүк сан)

Жак	Жекелик сан	Көптүк сан
Биринчи	тегирменчилерм	тегирменчилерибиз
Экинчи	тегирменчилерң	тегирменчилериңер
Экинчи (сыйлуу)	тегирменчилериңиз	тегирменчилериңиздер
Үчүнчү	тегирменчилери	тегирменчилери

Рис. 8. Пример синтеза существительного

Рис. 9. Результат синтеза глагола

Рис. 9а. Результат синтеза глагола

Заключение

В процессе реализации морфологического анализатора кыргызского языка разработаны модули морфологического анализа и синтеза.

Проведено тестирование, в результате которого морфологический анализатор показал устойчивую работу на множестве слов с различными морфохарактеристиками.

ЛИТЕРАТУРА

1. Исраилова, Н.А. Алгоритм образования словоформ для автоматизации процедуры пополнения базы данных словаря [Текст] : / Н.А. Исраилова, П.С. Бакасова // – Известия КГТУ 2016. – Т. 38. – № 2. – С. 23–27.

2. Исраилова, Н.А. Организация синтаксического анализа в трансляторах [Текст] / Н.А.Исраилова // Вестн. Вост.-Казахст. гос. техн. ун-та им. Д. Серикбаева. – 2012. – №1. – С. 101–104.

3. Исраилова, Н.А. Организация морфологического анализа в трансляторах [Текст] / Н.А.Исраилова // Вестн. Вост.-Казахст. гос. техн. ун-та им. Д. Серикбаева. – 2012. – №1. – С. 97–101.

4. Исраилова, Н.А. Алгоритмы отладки процесса трансляции [Текст] / Н.А. Исраилова // Изв. Кырг. гос. техн. ун-та им. И.Раззакова. – 2011. – №22. – С. 278–279.

5. Исраилова, Н.А. Принципы организации морфологического анализатора в трансляторе [Текст] / Н.А.Исраилова // Изв. Кырг. гос. техн. ун-та им. И.Раззакова. – 2010. – №20. – С. 233–236.

6. Мальковский М. Г., Старостин А. С. Система морфосинтаксического анализа Treeton и мультиагентный синтаксический анализатор Treevial: принцип работы, система правил и штрафов // Екатеринбург: Изд-во Уральского университета, 2007. – С. 135–143.

7. Крылов С.А., Старостин С. А. Актуальные задачи морфологического анализа и синтеза в интегрированной информационной среде Starling // М.: Диалог, 2003.

УДК 81'33

MATERIALS FOR MORPHOLOGICAL ANALYSIS AND PRESENTATION OF THE SPACE IN THE KYRGYZ LANGUAGE

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

We describe both well-known and new linguistic and pedagogical problems that can be solved using modern computer technology. A unified algorithm for word change and other materials for automating morphological analysis in the Kyrgyz language are described. We show that unlike some other languages, in Kyrgyz language, spatial concepts are represented not primarily in the form of «relations» of objects, but in the definition of virtual regions defined both by gravitation and by the relative location and movement of the object (s) and subject. The result of the experiment is also presented. On this basis, clearer mathematical and computer models of such concepts can be constructed within the framework of an independent computer representation of the language.

Keywords: morphological analysis, Kyrgyz language, word inflection algorithm, spatial concepts, virtual domain, mathematical model, computer model.

МАТЕРИАЛЫ ДЛЯ МОРФОЛОГИЧЕСКОГО АНАЛИЗА И ПРЕДСТАВЛЕНИЕ ПРОСТРАНСТВА В КЫРГЫЗСКОМ ЯЗЫКЕ

П.С. Панков[1], С.Ж. Карабаева[2]
[1]Институт математики Национальной академии наук КР
[2]Кыргызский государственный университет строительства, транспорта и архитектуры
pps50@rambler.ru

Приведены как общеизвестные, так и сформулированные с участием авторов лингвистические и педагогические задачи, которые решаются с использованием современной компьютерной техники. Описан единый алгоритм словоизменения и другие материалы для автоматизации морфо-

логического анализа в кыргызском языке. Установлено, что, в отличие от некоторых других языков, в кыргызском языке пространственные понятия представляются преимущественно не в виде «отношений» объектов, а в определении виртуальных областей, определенных как гравитацией, так и взаимным расположением и движением объекта (объектов) и субъекта. Изложен результат эксперимента. На этой основе, могут быть построены более ясные математические и компьютерные модели таких понятий в рамках независимого компьютерного представления языка.

Ключевые слова: морфологический анализ, кыргызский язык, алгоритм словоизменения, пространственные понятия, виртуальная область, математическая модель, компьютерная модель.

Введение

Развитие современной компьютерной техники и мультимедийных технологий дает возможность решать для различных естественных языков следующие взаимосвязанные лингвистические и педагогические задачи. Некоторые из них общеизвестны, другие сформулированы и частично решены для кыргызского языка с участием авторов, см. [4-12].

1) Разработка и программная реализация единого алгоритма словообразования и словоизменения (для тюркских языков).

2) Автоматизация морфологического анализа (для тюркских языков это облегчается использованием п. 1).

3) Автоматический синтез слов из основы и первичных форм аффиксов в соответствующем порядке (с использованием п. 1).

4) Составление электронных словарей, приспособленных для использования п. 2 и п. 3.

5) Разработка математических моделей и компьютерных моделей понятий естественных языков в виде виртуальных сред, объектов и действий пользователя с ними, без использования других понятий; установление естественной иерархии понятий.

6) Разработка независимых интерактивных компьютерных представлений естественных языков для объективной фиксации состояния языка и изучения языка без языка-посредника.

7) Разработка мультимедийных интерактивных комплексных экзаменов со случайным формированием заданий по естественным языкам, в которых проверяется не только знание грамматики

и связей понятий с другими понятиями, но и связей понятий с реальными объектами и процессами.

В настоящей статье излагаются п. 1, а также материалы к п. 2 и п. 5 для пространственных понятий кыргызского языка.

1. Единый алгоритм словоизменения в кыргызском языке

В кыргызском языке слова изменяются и новые слова образуются путем добавления соответствующих окончаний (аффиксов) к основе слова, при этом сама основа, как правило, меняется редко. Кроме того, аффикс при его соединении с основой слова меняется так, что его звучание уподобляется звучанию последних букв (звуков) в основе слова (сингармонизм).

Отметим, что отдельные разделы данного алгоритма используются также при изучении темы “Алгоритмы” на занятиях по информатике в средних и высших учебных заведениях Кыргызстана.

1.1. Алфавит, краткие сведения о произношении букв и классификация букв

Алфавит современного литературного кыргызского языка состоит из 36 букв: А, Б, В, Г, Д, Е, Ё, Ж, З, И, Й, К, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, Ү, Ф, Х, Ц, Ч, Ш, Щ, Ъ, Ь, Э, Ю, Я.

Буква Ң означает звук “заднеязычное носовое Н”, буква Ө – “смягченное О”, буква Ү – “смягченное У”. Буква Ж в кыргызских словах означает слитный звук “ДЖ”, в словах, заимствованных из русского языка – звук “Ж”. В слогах с гласными Е, И, Ө, Ү, Э буква Л читается мягко “ЛЬ”. В слогах с гласными А, Ё, О, У, Ы, Ю, Я буквы К, Г означают заднеязычные звуки “КЪ”, “ГЪ”, буква Л и другие согласные читаются твердо. (Мы рассматриваем именно классификацию букв, а не звуков, в виде, необходимом для построения алгоритма).

Гласные в кыргызском языке подразделяются по следующим признакам. Каждая из «йотированных» букв Ё = Й*О, Ю = Й*У, Я = Й*А, Е = Й*Э входит в ту же группу, что и основная буква. Шесть из восьми «основных» гласных могут быть долгими (Таблица1).

Таблица 1. **Классификация гласных букв**

	переднерядные (твердые)		заднерядные (мягкие)	
	широкие	узкие	широкие	узкие
неогубленные (негубные)	А (Я)	Ы	Э Е	И
	АА		ЭЭ	
огубленные (губные)	О (Ё)	У (Ю)	Ө	Ү
	ОО (ЁО)	УУ (ЮУ)	ӨӨ	ҮҮ

Если не использовать йотированные буквы и не различать буквы Э, Е, то нижеприведенный алгоритм упрощается и становится более близким к фонетическому строю кыргызского языка, но мы исходим из официальных документов. В таблице 2 приведена классификация согласных букв.

Таблица 2. **Классификация согласных букв**

звонкие согласные	Б, В, Г, Д, Ж, З, Л, М, Н, Ң	
если отдельно не оговорено, то включаются в звонкие согласные		Й, Р
глухие согласные	К, П, С, Т, Ф, Х, Ц, Ч, Ш, Щ	

1.2. Описание единого алгоритма словоизменения

В большинстве Окончаний как широкие, так и узкие гласные могут иметь все четыре формы, перечисленные в таблице 1. В исходной форме такие окончания будут записываться с гласными А, Ы.

Окончания, в которых гласные могут быть только губными, будут записываться в исходной форме с гласными О, У, ОО, УУ.

Если согласная в начале Окончания может меняться, то в исходной форме они будут записываться со звонкими согласными Б, Г, Д, Л, Н.

Исходные данные для алгоритма: Основа слова и Окончание в исходной форме.

Есть несколько окончаний, которые не изменяются.

1-й шаг алгоритма. Анализ Основы слова и Окончания.

Находим последнюю букву (ПБ) и последнюю гласную (ПГ) в Основе слова.

Находим начальную букву (или две начальных буквы, если это двойная гласная) (НБ) в Окончании.

Если ПГ отсутствует или ПБ = ”Ъ” или ПБ = ”Ь”, то алгоритм выдает сообщение: “Это слово – не кыргызское” (в том смысле, что не может обрабатываться по правилам кыргызского языка) и останавливается.

(Если какой-либо случай ниже отдельно не оговорен, то: либо буква не изменяется, либо в кыргызском языке такая ситуация возникнуть не может.) Замену обозначаем «→». Пустое слово обозначаем “ ”.

2-й шаг алгоритма. В Основе Ё → ЙО, Ю → ЙУ, Я → ЙА.

3-й шаг алгоритма. Изменение последней буквы Основы (ПБ)

3.1) (“Озвончение”). Если НБ = гласная, то: если буква перед ПБ – гласная и (если ПБ = “К”, то ПБ → “Г”; если ПБ = “П” , то ПБ → “Б”).

3.2) Если НБ = долгая гласная и ПБ = гласная, то ПБ → “ ” (значение ПБ сохраняется).

3.3) Если ПБ = “П” и окончание = “ЫП”, то ПБ → ПГ.

4-й шаг алгоритма. Изменение начальной буквы Окончания (или двух начальных одинаковых гласных) (НБ).

4.1) Если выполнены условия 3.2), то НБ изменяется следующим образом (Таблица 3):

Таблица 3. **Изменение долгих губных гласных в начале Окончания**

НБ \ ПБ	А	Ы	О	У	Е	И	Ө	Ү
УУ	ОО	УУ	ОО	УУ	ӨӨ	ҮҮ	ӨӨ	ҮҮ
ОО	ОО	ОО	ОО	ОО	ӨӨ	ӨӨ	ӨӨ	ӨӨ

4.2) Если НБ = “А” и Окончание ≠ “АК” и ПБ = гласная, то НБ → “Й”. Если при этом возникают подряд две буквы “ЙЙ”, то “Й”→“ ”.

4.3) Если НБ = Окончание = “Ы” и ПБ = гласная, то Окончание→“СЫ”.

4.4) Если НБ = “Ы” и Окончание (в начальном виде) состояло из более чем одной буквы и ПБ = гласная, то НБ→“ ”.

4.5) Если НБ = “Б”, “Г”, “Д”, “Н”, “Л”, то она изменяется следующим образом (Таблица 4):

Таблица 4. **Изменение букв Б, Г, Д в начале Окончания**

НБ \ ПБ	гласная или звонкая согласная	глухая согласная
Б	Б	П
Г	Г	К
Д	Д	Т

Вводится индикатор ЙР: если ЙР = 0, то Л переходит в Д для всех звонких согласных; если ЙР = 1, то Л сохраняется после Й; если ЙР = 2, то Л сохраняется после Й и Р. Изменение букв Л, Н в начале Окончания показан в таблице 5.

Таблица 5. **Изменение букв Л, Н в начале Окончания**

НБ\ПБ	Гласная	звонкая согласная	глухая согласная
Н	Н	Д	Т
НБ \ПБ	гласная, иногда Й, Р	звонкая согласная	глухая согласная
Л	Л	Д	Т

5-й шаг алгоритма. После выполнения 4-го шага, последовательно, исходя из НБ в случае 3.1 и ПГ в остальных случаях, преобразуем гласные в Окончании, ка показано в таблице 6:

Таблица 6. **Изменение гласных букв в Окончании**

Предыдущая гласная / Текущая гласная	А, Я, Ы О (Ё) У (Ю) Э, Е, И Ѳ, Ѵ
А	А О А Е Ѳ
Ы	Ы У У И Ѵ
У	У У У Ѵ Ѵ

6-й шаг алгоритма. Соединяем преобразованные Основу и Окончание в Слово.

7-й шаг алгоритма. В получившемся Слове свертываем сочетания букв “Й” с гласными: Й А → Я; Й О →Ё; ЙУ → Ю; ЙЕ → Е; ЕЕ→ЭЭ.

8-й шаг алгоритма. Если Окончание неизменяемое, то оно присоединяется к Основе без изменений.

На этом алгоритм работу заканчивает с результатом – Словом. Оно может рассматриваться, как Основа для присоединения последующего Окончания, но при этом нужно учитывать исключения, перечисленные ниже.

1.3. Списки аффиксов, для которых имеют место исключения из правил

Ниже приведены система местоимений и аффиксов, связанных с местоимениями, а также ряд других аффиксов вместе с кратким упоминанием об их применении (Таблицы 7-8). Запись « « обозначает отсутствие окончания.

Таблица 7. **Личные местоимения, личные и притяжательные аффиксы**

Число \ Лицо	Единственное			Множественное		
1-е	-МЕН	-МЫН, -М	-ЫМ	-БИЗ	-БЫЗ	-ЫБЫЗ или -К
2-е	-СЕН	-СЫҢ	-ЫҢ	-СИЛЕР	-СЫҢАР	-ЫҢАР
3-е	-АЛ	« « *или* - Т	-Ы	-АЛАР	« « *или* - Т	-Ы
2-е вежл.	-СИЗ	-СЫЗ	-Ы ҢЫЗ	СИЗДЕР	-СЫЗДАР	-ЫҢЫЗДАР

Таблица 8. **Аффиксы повелительного наклонения**

Лицо \ Число	Единственное	Множественное
1-е	не существует	-АЛЫ
2-е	« « или -ГЫН	-ГЫЛА
3-е	-СЫН	-СЫН
2-е вежливое	-ЫҢЫЗ	-ЫҢЫЗДАР

Для склонения местоимений единственного числа имеют место исключения, перечисленные ниже.

ИЛИК ЖӨНДӨМӨ – родительный падеж: -НЫН.

Исключения: МЕН*НЫН=МЕНИН; СЕН*НЫН=СЕНИН; АЛ*НЫН=АНЫН;

БАРЫШ ЖӨНДӨМӨ – дательно-направительный падеж: -ГА.

Исключения: МЕН*ГА=МАГА; СЕН*ГА=САГА; АЛ*ГА=АГА;

после -ЫМ, -ЫҢ окончание -А; после -Ы, -НЫКЫ окончание -НА.

ТАБЫШ ЖӨНДӨМӨ – винительный падеж: -НЫ.

Исключения: МЕН*НЫ=МЕНИ; СЕН*НЫ=СЕНИ; АЛ*НЫ=АНЫ;

после -Ы, -НЫКЫ окончание: -Н.

ЖАТЫШ ЖӨНДӨМӨ – местный падеж: -ДА.

Исключения: АЛ*А=АНДА; после -Ы, -НЫКЫ окончание: -НДА;

ЧЫГЫШ ЖӨНДӨМӨ – исходный падеж: -ДАН.

Исключения: АЛ*ДАН=АНДАН; после -Ы, -НЫКЫ окончание -НАН;

КӨПТҮК САН – множественное число: -ЛАР (ЙР = 2).

Исключение: БАЛА*ЛАР = БАЛДАР.

2. Проведение морфологического анализа

Он может быть проведен в соответствии с [16]. Приведем порядок присоединения возможных аффиксов [12].

Глаголы:

Корень * (Образование глагола из другой части речи) * (Уточнение смысла глагола – словообразование)* (Стандартное уточнение смысла глагола – словоизменение)* (Отрицание) * (Время или наклонение) * (Личный или притяжательный) *(Вопросительный).

Существительные:

Корень* (Образование существительного из другой части речи)*

(Уточнение смысла существительного – словообразование)* (Множественное число)* (Притяжательный)* (Падеж)*(Личный)*(Вопросительный)

Другие части речи имеют мало аффиксов или не меняются.

Наиболее полный список аффиксов составлен нами [14]. Отметим еще, что программа [15] осуществляет морфологический анализ слов с одним аффиксом.

3. Специфика пространственных понятий в кыргызском языке

Здесь в качестве управляемого объекта используется беспилотный управляемый летательный аппарат (дрон), как способный выполнять наиболее произвольные пространственные перемещения. (При компьютерной реализации различных глаголов нами установлено, что многие глаголы, которые раньше относились к «одушевленным» объектам, теперь относят и к сложным техническим объектам, в том числе – к дронам). См. также [3], [13].

3.1. Список пространственных областей

Отметим, что границы областей приводимого далее списка наиболее определены для прямоугольных объектов. Для объектов других форм они вызывают сомнения у самих носителей языка, в соответствии с [1], как показывают проведенные нами эксперименты, см. ниже раздел 5.

Термины для обще-топологических областей

ИЧ – внутренность (как в кыргызском, так и в русском языке это слово имеет дополнительный анатомический смысл).

ЧЕК – граница, для преимущественно двумерных объектов (по отношению к некоторой области).

БЕТ – поверхность, то есть двумерная внешняя сторона границы для трехмерных объектов.

ЧЕТ – отдаление от данной точки (употребляется также в смысле «большое отдаление» – заграница).

СЫРТ – внешность (в таком смысле этот термин употребляется в работах по топологии на русском языке).

ЖАН – близкое пространство, или окрестность – часть СЫРТ.

(Имеется слово АЛЫС – антоним слову ЖАН, но оно используется не как существительное, а как прилагательное – далекий).

ТҮПКҮР – самое удаленное от входа или края (ЧЕТ) место.

ТЫШ – наружность; используется, в отличие от СЫРТ, как не связанное с каким-либо конкретным объектом.

Термины для областей, определяемых гравитацией

ҮСТ или ҮСТҮ – верхнее пространство; АСТ – нижнее пространство.

Термины для областей по отношению к субъекту или при движении объекта

ОҢ – правое пространство; СОЛ – левое пространство;

(при движении объекта эти области определяются с учетом гравитации)

АЛД, АЛДЫ – переднее пространство;

АРТ – заднее пространство (объекта по отношению к субъекту);

АРКА – заднее пространство (в редких случаях при движении объекта, а также когда человек говорит «за моей спиной, за твоей спиной»).

Термины для областей, определяемых несколькими объектами

АРА – пространство между двумя или более объектами (в аналогичном смысле также используются слова ИЧ и ОРТ – последнее слово ближе к слову «середина», также в различных смыслах).

3.2. Пространственно-временные послелоги с обобщающим смыслом

Для перехода от выражения «(отдельный) элемент множества» к выражению «(все) элементы множества», кроме термина «все», в различных языках используются различные термины и грамматические конструкции. В кыргызском языке имеются два послелога (ЖАНДООЧ):

ЖЕТЕ (чаще для пространственных областей); ДЕЙРЕ (чаще для временных интервалов) обозначают, в отличие от послелога ЧЕЙИН – до, «все» объекты или события до указанного.

3.3. Использование терминов для пространственных областей

Используются следующие схемы.

Название объекта в родительном падеже (ИЛИК ЖӨНДӨМӨ) и название связанной с ним области пространства с притяжательным окончанием и падежным окончанием одного из трех

геометрических падежей: дательно-направительного (БАРЫШ ЖӨНДӨМӨ), местного (ЖАТЫШ ЖӨНДӨМӨ), исходного (ЧЫГЫШ ЖӨНДӨМӨ). В квадратных скобках будем приводить дословный перевод.

Примеры повествовательных предложений.

ДРОН КӨПҮРӨНҮН ҮСТҮНДӨ / АСТЫНДА [дрон в верхнем / нижнем пространстве моста] – дрон над / под мостом;

ДРОН КӨПҮРӨНҮН ЖАНЫНДА [дрон в ближнем пространстве моста] – дрон около моста;

ДРОН КӨПҮРӨНҮН ҮСТҮНӨН / АСТЫНАН ЧЫГЫП ЖАТАТ [дрон выходит из верхнего / нижнего пространства моста] – дрон двигается так, чтобы не быть над мостом / дрон выходит из-под моста;

ДРОН ҮЙЛӨРДҮН АРАСЫНДА ЖЫЛЫП (БАРА) ЖАТАТ [дрон двигается в промежуточном пространстве домов] – дрон двигается между домами;

ДРОН ҮЙЛӨРДҮН АРАСЫНА КИРИП ЖАТАТ [дрон входит в промежуточное пространство домов] – дрон двигается так, чтобы оказаться между домами;

ДРОН ҮЙЛӨРДҮН АРАСЫНАН ЧЫГЫП ЖАТАТ [дрон выходит из промежуточного пространства домов] – дрон находится между домами и двигается так, чтобы не быть между домами.

Примеры команд для объекта.

Здесь приведем команды, обращенные непосредственно к дрону.

КӨПҮРӨНҮН ҮСТҮНӨ / АСТЫНА БАР! [иди в верхнее / нижнее пространство моста] – двигайся так, чтобы оказаться над / под мостом;

КӨПҮРӨНҮН ҮСТҮНӨН / АСТЫНАН ЧЫК! [выйди из верхнего /нижнего пространства моста] – двигайся так, чтобы не быть над мостом / выйди из-под моста;

БУЛУТТУН СЫРТЫНДА / ИЧИНДЕ ЖҮР! [двигайся во внешнем / внутреннем пространстве облака] – двигайся вне / внутри облака;

ҮЙЛӨРДҮН АРАСЫНДА ЖҮР! [двигайся в промежуточном пространстве домов] – двигайся между домами;

ҮЙЛӨРДҮН АРАСЫНА КИР! [войди в промежуточное пространство домов] – двигайся так, чтобы оказаться между домами;

ҮЙЛӨРДҮН АРАСЫНАН ЧЫК! [выходи из промежуточного пространства домов] – двигайся так, чтобы не быть между домами.

Примеры команд управляющему дроном.

Здесь на первом месте стоит имя управляемого объекта в винительном падеже (ТАБЫШ ЖӨНДӨМӨ).

ДРОНДУ КӨПҮРӨНҮН ҮСТҮНӨ / АСТЫНА КОЙ! [поставь дрон в верхнее / нижнее пространство моста] – двигай дрон так, чтобы он оказался над мостом / двигай дрон под мост;

ДРОНДУ КӨПҮРӨНҮН ҮСТҮНӨН / АСТЫНАН ЧЫГАР! [выведи дрон из верхнего / нижнего пространства моста] – двигай дрон так, чтобы он не остался над мостом / выведи дрон из-под моста.

ДРОНДУ КӨПҮРӨНҮН ҮСТҮНДӨ / АСТЫНДА ЖЫЛДЫР! [двигай дрон в верхнем / нижнем пространстве моста] – двигай дрон над / под мостом;

ДРОНДУ КӨПҮРӨНҮН СОЛУНДА / ОҢУНДА ЖЫЛДЫР! [двигай дрон в левом / правом пространстве моста] – двигай дрон слева / справа от моста (по отношению к тому, кто смотрит и управляет дроном).

3.4. Измерение расстояний в пространстве временными интервалами

В кыргызском языке, как и в других тюркских языках, возможно измерение расстояний временными интервалами, например:

ТҮШТҮК ЖОЛ – [полуденный путь] – среднее расстояние, которое может проехать всадник от зари до полудня;

КҮНДҮК ЖОЛ – [дневной путь] – среднее расстояние, которое может проехать всадник от зари до вечера;

«... БЕШ КҮНДҮК ЖЕРГЕ УГУЛДУ « – [… было слышно на расстоянии пяти дней пути] (эпос «Манас»).

ҮЧ АЙЛЫК ЖОЛ – [трехмесячный путь] – (отличается от предыдущих примеров) – расстояние, которое проходит караван за три месяца (с остановками на ночлег и т.д.).

Также используется слово КЕРЕ (полностью):

КЕРЕ КҮНДҮК ЖОЛ – [максимальный дневной путь] – максимальное расстояние, которое может проехать всадник от зари до вечера (эпос «Манас»).

4. Виртуальные геометрические объекты, создаваемые глаголами

Известно понятие валентности глаголов (см. например [2]) – минимальное количество объектов, с которыми связан глагол в предложении. Как обобщение, мы предлагаем – минимальное количество реальных обстоятельств и объектов, при наличии которых носитель языка признает применение глагола корректным. Это необходимо для построения независимого интерактивного компьютерного представления глагола.

Глагол БАР – целенаправленно двигайся (иди и т.д.), удаляясь (*или* независимо) от говорящего. Подразумевается пространство для движения. Используется с дательно-направительным падежом (БАРЫШ ЖӨНДӨМӨ), возможно также указание среды передвижения с послелогом МЕНЕН. Подразумевается набор траекторий от данной точки до объекта назначения, (при отсутствии препятствий) не очень отклоняющихся от кратчайшего пути.

ШААРГА БАР! – (при значительном расстоянии) – поезжай в город.

Здесь обычно подразумевается единственная траектория от данной точки до объекта назначения.

Глагол БАРГЫЗ – понудительный глагол от глагола БАР. Используется также с винительным падежом (ТАБЫШ ЖӨНДӨМӨ), предполагает наличие объекта, который может передвигаться самостоятельно.

КИШИНИ ШААРГА БАРГЫЗ! – пошли человека в город.

При наличии препятствий больше употребляется глагол ЖЕТ – добирайся, достигай. Здесь носителем языка признаются траектории, близкие к минимальной не по длине, а по времени. Его понудительная форма глагол ЖЕТКИР- отличается от БАРГЫЗ- тем, что необходимо сопровождение объекта.

ООРУЛУУНУ ШААРГА ЖЕТКИР! – доставь *или* сопроводи больного в город.

КАЙПЫ – двигайся близко от, слегка коснись поверхности. Здесь подразумевается БЕТ – поверхность (двумерная) и ЖАН – ее окрестность, где происходит движение.

КАЙЫП УЧУУ – бреющий полет.

Глагол КЕЛ – двигайся сюда, по направлению к говорящему.

КЕЛ! – иди ко мне. *ҮЙГӨ КЕЛ!* – подойди к дому.

Здесь носителем языка признаются траектории, близкие к минимальной по длине, а при наличии препятствий – по времени.

Глагол ЧЫК – выходи. Подразумевается наличие двух пространственных областей: ИЧ – внутренность и СЫРТ – внешность. Используется с исходным падежом, названным по имени этого глагола (ЧЫГЫШ ЖӨНДӨМӨ).

ҮЙДӨН ЧЫК! – выйди из дома.

Здесь носитель языка требует исходное положение во «внутренности» объекта и заключительное положение во «внешности» объекта.

Глагол ЧЫГАР – понудительный от глагола ЧЫК.

МЫШЫКТЫ ҮЙДӨН ЧЫГАР! – выведи *или* заставь кошку выйти из дома.

Глагол КИР – входи. Также подразумевается наличие областей: ИЧ – внутренность и СЫРТ – внешность. Используется с дательно-направительным падежом.

ҮЙГӨ КИР! – войди в дом.

Здесь носитель языка требует исходное положение во «внешности» объекта и заключительное положение во «внутренности» объекта. Глагол КИРГИЗ – понудительный от глагола КИР.

МЫШЫКТЫ ҮЙГӨ КИРГИЗ! – введи *или* заставь кошку войти в дом.

Глагол ӨТ – близок по смыслу к глаголу «проходить» (геометрически). Также подразумевается наличие областей: 1) СЫРТ – внешность, 2) АРА – пространство между, 3) и 4) – соединения между 1) и 2), и передвижение в порядке 1)-3)-2)-4)-1).

КӨПҮРӨДӨН ӨТ! – перейди мост. *ҮЙЛӨРДҮН АРАСЫНАН ӨТ!* – [пройди пространство между домов] – пройди между домами.

Глагол КОН – абстракция от «приземлиться», «приводниться», «остановиться на ночлег» и т.д. Его можно также использовать в современном смысле *АЙГА КОН!* – прилунись, сделай посадку на Луну. Подразумевает наличие пространства, в котором возможно движение, и «более твердого, опорного» объекта.

Глаголы, связанные с явлением гравитации.

Глагол КӨТӨР – поднимай (например, гирю), носи поднятое (абстрактно: действуй против силы тяготения), в отличие от глагола ТУРГУЗ – придай вертикальное, стоячее положение *или* заставь встать.

Глагол ТӨМӨНДӨ – спустись, снизься.

Глаголы, связанные с прозрачностью пространства.

Глагол КӨРСӨТ – покажи (дословно: сделай так, чтобы (другой) увидел). Подразумевает наличие пространства, в котором возможно движение, и прозрачной зоны в пространстве, где действует зрение субъекта.

ОЮНЧУКТУ БАЛАГА КӨРСӨТ! – покажи игрушку ребенку (перемести игрушку в поле зрения ребенка).

ЖАП – закрой (в том числе – поставь преграду лучам света);

АЧ – открой (в том числе – убери преграду лучам света).

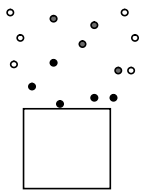
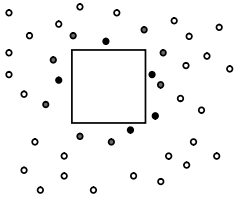
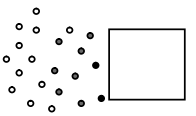
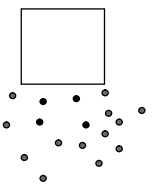
Глаголы ТАП – найди, ИЗДЕ – ищи. Подразумевают наличие непрерывного пространства или дискретного набора объектов.

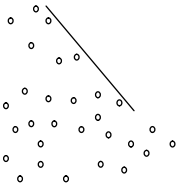
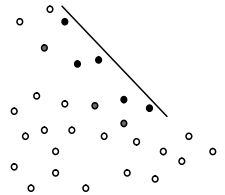
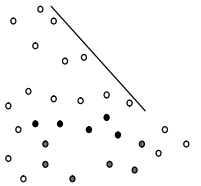
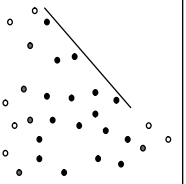
Глагол ИРГЕ – выяви наибольший из однородных твердых предметов с использованием синергетики. Здесь подразумевается большое (более 50) количество предметов, а также емкость, в которой их можно повергнуть вибрации.

Глагол СЫЗ – проведи черту (линию). Здесь геометрический объект уже заложен в глаголе.

5. Результаты экспериментов по определению пространственных областей с носителями языка

Участникам эксперимента предлагались рисунки нижеприведенных типов (все точки были неотмеченные) и предлагалось отметить точки, которые, по их мнению, удовлетворяют приведенному высказыванию.

			
1) Чекит чарчынын үстүндө…	*2) Чекит чарчынын жанында…*	*3) Чекит чарчынын солунда…*	*4) Чекит чарчынын астында…*

			
4а) Чекит кесиндинин астында…	*4а.1*	*4а.2*	*4а.3*

…деген сүйлөмгө дал келген чекиттерди белгилеңиз.

В приблизительном переводе: Отметьте точки, которые удовлетворяют высказыванию …

1) Точка над …	*2) Точка около …*	*3) Точка слева …*	*4) Точка под …*

На рисунках черным отмечены точки, которые отметили большинство участников эксперимента, серым – некоторые из участников.

Из результатов эксперимента можно сделать следующие выводы.

1. В общем подтвердилась идея Л.Заде [1] о нечетких множествах. Ответы, данные различными участниками эксперимента, очень различны, по «большинству» заданий у «большинства» участников ответы близки. Вместе с тем, по некоторым заданиям получены ответы, не укладывающиеся в схему нечетких множеств.

2. По отдельным заданиям:

– для округлых фигур (квадрат, круг): носители языка подсознательно исходят из числа, которое в математике называется «диаметром» D фигуры – максимальное расстояние между двумя точками фигуры. Под «ближним пространством, окрестностью» носители языка понимают множество точек, находящихся на расстоянии от фигуры, меньшем γD, где γ – некоторое положительное число. У разных участников эксперимента это число было от *1/5* до *1/3*;

– для фигур, имеющих ширину по горизонтали (квадрат, круг, горизонтальный отрезок), некоторые носители языка считают, что в понятия «верхнее пространство», «нижнее пространство» входит «окрестность» (приблизительно *D/2*), а другие носители языка считают, что эти области простираются на большие расстояния, при этом не было «промежуточных» результатов;

– некоторые носители языка считают, что «левое пространство» фигуры отмеряется от крайней левой точки фигуры, другие – от всех точек фигуры, отсюда возникло большое различие;

– для фигур, не имеющих ширину по горизонтали (вертикальный отрезок), некоторые носители языка считают, что «верхнее пространство», «нижнее пространство» отмеряются строго по вертикали, другие считают, что это – некоторые конусы от верхней и нижней точек;

– в 4а) получены три типа различных ответов (поэтому они изображены на отдельных рисунках) – некоторые носители языка считают, что «нижнее пространство» фигуры отмеряется от самой нижней точки фигуры, другие – от всех точек фигуры, отсюда возникли большие различия.

Таким образом, наблюдения ребенка над употреблением пространственных понятий на родном для него языке, кроме понятия «окрестности», вследствие разнородности случаев, не приводят к каким-либо определенным математическим выводам.

Явление, обнаруженное в 4а), названо нами «дискретизацией понимания пространственных понятий носителями языка».

Это следует учитывать при написании различных инструкций.

ЛИТЕРАТУРА

1. Zadeh L. A. The concept of a linguistic variable and its application to approximate reasoning // Information Sciences, 1975, Vol. 8, pp. 199–249, 301–357; Vol. 9, pp. 43–80.

2. Retz-Schmidt G. Various Views on Spatial Prepositions // Al Magazine, 1988, Volume 9, Number 2, pp. 95–105.

3. Ысакулов К. Английские предлоги со значением пространства и их функциональные соответствия в кыргызском языке. Автореф. дисс. … филолог. наук. – Бишкек, 1995. – 20 с.

4. Pankov P., Dolmatova P. Algorithmic Language and Classification of Verbs for Computer-Based Presentation // Academic Review, 2008, No. 1 (7). – American University of Central Asia, Bishkek, pp. 233–239.

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

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

7. Bayachorova B.J., Pankov P.S. Independent Computer Presentation of a Natural Language. In: Varia Informatica. Lublin, Polish Information Processing Society, 2009. – Pp. 73–84.

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

9. Karabaeva S., Dolmatova P. Mathematical and computer models of spatial relations in Kyrgyz language // Proceedings of V Congress of the Turkic World Mathematicians. – Bishkek: Kyrgyz Mathematical Society, 2012, pp. 175–178.

10. Karabaeva S. Peculiarities of spatial relations in Kyrgyz language // Abstracts of the Issyk-Kul International Mathematical Forum. – Bishkek: Kyrgyz Mathematical Society, 2015, p. 79.

11. Karabaeva S. Presentation of spatial-temporal relations in Kyrgyz language // Труды международной конференции по компьютерной и когнитивной лингвистике TEL-2016. – Казань: Изд-во Казан. ун-та, 2016. – С. 274–277.

12. Панков П.С. Обучающая и контролирующая программа по словоиз-менению в кыргызском языке на ПЭВМ. – Бишкек: Мектеп, 1992. – 20 с.

13. Сулейманов Д.Ш., Гатиатуллин А.Р., Карабаева С.Ж. К разработке функционально-структурной модели аффиксальных морфем языков тюркской группы (на примере киргизского и татарского языков) //Казанская наука, 2013, № 6. – Казань: Изд-во Казанский Издательский Дом, 2013. – 194 с. С. 120–122.

14. Карабаева С. Единый алгоритм словоизменения и представление пространства в кыргызском языке. – Saarbrűcken, Deutschland: Lap Lambert Academic Publishing, 2016. – 62 с.

15. Чодоев Р. http://tamgasoft.kg/morfo/ky

16. Дыбо А.В., Шеймович А.В. Автоматический морфологический анализ для корпусов тюркских языков // Филология и культура, 2014. – Выпуск № 2 (36). – С. 20–26.

УДК 81'33

MODEL OF MORPHOLOGICAL ANALYSIS OF THE KYRGYZ LANGUAGE

T. Sadykov[1], B. Kochkonbayeva[2]
[1]K. Karasayev Bishkek Humanitarian University, Bishkek, Kyrgyzstan
[2]Osh Technological University, Osh, Kyrgyzstan
tash_sadykov@mail.ru, buajar@mail.r

The article describes the issues related to the construction of the morphological analysis model for the Kyrgyz language.

Keywords: Morphological Analysis, Kyrgyz Language.

МОДЕЛЬ МОРФОЛОГИЧЕСКОГО АНАЛИЗА КЫРГЫЗСКОГО ЯЗЫКА

Т. Садыков[1], Б. Кочконбаева[2]
[1]Бишкекский гуманитарный университет им. К.Карасаева, Бишкек
[2]Ошский технологический университет, Ош
tash_sadykov@mail.ru, buajar@mail.ru

В статье рассматриваются вопросы, связанные с построением модели морфологического анализа для кыргызского языка.

Ключевые слова: морфологический анализ, кыргызский язык.

1. Морфология кыргызского языка

Морфология – это часть грамматики, функциональное назначение которой состоит в том, чтобы обеспечить процесс анализа и синтеза словоформ в тексте. С этих позиций модель морфологического анализа языка, как минимум, должна содержать информацию о перечне морфологических категорий и способах их выражения.

2. Морфологические категории

Их перечень определяется следующими категориями:

2.1. Имя существительное – Noun:

Категория числа – Number
1. Единственное число – singular
2. Множественное число – plural
Тэги:
1. **SG <=>**∅
∅
2. **PL <=> ЛАр**
-лар -дар -тар
-лер -дер -тер
-лор -дор -тор
-лөр -дөр -төр.

Категория притяжательности – Possessive
Единственное число – singular:
1. первое лицо единственного числа – 1st person singular possessive (‘my’),
2. второе лицо единственного числа – 2ndpersonsingularpossessive (‘your’),
3. второе лицо единственного числа ласк. – 2nd person sing.po**ss.** formal (‘your’),
4. третье лицо единственного числа – 3nd person singular possessive (‘his/her/its’),
Множественный падеж – plural:
5. первое лицо множественного числа – 1st person plural possessive (‘our’),
6. второе лицо множественного числа – 2nd person plural possessive (‘your’),
7. второе лицо множественного числа ласк. – 2nd person pl.po**ss.** formal (‘your’),
8. третье лицо множественного числа – 3nd person plural possessive (‘their’).
Тэги:

1. **POSS_1SG <=> [Ы]м**
-ым -им -ум -үм
-м;
2. **POSS_2SG <=> [Ы]ң**
-ың -иң -уң -үң
-ң;
3. **POSS_2SGF <=> [Ы]ң[Ы]з**
-ыңыз -иңиз -уңуз -үңүз
-ңыз -ңиз -ңуз -ңүз;
4. **POSS_3SG <=> [с]Ы[н]**
-ы -и -у -ү -ын -ин -ун -үн
-сы -си -су -сү -сын -син -сун -сүн;
5. **POSS_1PL <=> [Ы]б[Ы]з**
-ыбыз -ибиз -убуз -үбүз
-быз -биз -буз -бүз;
6. **POSS_2PL <=> [Ы]ң[А]р**
-ыңар -иңер -уңар -үңөр
-ңар -ңер -ңар -ңөр;
7. **POSS_2PLF <=> [Ы]ң[Ы]зд[А]р**
-ыңыздар-иңиздер -уңуздар -үңүздөр
-ңыздар -ңиздер -ңуздар -ңүздөр;
8. **POSS_3PL <=> [с]Ы[н]**
-ы -и -у -ү
-ын -ин -ун -үн
-сы -си -су -сү
-сын -син -сун -сүн;

Категория падежа – Noun Cases
1. Именительный падеж – nominative.
2. Родительный падеж – genitive.
3. Дательный (направительный) падеж– dative.
4. Винительный падеж – accusative.
5. Местный падеж– locative.
6. Исходный падеж– ablative.

Тэги:
1. NOM<=>∅
∅

2. GEN <=> [н]Ын

-нын -нин -нун -нүн
-дын -дин -дун -дүн
-тын -тин -тун -түн
-ын -ин -ун -үн;

3. DAT<=>[Г]А

-га -ге -го -гө
-ка -ке- ко –кө
-а -е- о -ө;

4. ACC<=> [н][Ы]

-ны -ни -ну -нү
-ды -ди -ду -тү
-ты -ти -ту -тү
-ы -и -у -ү;

5. LOC<=>ДА

-да -де -до -дө
-та -те -то -тө;

6. ABL<=> [Д]Ан

-дан -ден -дон -дөн
-тан -тен -тон -төн
-ан -ен -он -өн;

Категория лица – Personal

Единственное число – singular:

1. первое лицо единственного числа – 1stpersonsingular,
2. второе лицо единственного числа – 2ndpersonsingular,
3. второе лицо единственного числа ласк. – 2ndpersonsingularformal,
4. третье лицо единственного числа – 3ndpersonsingular,

Множественное число – plural:

5. первое лицо множественного числа – 1stpersonplural,
6. второе лицо множественного числа – 2ndpersonplural,
7. второе лицо множественного числа – 2ndpersonpluralformal,
8. третье лицо множественного числа – 3ndpersonplural,

Тэги:

1. **1SG<=>м[Ы]н**

-мын -мин -мун -мүн;

2. **2SG <=> с[Ы]ң**
-сың -сиң -суң -сүң;
3. **2SGF <=> с[Ы]з**
-сыз -сиз -суз -сүз;
4. **3SG <=>∅**
∅;
5. **1PL <=> б[Ы]з**
-быз -биз -буз -бүз;
6. **2PL <=> с[Ы]ң[А]р**
-сыңар -сиңер -суңар -сүңөр;
7. **2PLF <=> с[Ы]зд[А]р**
-сыздар-сиздер -суздар -сүздөр;
8. **3PL <=> [с]Ы[н]**
∅;

2.2. Имя прилагательное – adjective:

Сравнительная степень – comparative
Тэги:
COMP <=> [Ы]рААК
-ыраак -ирээк -ураак -үрөөк
-раак -рээк -раак -рөөк;

2.3. Имя числительное – Numeral:

1. Порядковое числительное – ordinalnumeral
2. Собирательное числительное – collective numeral,
3. Приблизительное числительное1 – approximatenumeral1,
4. Приблизительное числительное2 – approximatenumeral2,
5. Приблизительное числительное3 – approximatenumeral3,

Тэги:
1. NUM_ORD <=> [Ы]нчЫ
-ынчы -инчи -унчу -үнчү
-нчы -нчи -нчу -нчү;
2. NUM_COLL <=> ОО[н]
-оо -өө
-оон -өөн;

3. NUM_APPR1 <=> чА
-ча -че -чо -чө;
4. NUM_APPR2 <=> ДАй
-дай -дей -дой -дөй
-тай -тей -той -төй;
5. NUM_APPR3 <=> ДАгАн
-даган -деген -догон -дөгөн
-таган -теген -тогон -төгөн.

2.4. Глагол – Verb:

Категория залога – Voices
1. Основной залог – active.
2. Страдательный залог– passive.
3. Рефлексивный залог – reflexive.
4. Посредственный залог – causative.
5. Возвратный залог – reciprocal.

Тэги:
1. ACT<=>∅
∅
2. ACT <=> [Ы]л||н
-ыл -ил -ул -үл
-л;
-ын -ин -ун -үн
-л;
3. REFL<=>[Ы]н
-ын -ин -ун -үн
-л;
4. CAUS <=>Д[Ыр]
5. RECP <=> [Ы]ш
-ыш -иш -уш -үш
-ш;

Категория наклонения – Verb moods:

Повелительное наклонение – Imperatives
1. Первое лицо единственное число – Hortative: 1st person singular – ‘let me’.

2. Первое лицо множественное число – Hortative: 1st person plural – ‘let’s’.

3. Второе лицо единственное число ласк. – Imperative: 2nd person singular.

4. Второе лицо множественное число ласк. – Imperative: 2nd person plural.

5. Второе лицо единственное число форм.– Imperative: 2nd person singular formal.

6. Второе лицо множественное число форм. – Imperative: 2nd person plural formal

7. Третье лицо единственное число – Jussive: 3rd person singular – ‘let him/her/it’.

8. Третье лицо множественноечисло – Jussive: 3rd person plural – ‘let them’.

9. Просительный вежл. – precative (‘please’).

Тэги:

1. HOR_SG <=> [А]йЫн
-айын -ейин -ойун -өйүн
-йын -йин -йун -йүн;

2. HOR_PL <=> [А||й]лЫ[к]
-алык -елик -олук -өлүк
-йлык -йлик -йлук -йлүк;
-алы -ели -олу -өлү
-йлы -йли -йлу -йлү;

3. IMP_SG <=> ГЫн
-гын -гин -гун -гүн
-кын -кин -кун -күн;

4. IMP_ PL <=> ГЫлА
-гыла -гила -гула -гүла
-кыла -кила -кула -күла;

5. IMP_SGF <=> [Ы]ңЫз
-ыбыз -ибиз -убуз -үбүз
-быз -биз -буз -бүз;

6. IMP_PLF <=> [Ы]ңЫздАр
-ыңыздар-иңиздер -уңуздар -үңүздөр
-ңыздар -ңиздер -ңуздар -ңүздөр;

7. JUS_SG <=>сЫн
-сын -син -сун -сүн;

8. JUS_PL <=> [Ыш]сЫн
-ышсын -ишсин -ушсун -үшсүн
-сын -син -сун -сүн;
9. PREC_1 <=>чЫ
-чы -чи -чу -чү;

Категория времени – Verb tenses
1. Настоящее время – present.
2. Определенное прошедшее время – past definite.
3. Неопределенное прошедшее время – past indefinite.
4. Неожиданное прошедшее время – past evidentiality.
5. Обыкновенное прошедшее время – past iterative.
6. Определенное будущее время – future definite.
7. Неопределенное будущее время – future indefinite.
8. Неопределенное отрицательное будущее время – future indefinite negative.

Тэги:
1. **PRES<=> [А||й]**
-а -е -о -ө
-й;
2. PST_DEF <=> ДЫ
-ды -ди -ду -дү
-ты -ти -ту -тү;
3. PST_INDF<=> ГА[н]
-ган -ген -гон -гөн
-кан -кен -кон -көн;
-га -ге -го -гө
-ка -ке -ко -кө;
4. PST_EVID <=> чУ
-чу -чү;
5. PST_ITER <=>[Ы]п[тЫр]
-ыптыр -иптир -уптур -үптүр
-птыр -птир -птур -птүр;
-ып -ип -уп -үп
-п;
6. FUT_DEF <=> [А||й]
-а -е -о -ө
-й;

7. FUT_INDF<=>[A]p
-ар -ер -ор -өр
-р;
8. FUT_INDF_NEG <=> БАс
-бас -бес -бос -бөс
-пас -пес -пос -пөс;

Аспект – aspect
1. Отрицательный – negative,
2. Вопросительный – interrogative.

Тэги:
1. NEG <=> БА
-ба -бе -бо -бө
-па -пе -по -пө;
2. INT<=> БЫ
-бы -би -бу -бү
-пы пи -пу -пү.

Причастия – Participles
1. Причастие настоящего времени – present participle.
2. Причастие прошедшего времени – past participle.
3. Причастие будущего времени – future participle.
4. Причастие отрицания будущего времени – future participle negative.

Тэги:
1. PCP_PR <=> [УУ]чУ
-уучу -үүчү
-чу -чү;
2. PCP_PS <=> ГАн
-ган -ген -гон -гөн
-кан -кен -кон -көн;
3. PCP_FUT_DEF<=> [A]p
-ар -ер -ор -өр
-р;
4. PCP_FUT_ NEG <=> БАс
-бас -бес -бос -бөс
-пас -пес -пос -пөс;

Деепричастия – Converbs

1. Сопровождающие деепричастия – Adverbial verb (accompanist).
2. Долгие деепричастия – Adverbial verb (continuing).
3. Целевые деепричастия – Adverbial verb (intentional).
4. Отрицательная форма деепричастия – Adverbial verb (negative form).
5. Последовательные деепричастия – Adverbial verb (successive meaning).
6. Ограничительные деепричастия – Adverbial verb (limiting).

Тэги:

1. ADVV_ACC <=> [Ы]п
-ып -ип -уп -үп
-п;

2. ADVV_CONT <=> [А||й]
-а -е -о -ө
-й;

3. ADVV_INT <=> ГАнЫ
-ганы -гени -гону -гөнү
-каны -кени -кону -көнү;

4. ADVV_NEG <=> мАйЫн[чА]
-майынча -мейинче -мойунча-мөйүнчө;
-майын -мейин -мойун -мөйүн;

5. ADVV_SUC <=> ГЫчА
-гыча -гиче -гуча -гүчө
-кыча -киче -куча -күчө;

6. ADVV_ SUC <=> ГАнчА
-ганча -генче -гончо -гөнчө
-канча -кенче -кончо -көнчө.

Отглагольное существительное – Verbal nouns (masdars)

1. Отглагольное существительное на -оо – infinitive 1.
2. Отглагольное существительное на -уу – infinitive 2.
3. Отглагольное существительное на -ыш – infinitive 3.
4. Отглагольное существительное на -мак – infinitive 4.
5. Отглагольное существительное на -гы – infinitive 5.

Тэги:

1. INF_1 <=> OO
-оо -өө;

2. INF_2 <=> УУ
-уу -үү;
3. INF_3 <=> [Ы]ш
-ыш -иш -уш -үш
-ш;
4. INF_4 <=> мАГ
-мак -мек -мок -мөк
-маг -мег -мог -мөг;
5. INF_5 <=> ГЫ
-гы -ги -гу -гү
-кы -ки -ку -кү;

Модальные формы – Modal forms
1. Условно модальные – conditional.
2. Намеренно модальные – desiderative (intention).
3. Желанно модальные – optative1.
4. Желанно модальные – optative2.
5. Предостерегающие модальные – premonitive (warning).

Тэги:
1. COND <=> сА
-са -се -со -сө;
2. DESIDE <=> мАк[чЫ]
-макчы -мекчи -мокчу -мөкчү
-мак -мек -мок -мөк;
3. OPT <=> ГЫ+POSS келет||келди
-гы -ги -гу -гү
-кы -ки -ку -кү;
4. OPT <=> ГАй эле+PERS
-гай -гей -гой -гөй
-кай -кей -кой -көй;
5. PREM <=> БАГАй эле+PERS
-багай -бегей -богой -бөгөй
-пагай -пегей -погой -пөгөй;

3. Модель грамматической формы слов

Грамматическая форма слов кыргызского языка формируется правилами производства агглютинационной цепочки, которая может состоять из корня или основы и словоизменительных аф-

фиксов. Модель грамматической формы слов можно представить так:

$S=R+U_m$ (1); где S – словоформа, R – корень или основа слова, U_m – словоизменительные аффиксы.

Как видим, грамматическая форма слова, или словоформа в кыргызском языке, зависит от корня или основы и словоизменительных аффиксов.

Цепочка словоизменительных аффиксов может достичь до восьми показателей, т.е. $U_m=U_{m1}+U_{m2}+...+U_{m8}$ (2).

Правило 1: Если $U_m=\varnothing$, то S функция будет равна корню или основе слова, и вводимое слово, не членяясь на морфемы, соответствует лексеме – единицу словарной базы языка.

Словоизменительные аффиксы.

Словоизменительные аффиксы изменяют форму слова и выражают его грамматическое значение, но не изменяют лексическое значение.

Группируя перечень морфологических категорий во множества выражающих их аффиксов, получим следующее:

CAS={GEN,DAT,ACC,LOC,ABL} – множество тегов категории падежа;

POSS={POSS_1SG, POSS_2SG, POSS_2SGF, POSS_3SG, POSS_1PL, POSS_2PL, POSS_2PLF, POSS_3PL} – множество тегов категории принадлежности;

NUM={PL} – множество категории числа;

PERS={1SG, 2SG, 2SGF, 3SG, 1PL, 2PL, 2PLF, 3PL} – множество тегов категории лица;

TENS={PRES, PST_DEF, PST_INDF, PST_EVID, PST_ITER, FUT_DEF, FUT_INDF, FUT_INDF_NEG} – множество тегов категории времени;

MOOD={HOR_SG, HOR_PL, IMP_SG, IMP_PL, IMP_SGF, IMP_PLF, JUS_SG, JUS_PL, PREM} – множество тегов категории наклонения;

ASP={NEG, INT} – множество тегов аспекта глагола.

Таким образом, если U_m – это множество словоизменительных аффиксов, то он состоит из следующих компонентов:

U_m={CAS, POSS, NUM, PERS,TENS, MOOD, ASP}.

Правила вхождения аффиксов в именную слоформу.

Именные словоформы – это общее название грамматических форм существительных, прилагательных, числительных и местоимений.

Правила вхождения аффиксов в именную слоформу представлены в виде следующей таблицы (табл. 1).

Таблица 1

0	1	2	3	4	5
Именные основы: *ат, эт, от, өт, ата, казан, жүрөк* и др.	*-лар*	*-ым* *-ың* *-ы* *-ыбыз* *-ыңар* *-ныкы*	*-нын* *-га* *-ны* *-да* *-дан* *-сыз* *-сыздар* *-ыңыз* *-ыңыздар*	*-мын* *-м* *-быз* *-сың* *-сыңар*	*-бы* *-чы*

Правила вхождения аффиксов в глагольную слоформу.

Порядок вхождения аффиксов в глагольную слоформу представлены в виде следующей таблицы (табл. 2).

Таблица 2

0	1	2	3	4	5	6
Глагольные основы: *ал-, уч-, кел-, жаз-, иште-, сугар-, оку-* и др.	*-ын* *-ыл* *-ыш* *-тыр* *-кар* *-кыр* *-каз* *-ар* *-ыз* *-сөт* *-т*	*-ба*	*-ып* *-а /* *-й*	*-ды* *-ган* *-чу* *-ып* *-ыптыр* *-ар/-бас* *-гын* *-гыла* *-сын* *-гай* *-мак* *-макчы* *-са*	*-мын* *-м* *-быз* *-сың* *-сыңар* *-сыз* *-сыздар* *-ыңыз* *-ыңыздар* *-т*	*-бы* *-чы*

4. Правила сингармонизма

Кыргызский язык, впрочем, как и любой другой язык тюркской группы, подчиняется правилам сингармонизма, которые в равной мере распространяются как на моно, так и полисиллабические словоформы. То есть в агглютинативной словоформе корень слова, определяя тембровую окраску последующих аффиксов, как правило, сохраняется без изменения, а аффикс, изменяясь в рамках действия сингамонизма, варьируется набором определенных вариантов. Ср., например, аффикс родительного падежа, который в контексте словоформы реализуется набором, состоящим из 16 сингармонических вариантов: *-нын, -нин, -нун, -нүн; -дын, -дин, -дун, -дүн; -тын, -тин, -тун, -түн*; *-ын, -ин, -ун, -үн*.

Правила сингармонизма, а также процедуры изменения начальных согласных аффиксов формализованы в специальных работах [2].

5. Алгоритм морфологического анализа кыргызского языка

Соотествующий алгоритм представлен на рис. 1-2.

Можно рассмотреть 3 ступени проведения морфологического анализа:

1. Определение только грамматического значения слова.
2. Определение только основы слова.
3. Определение грамматического значения и основы слова.

Развернутое или неполное исследование морфологического разбора зависит от поставленной задачи.

Морфологический анализ является начальной ступенью различных задач, связанных с естественным языком, и поэтому его точное выполнение имеет большое значение.

Методы морфологического анализа можно разделить на 3 типа:

– анализировать со словарем аффиксов;
– анализировать с помощью словаря аффиксов и основ;
– анализировать с помощью словаря системы слов.

В методе анализирования с помощью словаря аффиксов [6] рассматривается выделение аффиксов из слова и поиск по словарю, и на этой основе раскрыть грамматическое значение слова. В

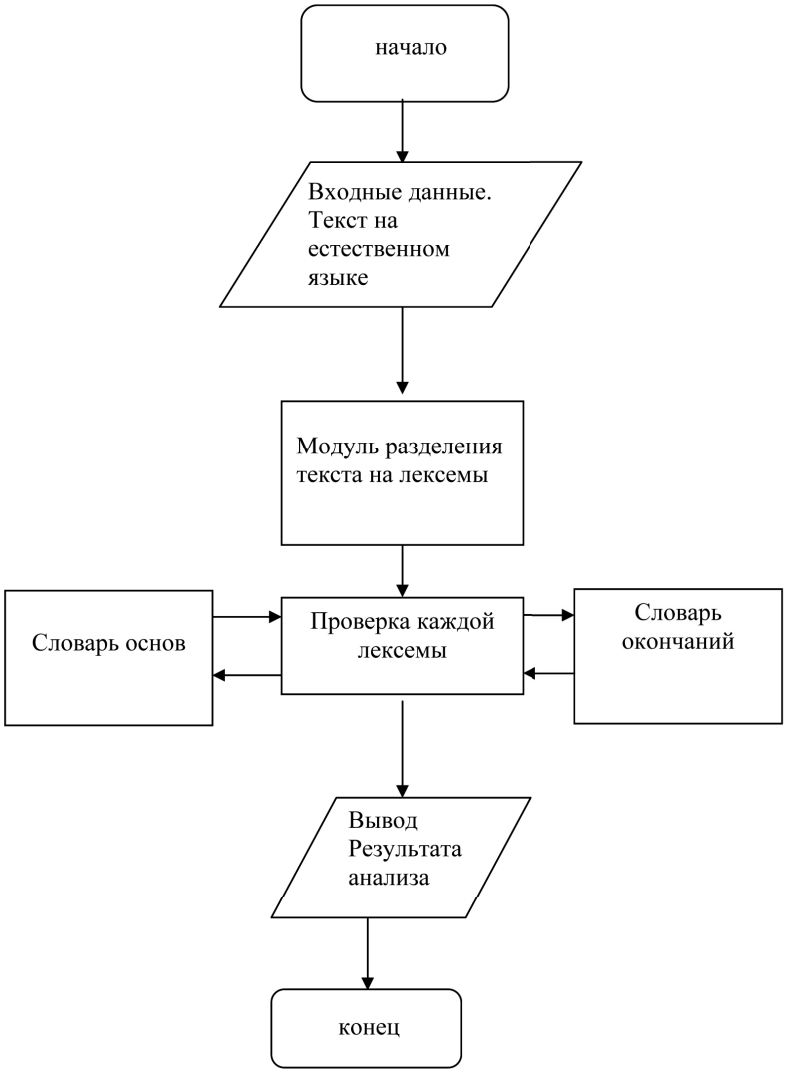


Рис. 1. Алгоритм морфологического анализа естественного текста

качестве результата такого вида анализа подбирается только грамматическое значение. В ходе морфологического анализа особо не пользуются словарем аффиксов и исследовательским

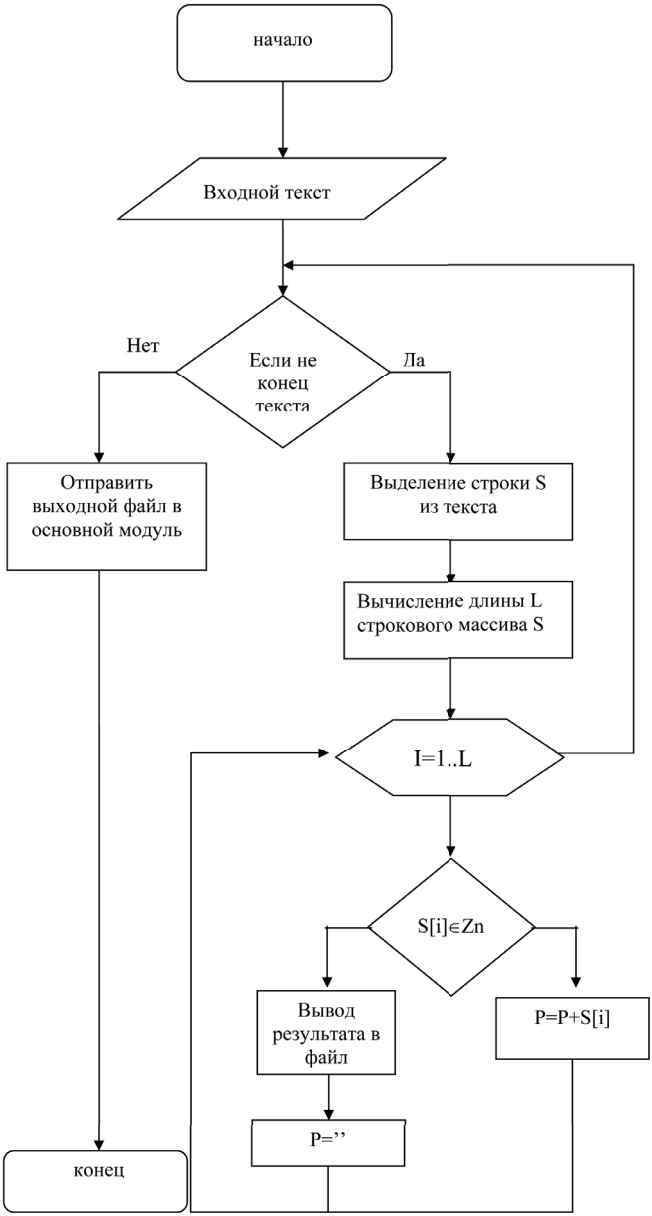


Рис. 2. Алгоритм модуля разделения текста на лексемы

методом в тексте естественного языка. Это связано с тем, что пользуясь одним только словарем аффиксов, нельзя проводить морфологический синтез.

В методике анализа, который состоит из словаря аффиксов и основ, основа и аффиксы слова выделяются и проводится поиск по словарю, исходя из этого берется грамматическое и семантическое значение слова.

Видом результата такого морфологического анализа считается только основа [8] и грамматическое значение [3,4,5].

Негативная сторона этого метода – при добавлении к некоторым основам аффиксов, из-за отпада букв, основа получается неполной.

Например, рассмотрим слово *балдар*, основа *бала+дар*, будет неправильным, если сделать анализ слова *бал*, потому что корень этого слова *бала*.

Третьим методом проведения морфологического анализа считается использование словаря системы слов [7]. Здесь нужно найти форму нужного слова и вернуть соответствующее ему грамматическое значение. При таком методе анализа приходится работать со словарем очень большого объема. Этот метод анализа выполняется для флективных языков, не предлагается для агглюнативных языков. Иначе говоря, не предлагается пользоваться этим методом в текстах кыргызского языка.

Среди вышерассмотренных методов анализа проведение морфологического анализа текстов естественного языка пользуясь словарем основ и аффиксов, применяется в среде программирования Embercadero RAD Studio. Словарь, составленный из основ, включает в себя класс основ и их различные формы. Это применено в целях устранения вышеназванных недостатков.

Как уже отмечалось, в кыргызском языке словоформу формирует агглютинативная цепочка, состоящая из словоизменительных аффиксов, а основа может состоят либо из корня, либо из корня, осложненного словообразовательным формантом. Задача морфологического анализа состоит в том, чтобы отделить основу и окончания, и приписать этим компонентам текста морфологическую информацию. Например, если в тексте встретится словоформа *тарбиячылар*, она расчленяется так: *тар-*

биячы+лар, где *тарбиячы* – основа, имя существительное ед.числа, осн.падежа; *-лар* – показатель множественного числа.

При составлении программы ряд специфических букв кыргызского языка отображался с помощью формата UTF8.

Фрагмент корпуса основ и аффиксов выглядит так (рис.3, 4):

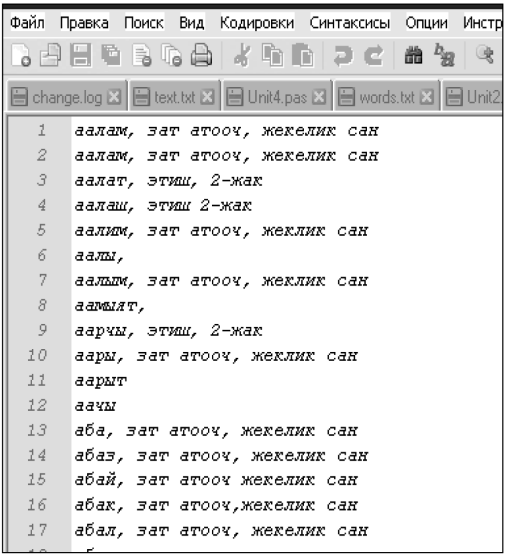


Рис. 3. Словарь основ

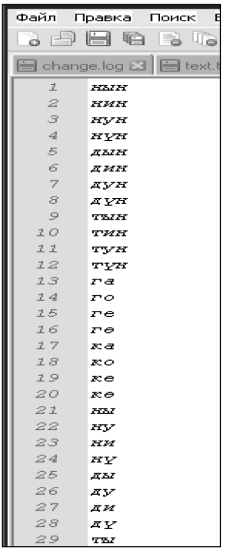


Рис. 4. Словарь аффиксов

6. Тестирование программы

На основе алгоритма в среде программирования Embercadero RAD Studio была составлена программа. Интерфейс системы показано на рисунке 4.

Морфологический анализатор можно применить во всех программах начального этапа работы с текстом на естественном языке.

Эти программы можно классифицировать так:

– системы антиплагиат;

– системы машинного перевода;

– системы поиска информаций;

– системы вопросов и ответов.

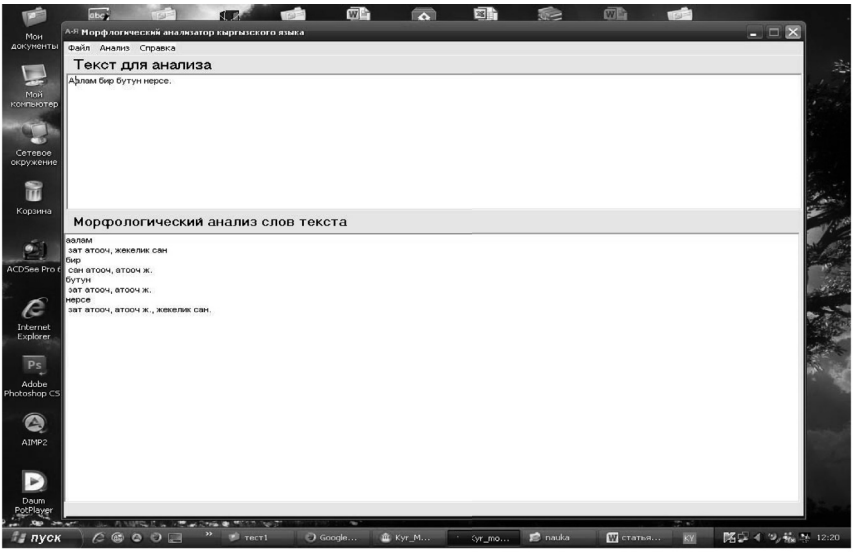


Рис. 4. Работа системы

7. Заключение

Учитывая особенности кыргызской морфологии, в сфере Embercadero RAD Studio была разработана начальная версия системы морфологического анализатора. Эта система была реализована на базе словарей основ и аффиксов. Она требует дальнейшего развития.

ЛИТЕРАТУРА

1. Абдувалиев И., Садыков Т. Современный кыргызский язык. (Морфология). – Бишкек, 1997.

2. Садыков Т. Проблемы моделирования тюркской морфологии. Фрунзе: Илим, 1987.

3. Белоногов, Г.Г. Автоматизированные информационные системы / Г.Г. Белоногов, В.И. Богатырев; под ред. К.В. Тараканова. – М.: Сов. радио, 1973. – 328 с.

4. Лингвистический процессор для сложных информационных систем / Ю.Д. Апресян, И.М. Богуславский, Л.Л. Иомдин и др. – М.: Наука, 1992. – 256 с.

5. Лингвистическое обеспечение системы ЭТАП-2 / Ю.Д. Апресян, И.М. Богуславский, Л.Л. Иомдин и др. – М: Наука, 1989. – 296 с.

6. Ножов, И.М. Прикладной морфологический анализ без словаря / И.М. Ножов // Тр. конф. по искусственному интеллекту КИИ-2000. – М.: Физматлит, 2000. – Т. 1. – С. 424–429.

7. Поляков, В.Н. Программа «Недоросль»: когнитивный подход к ис-следованию природы семантических связей естественного языка / В.Н. Поляков // Тез. 5-й нац. конф. по искусственному интеллекту КИИ-96. Казань, 1996. – С. 98–101.

8. Пруцков, А.В. Методы морфологической обработки текстов / А.В. Пруцков, А.К. Розанов // Прикаспийский журнал: управление и высокие технологии. – 2014. – № 3 (27). – С. 119–133.

УДК 81'33

MODELING GRAMMATICAL CATEGORIES OF VERB IN UZBEK AS STAGE OF MORPHOLOGICAL ANALYSIS IN MACHINE TRANSLATION

N. Abdurakhmonova
Tashkent State University of the Uzbek language and literature named after Alisher Navoi
abdurahmonova.1987@mail.ru

The article is devoted to how important the modeling of grammatical categories in Uzbek verbs and morpheme analysis for machine translation. In the article what aspects need representation in database.

The article examined the types of morphological analysis of grammatical categories of verbs and identified general paradigms and differences in source and target languages.

The verbs in the Uzbek language differentiate their analytical peculiarities by principles of word formations. Classifying of these particularities of verbs by compound, auxiliary and verbal word combinations gives very important issues in the special linguistic research.

Modeling of grammatical categories of verbs shows that the structure of verbs, types of connectivity, components, and definition of combination affixes of verbs. The article deals with creating database based inner opportunities of verbs, etc. constructing semantical models by modeling of verbs. Verbs in the Uzbek language differ their analytical characteristics and derivative forms. Compound forms of verbs and classification types of the verbs required very big linguistical analysis in computational morphology. On the modeling of grammatical categories in the verbs represented that all aspects as these the structures of verbs, the ways in communications and components and combinations in the verbs. On modeling of verbs of both of the languages analyzed by examples of morphological structures.

Keywords: Computational linguistics, the English language, the Uzbek language, machine translation, automated morphology, grammatical categories, morphotactics, analytical verbs, verbal word combination, morphological analysis, modeling, database of phrasal verbs.

МОДЕЛИРОВАНИЕ ГРАММАТИЧЕСКИХ КАТЕГОРИЙ И АНАЛИТИЧЕСКИХ ГЛАГОЛОВ УЗБЕКСКОГО ЯЗЫКА КАК ЭТАП МОРФОЛОГИЧЕСКОГО АНАЛИЗА В МАШИННОМ ПЕРЕВОДЕ

Н. Абдурахмонова
Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои
abdurahmonova.1987@mail.ru

В данной статье показан морфемный анализ моделирования грамматических категорий глаголов узбекского языка для машинного перевода. Также в статье почеркнуто именно на какие аспекты необходимо обратить внимание в момент их выражения на базе данных.

Также в статье описывается типы морфологического анализа грамматических категорий глагола, также обозначены общие парадигмы и различия в языке оригинала и переведенного языка. Глаголы в узбекском языке отличаются своими аналитическими особенностями и выделяются принципами формирования. Эти особенности глаголов выявляют необходимость проведения специальных лингвистических исследований по сложным, вспомогательным глаголам, по всем видам словосочетаний.

В моделировании грамматических категорий глагола указаны структура глаголов, виды соединений, состав, дефиниция комбинаций аффиксов глагола. Освещены проблемы создания базы данных на основе внутренних возможностей глагола, т.е. создание семантических моделей необходимо учитывать при моделировании глаголов. Глаголы на узбекском языке отличаются их аналитическим характером и особой формой образования, т. е. сочетанием глаголов в форме совместных, вспомогательных, Классификация видов требует крупного лингвистического анализа в компьютерной морфологии. При моделировании грамматических категорий в глаголе указывается, что все их аспекты, такие как структура глаголов, способ их общения, состав и комбинация дополнений в глаголах. При моделировании примеры примеров приводятся при анализе необходимости более глубокого изучения внутренних языковых способностей узбекской словесной семьи и их семантических моделей.

Ключевые слова: Компьютерная лингвистика, английский язык, узбекский язык, машинный перевод, автоматическая морфология, грамматическая категория, морфотактика, аналитические глаголы, глагольный словосочетание, морфологический анализ, моделирование, базы данных фразовый глаголов.

Введение

Информационные технологии играют большую роль в развитии науки и оказывают большое влияние на оптимизацию инфраструктуры в сфере накопления знаний, в основном на Интернет технологии. После обретения независимости в нашей стране каждая часть общества претерпела большие изменения. Наиболее сильно это проявилось в сфере образования. Широкое распространение вычислительных технологий и сети Интернет позволило решать серьезные научные проблемы. Появились новые научные направления, где используются компьютерные технологии. Одним из наиболее ярких примеров этого является то, что благодаря заслугам профессора А. Пулатова на факультете узбекской филологии Национального университета Узбекистана имени Мирзо Улугбека в 2000-х годах была открыта первая лаборатория компьютерной лингвистики. Основная цель этой лаборатории – сформулировать концептуальную идеологию узбекской компьютерной лингвистики. В те годы студентами этого цикла изучались такие дисциплины, как моделирование, машинный перевод, автоматическая орфография, создание информационного направления узбекского языка. PhD М.Хакимов проводил множество исследований по математическому моделированию в области многоязычного машинного перевода.

Под руководством М.Хакимова созданы следующие справочники и учебные пособия: «Компьютерная лингвистика» (А. Пулатов, 2011), «Основы компьютерной лингвистики» (А.Рахимов, 2011), «Лингвистические основы машинного перевода» (Н.Абдурахмонова, 2012) и др. Однако большинство работ были теоретическими, и не было никаких программных реализаций, основанных на реальных лингвистических базах данных. В настоящее время компьютерная лингвистика как наука преподается в нескольких государственных университетах Узбекистана: в Андижане, Намангане, Фергане, Хорезме, Самарканде, Бухаре. Самый большой центр по проведению исследований в области компьютерной лингвистики – Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои, который был создан 13 мая 2016 года указом первого президента Узбекистана И.А.Каримова.

Несмотря на разные отклонения в исследовании, после указа Каримова повысилось внимание к направлению компьютерной лингвистики и в течение короткого времени были созданы программы курсов по компьютерной лингвистике для бакалавриата и магистратуры. В соответствии с этим указом были поставлены акценты на следующие существенные вопросы: «... обеспечение соответствующего места нашего родного языка в мировой информационной сети Интернет, его компьютерное обеспечение и наличие научно-методических пособий, связанных с машинным переводом и электронными словарями, подготовка рекомендаций для широкого применения результатов на практике».

Мы должны выработать уважительное отношение к узбекскому языку, и поток исследований в этом направлении будет его положительной стороной. Узбекский язык – это язык великого Алишера Навои, который создал настоящие сокровища узбекского языка. Как известно, узбекский язык относится к тюркским языкам, который имеет давнюю историю со сменяющимся его состоянием по разным причинам. Его отличительные особенности от других языков мы можем видеть на каждом из языковых уровней.

Например, сохраненная сингармония гласных на турецком языке в таких словах, как üzüm, velâyet. Однако в узбекском языке больше заимствований, чем в турецком.

Например, *management, budget, test* из английского, *стол, поезд, бухгалтер* из русского, *vazir, maktab, maorif* из арабского.

С другой стороны, лексикология считается очень динамичной системой, которая связана с социальными и политическими ситуациями. Если говорить об узбекском языке, то его графическая система на протяжении веков также несколько раз менялась. После установления независимости Узбекистана в 1989 году узбекский язык получил статус государственного и были проведены реформы, направленные на его постоянное совершенствование. Одна из них – это орфографические правила на латинице, которые были установлены в 1995 году. Одна из важных задач компьютерной лингвистики в Узбекистане это создание грамматических анализаторов для узбекского языка.

1. Аннотирование узбекской грамматики

Грамматика состоит из двух частей: морфологии и синтаксиса.

Части речи узбекского языка			
самостоятельные части речи		**служебные части речи**	**отдельные группы слов**
Существительное	Наречие	Союз	междометия
Глагол	Числительное	Служебные слова	Подражательные слова
Прилагательное	Местоимение	Вспомогательные слова	Модальные слова

Представление грамматических значений, дериваций, словоизменительных правил и формальных моделей в морфологии рассматривается как лингвистическая процедура. Формальные морфологические модели появились в результате использования словосочетаний и отношений между ними в тексте. Формальные модели всегда существуют в синтагме.

Синтагма – семантико-синтаксическая единица, которая выражает некоторые унифицированные слова как значимую часть предложения. Лингвистическая база данных включает грамматику и словарь.

Как правило, синтаксический анализ, состоит из трех основных фаз:

1) Часть речи;
2) Член предложения;
3) Типы предложений.

Узбекский язык – это морфологически богатый язык с существительными, прилагательными и глаголами, изменяемыми по падежам, числам и другим формам слов. Это свойство требует добавления морфологической информации в системы машинного перевода для устранения недостатка множества флективных форм. Для машинного перевода важно создать формальную грамматику узбекского языка. Узбекский язык имеет агглютинативную морфологию с продуктивными флективными и деривационными суффиксами.

Суффиксы могут добавляться последовательно, и одно слово может содержать много параметров, такие как притяжательность, множественное / единственное число, падеж, модальность и т. д. Изменение по падежам – распространенная лингвистическая категория. В литературе, посвященной формальному синтаксису, есть два основных подхода к выделению падежей.

Первый подход в основном связан с работой Ноама Хомского, который рассматривает падеж как синтаксическое явление, известное в NLP; второй подход, предложенный в работе Алека Маранца, рассматривает падеж как предсинтактический, чисто морфологический феномен. [1, 461]

Существуют следующие деривационные модели узбекского языка:

W+A=>olma+zor
A+W=>be+foyda
W+W=>tez+yurar
W-W=>ota-ona
W W=>sotib olmoq
$W_{\text{-u/yu}}$ W=>Erta-yu kech

Из-за отсутствия грамматической информации для обработки естественного языка, осуществляется описание языка для лингвистической базы данных. Моделирование грамматических категорий для машинного перевода в узбекском языке производится в сравнении с английским языком. Английский и узбекский языки принадлежат разным языковым группам. Поэтому выделение уникальных свойств и различий обоих языков считается важным для морфологического анализа.

Рассмотрим этот процесс на примере глаголов узбекского языка.

2. Формальная модель традиционной морфологии в машинном переводе

Разумеется, что процесс перевода – трудная работа из-за ментальных и концептуальных различий, которые существуют в разных языковых семьях, обществе и культурах. На качество перевода влияют, как лингвистические (неоднозначность, синонимы), так и экстралингвистические (психологические) факторы. Даже человек-переводчик сталкивается с теми же проблемами в про-

цессе перевода, с которыми сталкиваются системы машинного перевода.

Как указано выше, независимо от того, производится перевод между родственными или неродственными языками, существуют определенные концептуальные различия между языками. Кемаль Алтынташ сравнивает крымскотатарский и турецкий языки, и считает, что «порядок слов и функции слов в предложении в большинстве случаев похожи. Корни обычно похожи, но иногда в разных языках они могут иметь разные значения». [2, 30].

Машинный перевод между тюркскими языками осуществляется легче, чем между неродственными языками.

Глаголы их меняются по числам, полу, лицу и времени, и два языка имеют сложные и похожие структуры глаголов и флективных систем.

Два языка имеют одни и те же глагольные формы:

1. Совершенная форма используется для прошедшего времени в узбекском языке.

2. Несовершенная форма используется для будущего времени в английском языке, но используется для выражения различных времен в узбекском языке (прошлого, настоящего и будущего) в сочетании с различными наклонениями и частицами.

3. Императив.

4. Активные и пассивные причастия используются с настоящим временем в английском языке и в меньшей степени с глаголами в узбекском языке.

Глаголы имеют следующие грамматические категории:

lexeme	->*o'qidim*
aspect	-> simple
person	->first
number	->singular
voice	->active
mood	- >indicative

Если посмотреть на агглютинативные языки, такие как финский, можно обнаружить, что морфосинтактические признаки систематически кодируются отдельными морфемами, которые расположены в линейном порядке. [3, 63]

Есть более 50 аффиксов, которые образуют новое значение флективных глаголов, и всего лишь 30 словоизменительных аффиксов; синтаксические аффиксы охватывают более 30 форм.

Полнозначные глаголы в узбекском языке состоят из более 6000 словарных слов. В узбекском языке есть около 207 типов аффиксов (включая вариации) частей речи, и 130 из них определяются как глагольные.

Чтобы добавить окончания к основам каждого слова, нужно выделить ту или иную часть речи из парадигмы [4, 12-17].

При морфологическом анализе основы слов даются в словарной форме с грамматической информацией и правилами.

Сравнивая основу «uchmoq», мы могли видеть некоторые примеры различных моделей глагольных структур:

1) Простой глагол – *uchmoq (fly)*

2) Составной глагол – *uchib ketmoq (fly away)*

3) Коллокация – samalyot *uchirmoq (fly the plane)*

4) Сочетание с глаголом – varrak *uchirib bermoq (fly the kite to smb.)*

5) Сочетание с модальным словом – *uchirish kerak (must fly)*

6) Идиома – *kapalagi uchib ketmoq (be afraid)*

Согласно Юлдашеву А.А. [5], существуют следующие типы аналитических форм глаголов в тюркских языках:

– *аналитическая условная форма;*

– *аналитическая модальная форма;*

– *аналитические формы другого вида.*

Кроме того, морфологический анализатор должен правильно анализировать каждый сегмент в тексте. В противном случае, при переводе единиц текста возникают проблемы омонимии. Например, комбинация слов *qo'yib berdi* используется во многих функциях как контекстная омонимия, как в следующих примерах:

U hujjatni stolga **qo'yib berdi**-> He gave document as **putting** on the table.

U bolani hovlida o'ynab olishiga **qo'yib berdi**-> He **let** the boy play in the yard.

Direktor ko'rsatilgan hujjatlarga darhol *imzo* **qo'yib berdi**-> The director **signed** abruptly brought documents.

U bolalar o'ynab olsin deb, sho'x ashula **qo'yib berdi**-> He **played music** so that to dance the children.

База данных и семантика глагольных сочетаний в узбекском языке мало изучены. Нет никаких вспомогательных глаголов, которые используются независимо от основных глаголов. Это около 30 типов глаголов, но они дают разные значения для разных глаголов. Это: *ber (ver), bil, bit (bitir), bor, boshla, boq, bo‘l, et, yoz, yot, kel, ket, ko‘r, ol, sol, tashla, tur, tush, chiq, yubor, yur, o‘l, o‘t, o‘tir, qara, qol, qo‘y. Существует две модели сочетания глаголов*:

– (MV+b/ib) HV

– (MV+a/y) HV

25 вспомогательных глаголов сочетаются с полноценными глаголами с помощью аффикса **-(i)b**:

O‘qi**b**	**berdi**	**bo‘ldi**	**ko‘rdi**	**(yoqib) tushdi**	**o‘tdi**
	Bildi	**(anglab) etdi**	**oldi**	**chiqdi**	**o‘tirdi**
	bitirdi	**yotdi**	**(aytib) soldi**	**yubordi**	**qaradi**
	bordi	**keldi**	**tashladi**	**yurdi**	**qoldi**
	boqdi	**ketdi**	**turdi**	**o‘ldi**	**qo‘ydi**

11 вспомогательных глаголов соединяются с помощью аффикса **-a/y** :

1) ber/ver (yoza berdi)
2) bil (topa bildi)
3) bor (o‘zgara bordi)
4) boshla (yoza boshladi)
5) yoz (yiqila yozdi)
6) ket (gapira ketdi)
7) ko‘r (ayta ko‘rma)
8) ol (unuta oldi)
9) sol (kela solib)
10) tur (yoza tur)
11) qol (ayta qol)

16 вспомогательных глаголов не могут соединяться с основным с помощью аффикса **-a/y**, но они связываются друг с другом только с помощью аффикса b/ib: **bit (bitir), boq, bo‘l, et, yot, kel, tashla, tush, chiq, yubor, yur, o‘l, o‘t, o‘tir, qara, qo‘y.**

9 вспомогательных глагола могут соединяться с помощью обоих аффиксов (b/ib и a/y): **ber** (aytib ber, ayta ber), **bil** (yozib biladi (in dialect), yoza biladi), **bor** (o‘qib bordi, o‘qiy bordi), **ket** (isib ketdi, gapira ketdi), **ko‘r** (aytib ko‘r, ayta ko‘rma), **ol** (yozib ol, yoza ol), **sol** (to‘kib soldi, kela solib), **tur** (o‘qib tur, yoza tur), **qol** (tugab qoldi, ayta qol). К этим глаголам могут присоединяться оба аффикса, однако они имеют разные значения: **yozib oldi (**помечать**) – yoza oldi (**мог написать**), aytib ko‘r (**попробуйте сказать**) – ayta ko‘rma (don’t tell any more).** Иногда они имеют одинаковые значения: **og‘irlashib bordi – og‘irlasha bordi**.

Значения глагольных комбинаций:

Форма	Значение	Пример
-(i)b ber	Прямое действие в отношении другого лица	qo‘yib ber, bilib ber
-a/y ber	Продолжение	ko‘chaver, o‘ylayver
-a/y ol/bil	Возможность	tuzata oldi, foydalana bil
-(i)b bit/bitir	Совершенное действие	yonib bitgan, ekib bitir
-i)b, -a/y bor/kel	Продолжение	unutib bordi, o‘zgartira bor
-a/y boshla	Начало	yura boshla, o‘qiy boshla
-(i)b boq/ko‘r/qara	Проба	o‘qib boq, o‘ylab qara, yozib ko‘r
-(i)b bo‘l	Совершенное действие	yozib bo‘l, yuvib bo‘l, qazib bo‘l
-(i)b et	Совершенное действие	anglab etdi, pishib etmagan
-a/y yoz	Завершенное действие	qula yozdi, yorila yozdi
-(i)b yot/tur/ o‘tir/yur	Продолжение	o‘qib yot, aytib tur, yozib o‘tir
-(i)b ket	Начало и продолжение	tarqab ketdi, isib ketdi, g‘ovlab ketdi, maqtab ketdi
-a/y ket	Начало и продолжение	o‘qiy ketdi, gapira ketdi
-(i)b sol	Совершенное действие	aytib sol, yuragini to‘kib soldi
-a sol	Очередное действие	ola solib ot, kela solib boshla, tura solib tashlan
-(i)b tashla/yubor	Полное и быстрое действие	o‘qib tashla, to‘qib tashla, haydab yubor
-(i)b chiq	Завершение	o‘qib chiq, so‘rashib chiq, aylanib chiq
-(i)b tush	Полнота	ag‘darilib tushdi, yiqilib tushdi, yoqib tushdi, yarashib tushdi
-(i)b o‘l	Продолжение и повторение	mudrab o‘lyapman, surishtirib o‘ldi, sog‘inib o‘ldi, chanqab o‘ldi
-(i)b o‘t	Совершенное действие	bo‘lib o‘tdi, gapirib o‘t

-(i)b qol	Неожиданность и продолжение	tikilib qol, serrayib qoldi; kelib qoldi, so‘rab qol; jonlanib qoldi
-a/y qol	Разрешение, согласие, пожелание	bora qol, yura qol, keta qol
-(i)b qo‘y	Продолжение и однократная активность	suyab qo‘y, ilib qo‘y; o‘ylantirib qo‘y, shoshirib qo‘y; qarab qo‘y, yo‘talib qo‘y

Вспомогательные глаголы иногда записываются в короткой форме или составной форме: *aytib yubor – aytvor, bora ber – boraver, unuta olmadi – unutolmadi.*

В таких глаголах содержится несколько глаголов: aytib berib tura qol, aytib bera olmay qoldi. Первая часть глагола всегда считается основным глаголом: Aytib (основной глагол) ber (вспомогательный глагол), aytib (основной глагол) berib tur (вспомогательный глагол).

Глагольное словосочетание похоже на составные глаголы, но только первый компонент глагола дает главное значение, а другие помогают этому главному глаголу, в составном глаголе сохраняются оба компонента в независимом значении. Мы видим три одинаковые модели:

1. (MV+PP) HV=> oshib tushmoq (Составной глагол)
2. (MV+PP) HV=> oshib ketmoq (Глагольное словосочетание)
3. (MV+PP) HV=> to‘pni oshirib uzatmoq (Коллокация)

1) если удалить с вспомогательный глагол из условных глаголов, то комбинация может иметь некоторые изменения в значении: **yozib bordi (write on) – yozdi (write), isib ketdi (warm up) – isidi (get warm)**. Тем не менее, удаление вспомогательного глагола из составного полностью изменит смысл слова, потому что вспомогательный глагол участвует в создании новых слов: **sotib ol (покупать) – sot (продавать)**; **ishlab chiqardi (изготовил) – ishladi (работал);**

2) в тексте имеется более двух независимых единиц коллокации: **Quvonib** (модификатор) **so‘zladi** (предикат) – *he talked joyfully*; в глагольном словосочетании есть только один предикат: **O‘ylab qoldi** (предикат).

Аффиксы залога и отрицания могут быть добавлены в обе части глагольного словосочетания: **to‘xtatib qo‘yishdi; aytib**

qo‘yma, aytmay tur, aytmay turma; аффиксы времени, наклонения, лица добавляются к вспомогательным глаголам. Отдельно от них существуют формы с аффиксами одновременно в обеих частях глагола: **tamomlashdi-qo‘yishdi.**

Обычно в дискуссиях по морфологии говорят об инфлективной версии словообразовательной морфологии на узбекском языке с точки зрения типов признаков каждого из этих кодов.

Важным вопросом является моделирование следующих грамматических форм в машинных переводах:

1) Моделирование глагольных словосочетаний;

2) Нахождение адекватного смысла глагольных словосочетаний на английском языке.

Основные модели глагольных сочетаний:

– MV HV->o‘qib berdi

– MV+HV->berolmadi<=> bera olmadi

– [MV]-[HV]->yozdi-qo‘ydi

1.1. Обозначения моделирования:

MV-основной глагол (MV – эти глаголы содержат собственные лексические значения);

HV1-вспомогательный глагол (HV1 – такие глаголы добавляются после аффиксов -b/-ib);

HV2 – вспомогательный глагол (HV1 – такие глаголы добавляются после аффиксов -a/-y);

– $MV+HV_1$=>aytib berdi

– $MV+HV_2$=>so‘zlay oldi

Глагольные словосочетания похожи на фразовые глаголы на английском языке (*look up, look forward*), так что некоторые из предлогов или наречий дают дополнительные значения условным глаголам:

o‘qib bo‘l-	Mushtariy kitobni o‘qib bo‘ldi.	Муштари закончил читать книгу
o‘qib chiq-	Mushtariy kitobni o‘qib chiqdi.	Муштари прочитал книгу.
o‘qib tur-	Mushtariy kitobni o‘qib turdi.	Муштари читал книгу.
o‘qib yubor-	Mushtariy kitobni o‘qib yubordi.	Муштари внезапно прочитал книгу.
o‘qib tashla-	Mushtariy kitobni o‘qib tashladi.	Муштари легко читал книгу.

o‘qib ol-	Mushtariy kitobni (qayta) o‘qib oldi.	Муштари читает книгу еще раз.
o‘qib ko‘r-	Mushtariy kitobni o‘qib ko‘rdi.	Муштари попытался прочитать книгу.
o‘qib qo‘y-	Mushtariy kitobni (o‘zi uchun) o‘qib qo‘ydi.	Муштари читал книгу про себя.
o‘qib ber-	Mushtariy kitobni (ukasiga) o‘qib berdi.	Муштари прочитал книгу своего брата.
o‘qib ket-	Mushtariy kitobni (to‘xtamay) o‘qib ketdi.	Муштар читал книгу без пауз.

Несмотря на некоторое сходство в обоих языках, есть один существенный аспект фразовых глаголов, который в узбекском языке определенно отличаются от фразовых глаголов на английском. Кроме того, он не может сравнивать две категории морфологических единиц из-за собственной специфики языков, моделей глагольных словосочетаний и отличия их позиций:

– MV+HV=> ko‘rib qoldi
– $MV+HV_{(1)} + HV_{(2)}$=>ko‘rsatib bera oldi
– $MV+HV_{(1)} + HV_{(2)} +HV_{(3)}$=>berib qo‘ya qoldi
– $MV+HV_{(1)} + HV_{(2)} +HV_{(3)} +HV_{(4)}$=>aytib berib qo‘ya qoldi
– $MV+HV_{(1)} + HV_{(1)}$=>o‘qib **tura tur**
– $MV_{(1)}+HV_{(1)}$=>**tura tur**
– $MV_{(1)}+HV_{(1)} +HV_{(2)}$=>**turib tura** qolgin

Как мы указывали выше, модели [глагол + глагол] включают одинаковые корни, которые могут появляться несколько раз, и они дают различные значения в тексте [6, 55].

Более того, они также выглядят как составные глаголы с внешней формой в соответствии со следующей структурой:

– Oshib tushmoq-> Составной глагол (climb over),
– Oshib ketmoq-> Глагольное словосочетание (rise up).

С другой стороны, вспомогательный глагол встречается как компонент в составных глаголах и идиомах:

– Nonushta qilmoq (завтракать) =>nonushta qilib berdi (составной глагол);
– Mashq qilmoq (делать упражнения) =>mashq qilib turdi (составной глагол);

– Kapalagi uchib ketmoq (бояться) => kapalagini uchirib yubordi (идиома).

Морфологический анализатор должен правильно идентифицировать каждую единицу в соответствии с их контекстным значением. В узбекском языке глаголы имеют следующие предикативные формы:

Отрицательная форма-N {-ma| -mas| -may}

Условная форма-CF {-sa}

Время-T {-a|-y|-yap|-moqda|-yotir…}

Предикативная форма-PF {-man|-san|-dir…}

Залог-VF {-t|-tir|-giz, |-kiz…}

Незаконченная формаNon-finite form-NF {-gan|-kan|-qan|-b|-ib…}

Личный-P {-im|-ing|-k|-ngiz|-lar| -man| -san|-k|-ngiz}

Цель-PS {-moqchi}

ПРИМЕР сочетания:

– MV+ N+PF=> O‘qimayman

– MV+ N+ T+PF=> O‘qimadim

– MV+ CF +PF=> O‘qisam

– MV+ N +CF +PF=>O‘qimasam

– MV+ T+PF=>O‘qiyapman

– MV+VF=>O‘qittir

– MV+NF=>O‘qigan

– MV+PS+PF=>o‘qimoqchiman

Аффикс может быть добавлен в оба компонента глагольного словосочетания:

– (MV+N+NF)+HV=> ko‘rmay qoldi

– (MV+ NF)+(HV+ N)=> ko‘rib qolmadi

– (MV+N+NF)+(HV+ N)=> ko‘rmay qolmadi

– (MV+V+NF)+HV =>ko‘rsattirib qo‘ydi

– (MV+NF)+(HV+ PF)=>ko‘rib borsam

– (MV+NF)+(HV+ N+PF)=>ko‘rib bormasam

– (MV+NF)+(HV+ VF+N+T)=>yeb ko‘rgizmadi

Кроме вышеупомянутых форм глагольных словосочетаний существуют следующие формы глаголов ***edi, ekan, emish*** and ***bo`lmoq, hisoblanmoq, sanalmoq, deyilmoq.*** Эти связывающие глаголы (эти глаголы создают предикативные формы) участвуют при формировании предикативных форм глаголов, которые имеют следующие модели:

a) N|Adj.|Num.|Pron.|Adv.+bo`lmoq=> Agar soat o‘n bo‘lsa, sizga qo‘ng‘iroq qilaman (Если будет 10 часов, я позвоню вам);

b) yo`q/bor/oz/ko`p/zarur/lozim/kerak+bo`lmoq=> O‘ylagan orzularim bir pasta yo‘q bo‘ldi (Мои , обдумываемые мечты внезапно исчезли);

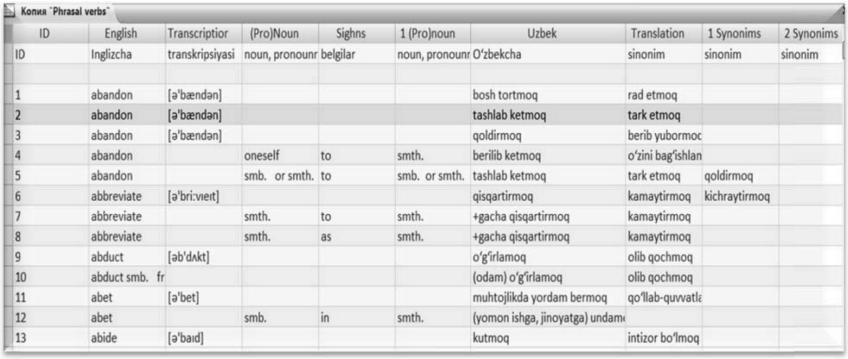
d) Infinitive+kerak/lozim/shart/darkor: xona tozalanishi kerak=> Xona tozalanishi kerak (Комната нуждается в уборке. – Комнату необходимо очистить).

4. Базы данных фразовых глаголов как аналитических моделей в англо-узбекском переводе

Для построения системы машинного перевода с английского языка на узбекский язык должен быть указан размер словаря, который был сохранен в базе данных. На английском и узбекском языках имеются очень большие базы данных, включающие все лингвистические уровни, и они очень разные. Глагольная категория на английском языке – фразовый глагол. Так или иначе, фразовые глаголы на английском языке, как глагольное словосочетание на узбекском языке имеют свои особенности. Это проблема для структурных компонентов предложения. Фразовые глаголы считаются очень важной и часто встречающейся особенностью английского языка. Во-первых, они настолько распространены в повседневном разговоре, и иностранцы, которые хотят казаться естественными, когда говорят на английском языке, должны изучать грамматику, чтобы знать, как правильно их произносить. Во-вторых, привычка изобретать фразовые глаголы была источником большого обогащения языка. С помощью фразовых глаголов описывается наибольшее разнообразие человеческих действий и отношений [7, p 16]. Следовательно, в частности, глагольные конструкции английского языка очень сложны для анализа и когерентного описания в синхронных терминах.

Конструирование базы данных считается информационной стадией цикла оазработки, и очень важной задачей является нормализация каждой единицы в процессе проектирования базы данных. Мы собрали более 12 тысяч фразовых глаголов и 3 тысячи отдельных смысловых глаголов. Каждая единица, размещена в отдельной ячейке базы данных и они составляют более 80 ты-

сяч фразовых и смысловых глаголов. Это система английско-узбекско-английского перевода.



Копия "Phrasal verbs"

ID	English	Transcriptior	(Pro)Noun	Sighns	1 (Pro)noun	Uzbek	Translation	1 Synonims	2 Synonims
ID	Inglizcha	transkripsiyasi	noun, pronounr	belgilar	noun, pronounr	O'zbekcha	sinonim	sinonim	sinonim
1	abandon	[ə'bændən]				bosh tortmoq	rad etmoq		
2	abandon	[ə'bændən]				tashlab ketmoq	tark etmoq		
3	abandon	[ə'bændən]				qoldirmoq	berib yubormoc		
4	abandon		oneself	to	smth.	berilib ketmoq	o'zini bag'ishlan		
5	abandon		smb. or smth.	to	smb. or smth.	tashlab ketmoq	tark etmoq	qoldirmoq	
6	abbreviate	[ə'bri:vieıt]				qisqartirmoq	kamaytirmoq	kichraytirmoq	
7	abbreviate		smth.	to	smth.	+gacha qisqartirmoq	kamaytirmoq		
8	abbreviate		smth.	as	smth.	+gacha qisqartirmoq	kamaytirmoq		
9	abduct	[əb'dʌkt]				o'g'irlamoq	olib qochmoq		
10	abduct smb.	fr				(odam) o'g'irlamoq	olib qochmoq		
11	abet	[ə'bet]				muhtojlikda yordam bermoq	qo'llab-quvvatla		
12	abet		smb.	in	smth.	(yomon ishga, jinoyatga) undam			
13	abide	[ə'baıd]				kutmoq	intizor bo'lmoq		

Существует возможность использования в электронных словарях таблицы данных с транскрипцией основного глагола. В базу данных включены почти все значения основного глагола с их синонимами, определенными в той же строке. Это помогает пользователям искать все синонимы глаголов не только для основных значений, но и для вторичных, а также для фразовых глаголов. Также учитываются разделяемые или неразделяемые фразовые глаголы. Эта база данных сформирована в соответствии со следующими моделями фразовых глаголов (V-глагол; P-причастие (предлог или наречие); – возможно или невозможно; smth. – что-то, smb. – кто-то)):

– V + oneself +P+smth. => присоединяться к smb. или smth.

– V+ oneself +P=> arch (oneself) over

– V + oneself +P+smb.=> attach oneself to smb

– V + P+ smb. or smth.+P smth. =>arrange with smb. about smth.

– V+ smb. or smth. + P+ smb. or smth. => associate smb. or smth. with smb. or smth.

– V+ smth. +P+ smth.=> balance smth. against smth.

– V+ smth. +P=> bail smth. out

– V+ smb. +P+ smth.=> astound smb. with smth.

– V+smb. +P+ smb. or smth.=> bias smb. against smb. or smth.

– V+smb. +P=> beat smb. up

– V +P+smb.=> bet with smb.

– V +P+smb. or smth.=> attend to smb. or smth.
– V +P+ smth.+P+ smth.=>
– V +P+ smb.+P+ smb.or smth.=> book smb. through (to some place)
– V+P+P=> be in for
– V+P+P+it=> be in for it
– V+P+P+smth.=> be off for smth.
– V+P+P+ smb.or smth.=> bound up with smb. or smth.
– V+P+ P+ smth=>bear up (against smth.)

Мы также включили следующие поясняющие символы:
– P1, P2, … , PN –N значения основных глаголов;
– N1, N2, …, Nk –k формы фразовых глаголов;
– B1, B2, …, Bl –l значения фразовых глаголов;
– P1_, P2_, … , PN_ – синонимы основных глаголов;
– B1_, B2_, …, Bl_ – синонимы значений фразовых глаголов;
– P – это означает, что он не принадлежит ни одному основному глаголу.

После анализа фразовых глаголов мы можем заключить, что фразовые глаголы соответствуют простым глаголам, сложным глаголам, словосочетанию, комбинации слов и идиомам на узбекском языке. Согласно Йорику Уилксу «Хотя мы согласны с тем, что маловероятно, чтобы информации в базы данных машиночитаемых словарей было достаточно для полной поддержки NLP, мы с оптимизмом смотрим на использование информации, которую они предоставляют для поддержки создания лексических записей в конкретных системах обработки естественного языка» [8, с. 139].

5. Глаголы узбекского языка для систем морфологического анализа

Автоматический морфологический анализ относится к самым первым работам по вычислительной лингвистике и машинному переводу с 1950-х годов (Andron, 1962;Woyna, 1962; Bernard-Georges et al., 1962; Boussard and Berthaud, 1965; Vauquois, 1965; Schveiger and Mathe, 1965; Matthews, 1966; Brand et al., 1969; Hutchins, 2001). Было создано множество приложений, в том числе стеммер Портера (Porter, 1980) широко используемый в

информационно-поисковых системах (Dolby et al., 1965; Attar et al., 1978; Choueka, 1983; Büttel et al., 1986; Meya-Lloport, 1987; Choueka, 1990; Koskenniemi, 1984), орфографических корректорах (McIlroy, 1982; Hankamer, 1986), системах текстового ввода (Becker, 1984; Abe et al., 1986) и в системах синтеза речи (Allen et al., 1987; Church, 1986; Coker et al., 1990). Многие из этих ранних приложений использовали довольно специфические подходы, включая кодирование большой части лингвистической информации в программе. Например, в системе, описанной в Coker et al. (1990), большая часть морфологического анализа представляется таблицами, закодированными как файлы операторов языка С и правилами изменения правописания, записанными как функции языка С [9, 100].

В системах машинного перевода требуется Word-менеджер. Word-менеджер (WM) система для работы с морфологическими словарями [10, 88].

С нашей точки зрения, морфологический анализ не должен ограничиваться только анализом категорий, который является основным в машинном переводе. Что касается Радольфо Дельмонте, он определил следующие лингвистические категории в итальянском языке [11, 4-5]:

- грамматические категории – полученные из категоризации реальности: сущности – существительные, события – глаголы и номиналы, атрибуты – прилагательные, наречия и существительные;
- семантические категории, как отрицание, кванторы;
- категории уровня дискурса, такие как дейктика, определенность, союзы для координации и подчинения на пропозициональном уровне;
- синтаксические категории – кодирование валентности предикатно-аргументных структур, поскольку они интерпретируются в ситуациях;
- аспектные категории – кодирование внутренней временной структуры событий (как выраженные как вербальными и невербальными определениями);
- семантические концептуальные категории – классификация типов событий по отношению к (не)реальности, которую они кодируют;

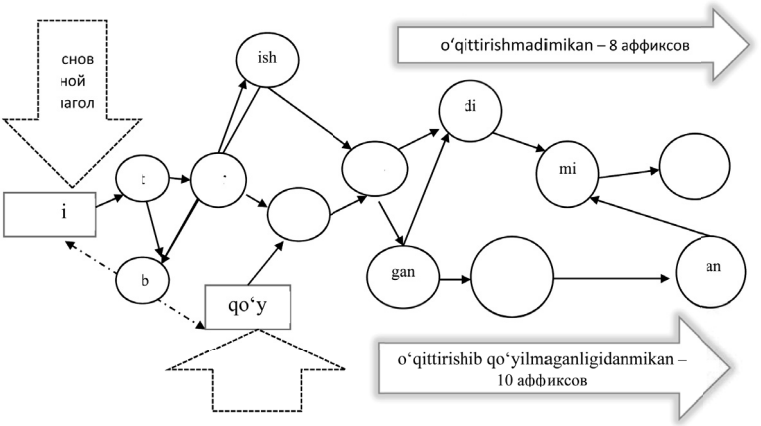
➢ выборные ограничения – кодирование типичности участников событий в неотъемлемых семантических признаках, поскольку они представлены в онтологии или связанной энциклопедической базе данных сущностей и их семантических отношениях;

➢ грамматические ограничения – кодирование так называемых синтаксических и анафорических ограничений привязки аргументов предиката и зависимых предикатов только для аргументов пропозиции.

Как видим, грамматические категории являются основной частью морфологического анализа. Согласно ряду работ, автоматический анализ имеет следующие типы [12, 65]:

- Стемминг
- Анализ словоформ со словарем
- Анализ с помощью логического подхода
- Анализ с помощью таблицы без словаря

Какой тип подходит для узбекского языка – рассмотрим морфотактические правила глагола:



Здесь глаголы с основой <ko’n> дают частичные определения:

ko‘nmoq<=>ko‘nikmoq

1. V->ko‘n (agree) -> U mening shartimga ko‘ndi (Он принял мои условия).

2. stem +ik=Verb ->ko‘nikmoq (used to)– Men shahar hayotiga ko‘nikdim (Я привык жить в урбанистическом образе жизни).

Прежде всего, это должны быть направленные глаголы в морфологическом анализе, которые включены в лемму. “Записи Lexicon токенизизируются с помощью простого алгоритма токенизации слева-направо. Запись токенизируется путем прохождения по строке ввода, шаг за шагом, и поиска самых длинных доступных цепочек с помощью жадного алгоритма токенизации. Если токенизатор является возрастающим, то он запоминает новые токены при анализе входной строки, предполагая, что многосимвольные токены были описаны ранее. Альтернативная, но менее эффективная стратегия состоит в том, чтобы определить все токены за один проход, чтобы образовать строку ввода с помощью токенизатора-преобразователя, реализующего жадный алгоритм слева-направо или какую-либо другую стратегию для достижения желаемых результатов» [13, 31]. Приведем еще один пример анализа глагола в тексте: *Men hali ham* ***tuzalganim yo‘q.*** – Если автомат анонсирует подчеркнутое слово как предикат, основная проблема заключается в том, как он будет правильно разбирать морфемы, чтобы придать правильный смысл на другом языке.

TUZALMOQ: 1) заживать, вылечиться; 2) уложить ткань; 3) ремонтироваться; 4) исправляться

Как мы видим, есть несколько значений слова “tuzalmoq”, поэтому мы должны определить правильное значение лексемы в соответствии с контекстной позицией.

1-step t+u+z+a+l+g+a+n+i+m

2-step tu+za gl+an +im

3-step tuz+alg+ani+m

4-step tuza+lgan+im

5-step tuzal+ganim

6-step tuza+l+gan+im

7-step tuzal+gan+im=>true

8-step tuzalganim

Какой шаг в морфемном анализе глагола верный?

Минимальная длина глагола в узбекском языке – две буквы: ich (пить), ek (сажать), ur (бить). Поэтому основа приращивается на одну букву.

Мы видим корни трех глагольных основ: 1) tuzmoq (расставлять), 2) tuzamoq (украшать), 3) tuzalmoq (заживать).

Word-менеджер ищет список аффиксов в базе данных. Таким образом рассмотрим, какие аффиксы правильной формы могут быть добавлены в глагол. В базе данных имеется следующий список:

ID	Affixes	Function	Abbreviation
1.	ga	Падеж	C
2.	gan	Время (Прошедшее время)	T_{past}
3.	gan	Причастие (Прошедшее время)	PP
4.	l	Залог (Активный)	$V_{act.}$
5.	l	Залог (Пассивный)	$V_{pass.}$
6.	a	Время (Настоящее время)	$T_{pres.}$
7.	m	Лицо (первое)	P_1
8.	i	Притяжательное местоимение (3-лицо)	$PossP_3$
9.	im	Притяжательное местоимение (1-лицо)	$PossP_1$

После проверки аффиксов тестируется модуль формирования и комбинации аффиксов.

– W_{stem}+V+PP+PossP=>tuza+l+gan+im=> Пассивный залог

– W_{stem} +PP+PossP=>tuzal+gan+im=> Активный залог

Здесь может возникнуть проблема определения правильного варианта. Принимая это во внимание считаем, что необходим семантический анализ этих данных. Предположим, что аффиксы принимают разное значение в зависимости от функциональной позиции. Отрицательные формы глагола также считаются одной из важных парадигм в узбекском языке. Потому что вариационные формы также приводят к изменению значения. Следующие модели демонстрируют отрицательные формы в разных коллокациях:

I. MV+ma=>o‘qimadi (Он не читал)

II. (MV_+may) HV=> O‘qimay qo‘ydi (Он не использовал для чтения)

III. MV (HV+ma)=>O‘qib qo‘ymadi (Он больше не читал)

IV. (MV_+may) (HV+ma)=>O‘qimay qo‘ymadi – **положительное значение** (Конечно, он прочитал (в прошедшем времени))

V. (MV+PP)+emas=>O‘qigan emas (Он никогда не читал)

VI. (MV+PP+Poss) yo‘q=>O‘qigani yo‘q (Он не читал)

VII. Na MV{CV, MV, VC}=>na habar oldi | na o‘qidi | na berib ketdi (Никто не читал)

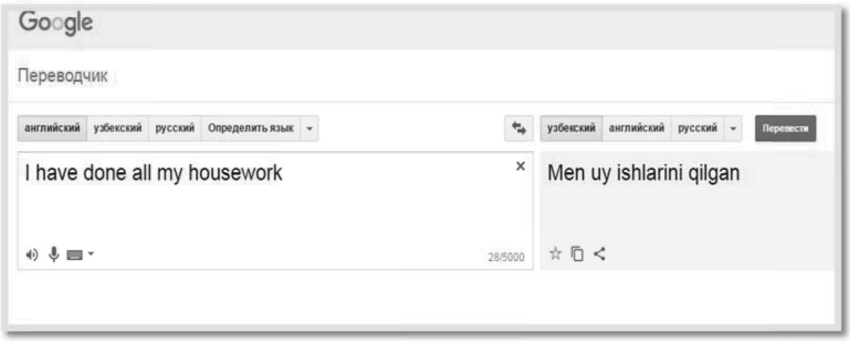
VIII. (MV+mas) edi=>O‘qimas edi (Он не использовал для чтения)

IX. (MV+ma+gan)+ekan+P=>O‘qimagan ekanman (Я не читал)

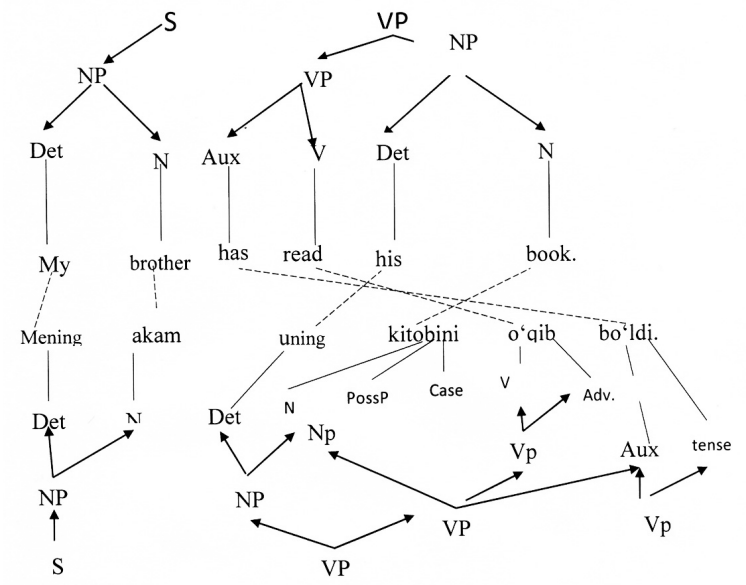
X. Na (MV+ma)=>Na o‘qimadi (Он не читал)

Композиция регулярных отношений – это основная операция, которую могут обрабатывать формальные устройства, найденные в морфологии естественного языка.

Проблема в том, что среди систем машинного перевода (Solver.uz, Google.translator и т.д.), ни один не может правильно переводить с узбекского на английский и наоборот. В качестве подтверждения мы можем привести корректность частей речей в переведенном тексте:



Предполагается, что семантическое поле лексики и контекстного значения словоформ необходимы для лингвистической базы данных в системе машинного перевода. Из-за отсутствия соответствия слов и сочетаний аффиксов, грамматических категорий в исходном и целевом языках, получаются только словарный, но не машинный перевод. В тексте, переведенном на узбекский язык, должно было быть: *Men barcha ishlarimni qilib bo‘ldim.* Проанализируем аналогичный текст в контекстно свободной грамматике:



Как показано в модели предложения, должно быть ясно, как в этих языках части предложения связаны между собой. В частности, некоторые категории, показанные на диаграмме, в одном из языков отсутствуют. Соответствующие части речи и семантические поля слов должны учитываться и в процессе машинного перевода.

Более того, мы могли видеть в таблице, как соответствуют части речи между языками:

Англо-Узбекско-Английский перевод					
	Kecha	ba'zi	talabalar	darsga	kelmadilar
Yesterday					
Some					
of					
the					
students					

did					
not					
come					
to					
the					
lesson					

В основном это результат исследований, основанных на статистическом машинном переводе. Вероятностные и математические подходы к МТ полезны, но сущность естественного языка не дает избежать этих проблем. В частности, английский и узбекский языки различаются особенностями связей разных языковых групп. Машиночитаемые словари теперь можно адаптировать к системам машинного перевода, но создание лингвистических баз данных могут улучшить их качество. Очень важно создавать параллельные базы данных на английском и узбекском языках. В противном случае, будет неправильный перевод. С другой стороны, очень сложно создать семантическую базу данных с аналогичным контекстным значением словосочетания. Следовательно, мы предлагаем создать базу данных соответствия словосочетаний с богатой лексикой:

read somebody like a ***book*** *– biror kishini juda yaxshi tushunmoq*
book *club – kitobxonlar to'garagi*
speak by the ***book –*** *aniq ma'lumotga tayanib gapirmoq*

Если наше исследование основано на фразеологических принципах совместно с подходом основанном на правилах, тогда машинный перевод способен придать смысл полностью концептуализированному со значением дискурса. Например, *It is a piece of cake* – 1) *Bu tort bo'lagi*; 2) *Bu juda oson.* Что мы должны делать для решения таких проблем? В перспективе нам потребуются параллельные корпуса. Однако для того, чтобы начать исследования по созданию корпуса, мы анализируем структуру предложения.

Завершив морфологический анализ, найдя соответствие между языковыми моделями, текст формируется таким образом, чтобы

сохранить контекстную информацию. В большинстве случаев порядок лексем в тексте важен и смысл текста напрямую зависит от этого порядка. Когда слова неоднозначны, а языковые единицы состоят из нескольких частей, таких как коллокации, идиомы или фразовые глаголы, то все возможные комбинации со словами генерируются одновременно.

Заключение

В общем, лингвистические модели и семантические отношения каждой языковой единицы играют важную роль при создании баз данных для систем машинного перевода. Из-за процессов глобализации все меняется; также нет никаких препятствий для унификации культурных и социальных отношений между людьми.

Поэтому понимание иностранных языка очень важно и мы не можем это игнорировать. Сегодня результат машинного перевода, который появился в последней половине 20-го века, повлиял на расширение исследований в этом направлении компьютерной лингвистики и новейших информационных технологий, что дает возможность использовать их в любых сферах общественной деятельности. Используя только грамматические модели нельзя решить проблему машинного перевода. Если не использовать полную информацию о языке, системы машинного перевода останутся только вспомогательным инструментом переводчика.

ЛИТЕРАТУРА

1. Лютикова Е. А. Формальное Моделирование падежного варьирования: параметрический подход // Компьютерная лингвистика и интеллектуальные технологии По материалам ежегодной международной конференции «Диалог» (2016) Выпуск 15, С. 461.

2. Kemal Altıntaşh Turkish to Crimean Tatar machine translation system (a thesis submitted to the department of computer engineering and the institute of engineering and science of Bilkent university in partial fulfillment of the requirements for the degree of master of science) 2001, P. 30

3. Brian Roark, Richard Sproat. Computational Approaches to Morphology and Syntax. Oxford University Press Inc., New York, 2007, P. 63.

4. N.Abdurakhmonova, The bases of automatic morphological analysis for machine translation, Известия Кыргызский государственный

технический университет им. И.Раззокова теоритической и прикладной научно-технический журнал, 2016 № 2 (38)], С. 12–17.

5. Yuldashev A.A. Аналитические формы глагола в тюркских языках. Москва, – Наука, 1965.

6. N.Abdurakhmonova. O‘zakdosh fe’llarning ketma-ket qo‘llanilishiga doir – Tilshunoslikka ilk qadam (to‘plam III), T., 2007, B. 55–59.

7. Andreea-Rosalia Olteanu. A holistic approach to phrasal verbs, Editura Sfântul Ierarh Nicolae 2012, P 16.

8. Yorick Wilks. Machine translation. Its scope and limits. Spring science+Business Media LLC. 2009 UK, P 139.

9. Brian Roark and Richard Sproat. Computational Approaches to Morphology and Syntax, 2007, OXFORD, P.100.

10. State of the Art in Computational Morphology, Workshop on Systems and Frameworks for Computational Morphology, SFCM 2009 Zurich, Switzerland, September 4, 2009 Proceedings, P. 88.

11. Rodolfe D. Computational Linguistic Text Processing: Lexicon, Grammar, Parsing and Anaphora Resolution. Nova Science Publishers, Inc. New York, 2008, P. 4–5.

12. Марчук Ю. Компьютерная лингвистика. Москва, 2006 С. 65.

13. State of the Art in Computational Morphology Workshop on Systems and Frameworks for Computational Morphology, SFCM 2009 Zurich, Switzerland, September 4, 2009, Proceedings, P. 31.

14. Дрейзин Ф.А. Об алгоритмизации составление алгоритма анализа языка. (на примере морфологии агглютинативного узбекского языка)//Научные труди Ташкентского университета, 1961, вып. 189. Матем. Науки, кн. С. 121.

15. Абдурахмонова Н.З. Машина таржимасининг лингвистик асослари. Тошкент, «Академнашр», 2012 Б. 160.

УДК 81'33

APPLICATION OF COMPARATIVE ANALYSIS OF ENGLISH AND AZERBAIJAN LANGUAGES FOR CREATION OF THE MORPHOLOGICAL ANALYZER KNOWLEDGE BASE IN EXPERT SYSTEM OF MACHINE TRANSLATION SUPPORT

Z. Kuliyeva
Oxbridge Academy, Baku, Azerbaijan
guliyeva_z_y@hotmail.com

Raising of the national language status, integration of Azerbaijan into the international community, along with the elimination of economic and political barriers, the elimination of language barriers and the dynamic development of the information technology sector, have created an urgent need to create machine translation systems and computer dictionaries from various natural languages to the Azerbaijani language and vice versa. This, in turn, stimulated the development of such scientific areas as computer lexicography, machine translation and automatic text processing systems. Since the dominant language in the Internet is English, translating texts into national language and vice versa acquires special significance. The principles of compiling an automatic dictionary for machine translation systems play an important role in obtaining a quality translation, improving its quality and reducing the loss of the meaning of a text in a translated language. The current machine translation systems focus on such issues as the formation of a vocabulary structure and lexical composition, taking into account the structural and lexical differences between languages, the problems of disambiguation, the selection of transfer equivalents of grammatical constructions and the corresponding characteristics of the source language, etc. The proposed material describes the process of comparative analysis of the morphological, syntactic and semantic systems of a pair of studied languages (English and Azerbaijani languages) necessary for developing the morphosyntactic analyzer knowledge base in the expert system of machine translation support.

An expert system of machine translation support has been practically implemented. The developed system is implemented at the Institute of Information Technologies of the National Academy of Sciences of Azerbaijan on the basis of the Delphi 7 program. The volume of the dictionary in the implemented version is 3000 entries for each language. On the basis of the dictionary, a database is implemented, it is oriented to a specific subject area, in which at the initial stage words of neutral vocabulary are included to check the rules that make up the knowledge base. The automatic dictionary as part of the expert system is an information store used to process text based on

knowledge, presented as transformation rules of recognition and generation of grammatical, phonetic and semantic phenomena of the language. An automatic binary dictionary, developed as part of an integrated translation system, serves as the main tool for finding (fixing) lexical translation equivalents; it is used for working in the interactive mode, being integrated into the general lexicographic base and is the main informative-reference database. Knowledge is represented by a set of rules, each of which consists of a condition and a result.

Formally, the rules that form the basis of the knowledge base of the expert system and ensure the functioning of databases and knowledge are classified into recognition, generation and substitution rules. In accordance with this classification, the knowledge base, respectively, can be divided into 3 blocks, which in turn continuously interact with the database as part of the expert system analyzer.

The knowledge base of the expert system, including a wide range of formalized linguistic information and contributing to the selection of more accurate translation equivalents of dictionary entries, makes it possible to improve the efficiency of translation processes. This system can be used in the machine translation system, as well as in any text processing system in the English-Azerbaijani language pair. Since the rules in the knowledge base of the expert system are based on the lexical and grammatical information of the English and Azerbaijani languages, the latter can also be used as a training system providing the opportunity to teach grammar and vocabulary of both Azerbaijani and English.

Keywords: Azerbaijani language, knowledge base, morphological analyzer.

ПРИМЕНЕНИЕ СРАВНИТЕЛЬНОГО АНАЛИЗА АНГЛИЙСКОГО И АЗЕРБАЙДЖАНСКОГО ЯЗЫКОВ ДЛЯ СОЗДАНИЯ БАЗЫ ЗНАНИЙ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА В ЭКСПЕРТНОЙ СИСТЕМЕ ПОДДЕРЖКИ МАШИННОГО ПЕРЕВОДА

З. Ю. Кулиева
Оксбриджская академия, Баку, Азербайджан
guliyeva_z_y@hotmail.com

Повышение статуса национального языка, все более тесная интеграция Азербайджана в международное сообщество, предполагающая наряду с устранением экономических и политических барьеров, устранение и языковых барьеров динамичное развитие сектора информационных технологий вызвали острую потребность в создании систем машинного перевода и компьютерных словарей с различных естественных языков на

азербайджанский язык и наоборот. Это, в свою очередь, стимулировало развитие в стране таких научных направлений как компьютерная лексикография, машинный перевод и системы автоматической обработки текста (АОТ). Поскольку доминирующим в сети языком является английский, то реализация перевода текстов на национальные языки, и наоборот приобретает особую значимость. Важную роль в получении качественного перевода, улучшения его качества и уменьшения потерь смысла текста на переводном языке играют принципы составления автоматического словаря для систем машинного перевода. В действующих системах машинного перевода основное внимание уделяется таким вопросам, как формирование структуры словаря и лексического состава с учетом структурных и лексических различий между языками, проблемы устранения неоднозначности, отбор переводных эквивалентов грамматических конструкций и соответствующих характеристик исходного языка и т.п. Предлагаемый материал описывает процесс проведения сравнительного анализа морфологических, синтаксических и семантических систем пары исследуемых языков (английского и азербайджанского языков) необходимого для разработки базы знаний морфо-синтаксического анализатора в экспертной системы поддержки машинного перевода.

Практически реализована экспертная система поддержки машинного перевода. Разработанная система реализована в Институте Информационных Технологий Национальной Академии Наук Азербайджана на основе программы Delphi 7. Объем словаря в реализованной версии составляет 3000 записей для каждого языка. На основе словаря реализуется база данных, ориентируемая на конкретную предметную область, в которую на начальном этапе включены слова нейтральной лексики, для проверки правил, входящих в состав базы знаний. Автоматический словарь в составе экспертной системы представляет собой хранилище информации, используемое для обработки текста на основе знаний, представленных в виде трансформационных правил распознавания и порождения грамматических, фонетических и семантических явлений языка. Автоматический бинарный словарь, разработанный как часть интегрированной системы перевода, служит основным инструментом поиска (установления) лексических переводных эквивалентов; используется для работы в диалоговом режиме, будучи интегрированным в общую лексикографическую базу и является основной информативно-справочной базой. Знания представлены набором правил, каждое из которых состоит из условия и результата.

Формально правила, составляющие основу базы знаний экспертной системы и обеспечивающие функционирование баз данных и знаний, классифицированы на правила распознавания, порождения и подстановки. В соответствии с данной классификацией база знаний, соответственно, может быть разделена на 3 блока, которые в свою очередь непрерывно взаимодействуют с базой данных в составе анализатора экспертной системы.

База знаний экспертной системы, включающая широкий спектр формализованной лингвистической информации, способствующая отбору более точных переводных эквивалентов словарных статей, позволяет повысить эффективность процессов перевода. Данная система может быть использована в системе машинного перевода, а также в любой системе обработки текста в англо-азербайджанской языковой среде. Поскольку правила в базе знаний экспертной системы основаны на лексико-грамматической информации английского и азербайджанского языков, то последняя может быть также использована в качестве обучающей системы, предоставляющей возможность обучения грамматике и лексике как азербайджанского, так и английского языка.

Ключевые слова: азербайджанский язык, база знаний, морфологический анализатор.

Введение

Повышение статуса национального языка, все более тесная интеграция Азербайджана в международное сообщество, предполагающая наряду с устранением экономических и политических барьеров, устранение и языковых барьеров динамичное развитие сектора информационных технологий вызвали острую потребность в создании систем МП и компьютерных словарей с различных естественных языков на азербайджанский язык и наоборот. Это, в свою очередь, стимулировало развитие в стране таких научных направлений как компьютерная лексикография, машинный перевод и системы автоматической обработки текста (АОТ). Поскольку доминирующим в сети языком является английский, то реализация перевода текстов на национальные языки, и наоборот приобретает особую значимость. Важную роль в получении качественного перевода, улучшения его качества и уменьшения потерь смысла текста на переводном языке играют принципы составления автоматического словаря для СМП. В действующих системах машинного перевода (МП) основное внимание уделяется таким вопросам, как формирование структуры словаря и лексического состава с учетом структурных и лексических различий между языками, проблемы устранения неоднозначности, отбор переводных эквивалентов грамматических конструкций и соответствующих характеристик исходного языка и т.п. Предлагаемый материал описывает процесс проведения

сравнительного анализа морфологических, синтаксических и семантических систем пары исследуемых языков (английского и азербайджанского языков), необходимого для разработки базы знаний морфо-синтаксического анализатора в экспертной системе поддержки машинного перевода. Предлагаются принципы отбора формальных признаков и методология создания АС посредством отбора формальных признаков в словарные статьи для различных частей речи, представляющих собой формализованную лингвистическую информацию языковой пары (английский и азербайджанский), приемлемую для хранения в базе данных и переработке в базе знаний анализатора. Дается краткое описание составленного словаря, применяемого в качестве базы данных экспертной системы поддержки машинного перевода.

На базе теоретических принципов разработки оптимальной структуры АС, предложенных в данной работе, практически реализована экспертная система поддержки машинного перевода. Разработанная ЭСПМП реализована в Институте Информационных Технологий Национальной Академии Наук Азербайджана на основе программы Delphi 7, применяемой для создания систем управления базами данных и знаний. Объем словаря в реализованной версии составляет 3000 входов на каждый язык в словаре комбинированного типа. На основе словаря реализуется база данных, ориентируемая на конкретную предметную область, в которую на начальном этапе включены слова нейтральной лексики, для проверки правил, входящих в состав базы знаний. Автоматический словарь в составе экспертной системы представляет собой хранилище информации, используемое для обработки текста на основе знаний, представленных в виде трансформационных правил распознавания и порождения грамматических, фонетических и семантических явлений языка. Автоматический бинарный словарь, разработанный как часть интегрированной системы перевода, служит основным инструментом поиска (установления) лексических переводных эквивалентов в ЭСПМП; используется для работы в диалоговом режиме, будучи интегрированным в общую лексикографическую базу ЭСПМП и является основной информативно-справочной базой; и наконец, в ЭСПМП, как в одной из систем автоматической обработки текста, АС служит источником грамматической информации, необходимой для ра-

боты алгоритмов автоматического морфологического и синтаксического анализов, а также для работы алгоритмов лемматизации и правил базы знаний. Знания ЭСПМП представлены набором правил, каждое из которых состоит из: антецедента (условия) и консеквента (результата).

Формально правила, составляющие основу базы знаний экспертной системы и обеспечивающие функционирование баз данных и знаний, классифицированы на правила распознавания, порождения и подстановки. В соответствии с данной классификацией база знаний, соответственно, может быть разделена на 3 блока, которые в свою очередь непрерывно взаимодействуют с базой данных в составе анализатора экспертной системы.

База знаний экспертной системы (ЭС), включающая широкий спектр формализованной лингвистической информации, способствующая отбору более точных переводных эквивалентов словарных статей, позволяет повысить эффективность процессов перевода. Данная система может быть использована в системе машинного перевода, а также в любой системе обработки текста в англо-азербайджанской языковой среде. Поскольку правила в базе знаний экспертной системы поддержки МП (ЭСПМП) основаны на лексико-грамматической информации английского и азербайджанского языков, то последняя может быть также использована в качестве обучающей ЭС, предоставляющей возможность обучения грамматике и лексике как азербайджанского, так и английского языка.

1. Краткое описание морфо-синтаксического анализатора ЭСПМП

Теоретические исследования и практические разработки в области машинного перевода (МП) и компьютерной лингвистики в целом с начала 21 века в Азербайджане были направлены на некоторые проблемы морфологического и синтаксического анализа, являющегося одним из важных этапов системы МП. В процессе исследований были рассмотрены орфографические особенности азербайджанских текстов и выделены основные этапы автоматического выявления ошибок. При этом в процесс автоматического редактирования азербайджанских текстов включены такие

уровни контроля, как обеспечение правильности слов текста, автоматическое распознавание словосочетаний, контроль правильности переноса словоформ, контроль правильности употребления знаков препинания, контроль структуры предложения и др. Были осуществлены попытки создания систем перевода с и на азербайджанский язык, созданы электронные словари в помощь переводчику и словари, направленные на определенную предметную область.

Как логический итог этих работ, появились системы, проверяющие правильность текстов азербайджанского языка и электронный словарь азербайджанского языка, вышедшие на промышленный уровень эксплуатации. Хотя они еще окончательно не завершены, но и это можно считать определенным успехом в создании таких систем.

С 2002 года начали проводиться системные исследования по разработке теоретических основ и принципов реализации автоматической обработки азербайджанского текста на всех языковых уровнях: лексическом, морфологическом, морфонологическом, синтаксическом и семантическом. В течение нескольких лет был осуществлен сравнительный анализ английского и азербайджанского языков, выявлены лингвистические сходства и различия в морфологической и синтаксической системе этой пары языков. К 2011 году в Институте Информационных Технологий НАНА было положено начало создания базы знаний (БЗ) морфосинтаксического анализатора, в котором БЗ представляет собой рекурсивные правила анализа словоформ выходного языка, трансформированные в виде формул, с последующим выходом, то есть конечным синтезом словоформ на основе правил выходного языка. В результате сопоставления двух языков и проведенного сравнительного анализа были выявлены те признаки частей речи, которые могут быть представлены в формальном виде и на основе которых может быть создана база знаний морфосинтаксического анализатора. Так, реализуемая база знаний, создавалась на языке программирования DELPHI-7, включая в себя различные объекты и связывающие их правила, куда так вводятся формальные правила порождения правил распознавания словоформ множественного числа, категории принадлежности, сочетаний существительного с личными местоимениями, глаголом и предлогами в англо-

азербайджанской среде. Исследование процесса формального описания синтеза слова позволило разработать правила автоматического синтеза слова и учесть те изменения, которые возникают при присоединении заданных аффиксов к основе. Далее были описаны закономерности выпадения букв, чередования согласных и сингармонического уподобления гласных на стыках морфем в соответствии с требованиями азербайджанской орфографии и автоматической расстановки морфем, автоматическое определение порядка слов между компонентами словосочетаний, автоматическое определение порядка слов в предложении, автоматическое определение типов предложений.

На сегодняшний день наиболее слабым звеном Экспертной Системы Поддержки Машинного Перевода (ЭСПМП) является формальный анализ смысла текста, т.е. обеспечение автоматического понимания последнего. Формальный анализ смысла текста предполагает выполнение двух основных операций: выявление семантических связей между словами и их представление в той или иной форме. Необходимо создание банков лингвистических данных, в которых хранение, обработка и выдача терминологической информации осуществляется с использованием ЭВМ. С другой стороны банки терминов на базе ЭВМ могут интенсифицировать переводческую работу, улучшить качество переводов и облегчить поиск терминов.

Морфосинтаксический анализатор позволяет рассмотреть подчинительные связи и грамматические значения слов в предложении. Он реализует универсальные языковые зависимости, что позволяет применять его для разработки анализаторов на единой платформе. Таким образом, учитывая особенности, сходства и различия английского и азербайджанского языков, составлены списки продукционных правил. В процессе разработки ЭСПМП в базу данных морфосинтаксического ее анализатора были введены:

– списки алфавитов двух языков с разделением на поля (гласные, согласные) – для того чтобы машина могла определить, что слово завершается на гласную или согласную; азербайджанский алфавит – на поля: мягкие и твердые гласные и согласные;

– морфонологические правила для различных частей речи английского языка, которые выявляют фонетические изменения,

возникающих на стыке основы и аффиксов при анализе словоформ;

– морфонологические правила для различных частей речи азербайджанского языка, которые выявляют фонетические изменения, возникающих на стыке основы и аффиксов при синтезе словоформ;

– список исключений при образовании множественного числа существительных с морфонологическими изменениями в основе словоформ;

– список исключений при образовании множественного числа существительных, оканчивающихся на – о;

– список неправильных глаголов;

– список существительных азербайджанского языка с выпадением корневой гласной в форме Р.п., Д.п. и В.п.;

– список существительных азербайджанского языка, не подчиняющихся правилам чередования [k~y];

– список существительных азербайджанского языка, не подчиняющихся правилам чередования [q~ğ];

– правила чередования гласных в словоформах и аффиксах по закону сингармонизма;

– формальные правила образования грамматических форм частей речи английского языка с последующим выходом на формальные правила образования грамматических форм частей речи азербайджанского языка;

– правила построения порядка слов предложения английского языка;

– правила построения порядка слов предложения азербайджанского языка.

В настоящее время продолжается поиск путей решения проблем машинного анализа, синтеза и перевода, в результате которого, были построены формальные модели языка на синтаксическом и морфологическом уровнях, а также формальные модели различных типов предложений. Необходимость каталогизировать все морфосинтаксические элементы языка привела к созданию базы знаний на основе грамматики двух языков. Сегодня становится актуальным построение анализатора более широкого круга морфологических знаний как инструмента для обобщения, применения и оценки разнообразных морфологических концепций.

2. Автоматический словарь как база данных ЭСПМП

В любой системе перевода основным средством хранения языковой информации является автоматический словарь (АС), непосредственно связанный со всеми уровнями языковой иерархии и принимающий участие на всех этапах переводческого процесса. Данные, содержащиеся в АС, используются как для морфологического, синтаксического, так и для семантического анализа. Однако, будучи составным элементом системы, АС взаимодействует с базой знаний, без которой данные словаря не будут иметь особой значимости в СМП. Понимание языка как знаковой системы с четко выделенными структурными элементами (единицами языка, их классами и др.) облегчает и, более того, способствует проведению типологического анализа пары языков. Для типологического анализа обнаружение параллелизма языковой структуры в отдельных микро- и макро- системах имеет первостепенное значение, так как это во многом облегчает осуществление формального описания языка. Схожие явления в системах языков называют изоморфами. Язык представляет собой сеть отношений между элементами, упорядоченных и находящихся в иерархической зависимости в пределах определенных уровней. Все языки имеют деление на гласные и согласные звуки, части речи, разветвленная вербальная система и т.д. Сходные черты рассматриваемых языков представляют собой языковые универсалии, отличительные черты называют языковыми дифференциями.

В рамках составления морфо-синтаксического анализатора первоначальным этапом сравнительного анализа двух различных по своей структуре и происхождению языков является выявление генеалогического происхождения рабочих языков. По своему строению языки часто делятся на 4 вида: изолированные, агглютинативные, флективные и аморфные языки. Чистых языков, то есть языков только одного вида не существует. Английский язык относится к западногерманской подгруппе германской группы индоевропейской языковой семьи. Английский, немецкий и другие языки германской группы относятся к языкам флективно-аналитического типа. Здесь флективность характеризуется полифункциональностью грамматических морфем, а аналитичность отражает грамматические знания, выражаемые вне слова – слу-

жебные слова, порядок слов, интонация. Поэтому при составлении словников для этих языков в большинстве случаев в качестве словарной единицы используются словоформы. В английском языке доминируют аналитические формы. К ним относятся вспомогательные глаголы, частицы для образования временных и других форм глагола. Однако в некоторых случаях имеет место процесс агглютинации. Азербайджанский язык относится к огузо-сельджукской подгруппе огузской группы в тюркской языковой семье. Азербайджанский, турецкий, узбекский и др. языки тюркской группы относятся к языкам агглютинативного типа, т.е. словоизменение в этих языках осуществляется путем агглютинации (механического присоединения однозначных стандартных аффиксов к неизменяемым основам или корням). В азербайджанском языке практически 90% грамматических форм образуется при помощи агглютинации. Очевидно, что для составления АС требуется исследование этой пары языков, их морфологических, синтаксических, лексических и семантических особенностей; установление их языковых сходств и различий.

3. Типологическое сопоставление и определение универсалий и дифференций рабочей пары языков

Для выявления сходств и различий английского и азербайджанского языков изучена структура морфологической и синтаксической парадигматической систем языка. Парадигму можно определить как систему грамматического изменения слова для некоторой части речи в конкретном языке, которая представляет собой структуру словоизменительных реализаций у всех комбинаций морфосинтаксических свойств, причем каждое слово этой части речи обладает всеми реализациями и только ими. Каждая из частей речи в языке называется парадигматической группой, или парадигмой, а каждая форма, входящая в группу, называется словоформой (inflected form) или флексией.

В структурной лингвистике парадигматическими отношениями называют «отношения между единицами, принадлежащими одному и тому же уровню и одному и тому же формальному классу и допускающими замену (субституцию) одна другой» [1]. Язык есть парадигматическая система, в которой категории, и элементы или признаки сосуществуют одновременно и в комплексе сочета-

ются при образовании единиц. Это такая система форм, при которой одно слово имеет различные формы словоизменения, а каждая из этих форм служит для обозначения определенных одних и тех же синтаксических отношений. В то же время тождественные своими формальными показателями словоизменения разных слов одного класса служат для обозначения тождественных синтаксических отношений, в которые вступают эти слова. Таким образом, в результате сравнения парадигматических систем английского и азербайджанского языков были выявлены следующие дифференции и универсалии:

1) Морфонологические изменения в результате присоединения словоизменительных аффиксов имеют место, как в английском языке, так и в азербайджанском.

2) Выпадение гласных звуков и удвоение согласных звуков характерно для английского языка, тогда как выпадение гласных и согласных, чередование согласных при присоединении соответствующих аффиксов имеет место в азербайджанском языке.

3) Заглавное слово парадигмы считается отправной точкой для описания остальных форм той же парадигмы. В английском и азербайджанском языках существительные в им. п. могут выступать в чистом виде в качестве заглавного слова, благодаря чему становится организующим центром всей парадигмы склонения. Структура парадигмы, т.е. такая, в основе которой лежит один структурный стержень, оказала влияние на характер фонетических процессов (тенденция к сохранению четких границ между морфемами, препятствие к деформации самой оси парадигмы, к деформации основы слова и т. д.).

4) Спутником агглютинации в азербайджанском языке является сингармонизм. Наличие гармонии гласных и связанное с ней противопоставление переднеязычных согласных заднеязычным, отсутствие в исконно тюркских словах сочетаний нескольких согласных в начале слова, на стыках морфем или в абсолютном исходе слова, особая типология слогов обусловливают относительную простоту дистрибутивных отношений фонем в данном языке. Такого явления в английском языке не существует.

5) Категория имени в английском языке имеет 2 падежа (именительный и притяжательный), тогда как в азербайджанском их 6 (Именительный, родительный, дательный (направительный), винительный, местный и исходный падежи).

6) Категория рода в обоих языках отсутствует, разве что при употреблении одушевленных лиц и в редких случаях применение личных местоимений he или she для какого-то вида собственности и домашних питомцев. Гендерной зависимости в сочетаниях, где существительное согласуется с прилагательным и глаголом в мужском, женском или среднем роде, как в русском языке не существует.

7) В азербайджанском языке множественное число выражается при помощи аффикса *–lar/lər*, тогда как в английском языке данная категория имеет более сложную аффиксальную систему, случаи чередования гласных в корне слов или же отсутствие изменений.

8) Категория принадлежности в азербайджанском языке передается при помощи системы личных аффиксов, присоединяемых к основе. В английском языке данная категория образуется присоединением ‘s, s’ или посредством предлога of.

9) В состав числительных обоих языков входят лексические единицы для обозначения чисел первого десятка, для чисел двадцать, тридцать, сорок, пятьдесят, сто, тысяча; для чисел шестьдесят, семьдесят, восемьдесят и девяноста употребляются сложные слова, первая часть которых представляет фонетически видоизмененные названия соответствующих единиц первого десятка.

10) Парадигма личных местоимений обоих языков включает формы трех лиц ед. и мн. ч., при их склонении в азербайджанском языке происходят изменения гласного основы.

11) Грамматическая категория такого определителя существительного как артикль в Азербайджанском языке не нашла свое выражение.

12) Роль предлогов в английском языке выделена особо, так как за отсутствием падежных форм слова и отношения между ними выражаются посредством предлогов. Для азербайджанского языка характерны послелоги, и они немногочисленны.

13) Сравнительные степени прилагательных и наречий в английском языке строятся как присоединением аффиксов (-er,-est), так и добавлением вспомогательных элементов more/most перед прилагательным или наречием. В азербайджанском языке данная категория строится посредством вспомогательных элементов (daha/ən) перед прилагательным или наречием. Это единичный

случай аналитического элемента именных категорий в азербайджанском языке.

14) Парадигма английского глагола имеет разветвленную систему – около 64 форм изменений основной формы глагола. Для образования глагольных временных форм применяется как метод флексии, так и метод присоединения вспомогательных элементов (**is** com**ing, will** come, **would have been** *done*). В парадигме глагола азербайджанского языка около 66 грамматических форм, образующиеся при помощи агглютинации. Данные цифры приводятся, во-первых, с учетом переходности глаголов, а также учитываются как утвердительные, так и отрицательные глагольные формы.

15) В английском языке существует 3 вида наклонения: изъявительное, повелительное и сослагательное. В азербайджанском языке помимо изъявительного, повелительного и условного (сослагательного) наклонения имеется желательное наклонение и долженствовательное наклонение. В английском языке желательное и долженствовательное наклонения можно отождествить с модальными глаголами, выполняющими функции как вспомогательных так смысловых глаголов.

16) Оба языка имеют действительный и страдательный залоги, однако в азербайджанском языке помимо вышеуказанных существуют также взаимно-совместный, возвратный и понудительный залоги.

17) Неличные формы глагола исследуемой пары языков также совпадают в понятии инфинитива – неопределенной формы глагола и причастия. В азербайджанском языке имеется также и деепричастие, однако, в английском языке неличных форм гораздо больше и их использование и распознавание в языке гораздо сложнее. Для последнего характерными понятиями являются причастия прошедшего и настоящего времени, герундий (субстантивированный глагол), отглагольное существительное.

18) Твердый закон порядка слов «определение + определяемое» проявляется в структуре всех синтаксических категорий азербайджанского языка. Любое словосочетание в тюркских языках строится по принципу «зависимый член + главный член» (*yaxşı qız- yaxşı yazır*)

19) Способность имен существительных выступать в роли определения при отсутствии согласования с определяемым

явилась почвой для развития в азербайджанском языке особого типа словосочетаний – изафета. 1-ый тип изафета основан на простом примыкании *dəmir qapı.* В них развиваются адъективизированные сочетания слов. В английском языке также имеет место понятие адъктивации существительных, например, stone wall, где наблюдается левосторонняя сочетаемость существительного с существительным-прилагательным. На базе отношений *«существительное+ существительное»,* то есть «определение + определяемое» в тюркских языках развивались притяжательные аффиксы, синтаксические конструкции по второму типу изафета *tibb bacısı*.

20) В английском предложении строго фиксированный порядок «подлежащее-сказуемое-дополнение-обстоятельство». В азербайджанском языке тоже строгий порядок слов «подлежащее – дополнение-обстоятельство-сказуемое». Конечная позиция глагола в азербайджанском языке и вытекающая отсюда препозиция дополнения, включающая развернутые конструкции, продиктованные тем же законом порядка слов «определение + определяемое». Преобладание именного строя предложения, замена личных глагольных форм отглагольными именами при господстве способа примыкания обеспечивали на ранних этапах развития агглютинативных языков структуру тесно спаянных комплексов «объект + отглагольное имя».

В процессе исследования морфологической структуры рабочей пары языков для создания морфосинтаксического анализатора, в первую очередь, необходимо определить задачи, методы, и наконец, этапы функционирования блока морфологического анализа.

Потребность в морфологических операциях над текстом возникает в системах самого различного масштаба и назначения, обрабатывающих текст на естественном языке. Достаточно назвать задачи автоматического и автоматизированного перевода, извлечения информации из текста, автоматического индексирования баз данных в информационно-поисковых системах, сжатия текстовых баз данных, проверки грамматической правильности текста, создания электронных словарей и обучающих систем, проведения лингвистических исследований, и многие другие задачи.

Под морфологическими операциями понимаются операции над текстом, связанные с явлением словоизменения (более точно – морфологического словоизменения), то есть образования различных форм слов. Образование форм слов свойственно пода-

вляющему большинству крупных языков мира, в частности, всем европейским языкам. Даже в языках со слабо развитой системой словоизменения большая часть слов образует несколько различных форм, называемых словоформами данного слова. Например, англ. *book – books – book's – books', ask – asks – asked – asking*. В азербайджанском языке типичное прилагательное – до 10 форм, глагол – около 70 форм. Так, в информационных системах любая информация обычно привязана к собственно слову, лексеме, а не к отдельной словоформе. Однако в текстах, обрабатываемых системой, слова представлены в виде различных словоформ. Первой задачей морфологического анализатора является сведение всех форм слова к единому виду, представленному либо числом, либо стандартизированной буквенной цепочкой. Морфологическая система должна «вычислять» различные формы слов. Более того, наличие в системе правил изменения слов играет принципиальную роль при решении некоторых задач морфологической обработки текста. Такие правила действительно существуют в естественном языке и могут быть сформулированы в виде, необходимом для автоматического образования словоформ.

Например, для азербайджанского языка, правила состоят в присоединении к соответствующей цепочке справа буквенных цепочек. Соответственно, правило автоматического определения слова по словоформе состоит в отделении такой цепочки от словоформы. Языки с подобным типом словоизменения называются – агглютинативным – типом словоизменения, Однако формально он может рассматриваться как упрощенный частный случай флективного.

При сопоставительном анализе английского и азербайджанского языков были выявлены те признаки их частей речи, которые возможно представить в формальном виде в базе данных, представленной автоматическим словарем, и на основе которых может быть создана база знаний морфосинтаксического анализатора. Однако, следует отметить, что для каждой части речи помимо информации, собранной в базе данных АС, приписывается определенное количество правил распознавания и порождения грамматических форм. Правила группируются по категориям, характерным для той или иной части речи, со своими списками исключений, дополнений и т.д. Таким образом, строятся такие рекурсивные правила анализа словоформ входного языка, которые трансфор-

мированы в виде формул, с последующим выходом, то есть конечным синтезом словоформ на основе правил выходного языка.

Морфологическая система должна обладать моделью морфологического строения данного естественного языка. Кроме того, для описания свойств отдельных слов система должна обладать словарем большого объема. Однако для практически работающей морфологической системы существенным является не столько само наличие такого аппарата, сколько высокоэффективная программная реализация алгоритмов, решающих на его основе разнообразные задачи обработки текста.

4. Особенности формальных признаков частей речи для базы данных автоматического словаря

Фундаментальным этапом при построении базы данных (БД) морфосинтаксического анализатора является отбор формальных признаков морфосинтаксических систем рабочей пары языков для их ввода в словарную статью. Структура автоматического словаря в системе МП обеспечивает наиболее тщательную обработку переводимого текста, введенного в СМП. Обосновано, что основным эффективным методом отбора, которым следует руководствоваться при построении автоматического словаря, функционирующего в составе системы машинного перевода, для любой пары языков является установление переводных соответствий на всех языковых уровнях. Разработка такого метода позволяет выбрать наиболее точные переводные соответствия на всех языковых уровнях, и тем самым обеспечить оптимальность структуры автоматического словаря. При определении понятия переводного соответствия автор исходит из той посылки, что переводное соответствие выражает не только переводной эквивалент слова с одного языка на другой, так как этот фактор относится только к лексическому значению слова в языковой иерархии. В предложенном подходе установление переводных соответствий для грамматических категорий, синтаксических конструкций и морфосинтаксических функций слов, словосочетаний и предложений рабочей пары языков является необходимым условием, обеспечивающим адекватность и точность перевода. Чем шире спектр установленных переводных соответствий с учетом многозначности и многофункциональности языковых единиц, тем

полнее будет информация, вводимая в словарь в виде кодов, что предопределяет условия для получения более точного перевода. В исследуемой паре языков основу морфологической системы составляют ее части речи. Так, в азербайджанском языке части речи по смысловому содержанию, принадлежности к той или иной системе образования и изменения морфологических особенностей и их синтаксических функций делятся на главные, служебные и специальные, тогда как в английском языке они делятся на самостоятельные и служебные. В машинной морфологии части речи подразделяются на 2 группы по признаку отношения к словоизменительным аффиксам: 1. Изменяемые – это части речи, в которые входят словоформы, содержащие словоизменительные аффиксы. К ним относятся главные/ самостоятельные части речи.

2. Неизменяемые – которые охватывают словоформы, не содержащие словоизменительные аффиксы. К ним относятся служебные/специальные части речи.

Таблица 1. **Части речи азербайджанского языка**

Главные	**Служебные**	**Специальные**
Существительное	Послелог	Звуко-образо подражательные слова
Прилагательное	Союз	Междометия
Наречие	Частицы	Модальные слова
Числительное		
Глагол		
Местоимение		

Таблица 2. **Части речи английского языка**

Самостоятельные	**Служебные**
Существительное	Предлог
Прилагательное	Вспомогательные глаголы
Местоимение	Артикль
Числительное	Союз
Глагол	Междометия
Наречие	

Формальные признаки частей речи двух языков, английского и Азербайджанского, были изучены параллельно и их особенности ниже изложены в комплексе.

Каждая часть речи имеет характерные ей признаки, определяющие ее место и функции в машинной морфологии. Эти признаки можно классифицировать на 4 группы:

– морфонологические;
– морфологические;
– синтаксические;
– семантические.

Морфонологические признаки – разделение определенных частей речи на группы, оканчивающиеся на согласную или гласную морфонему, или сочетание фонем, ведущие к преобразованиям, выпадению или чередованию фонем. Они составляют граничную зону между морфологией и фонологией, однако, для каждого отдельного языка составляет особую и самостоятельную область грамматики. В английском языке наблюдаются следующие морфонологические процессы:

1. При образовании степеней сравнения односложных прилагательных с предшествующим кратким гласным, в сравнительной и превосходной степени имеет место редупликация – удвоение согласных (big – bi**gg**er).

2. При образовании степеней сравнения если прилагательное оканчивается на немое **e**, то при прибавлении -er, -est немое **e** опускается, иными словами выпадает. (larg**e**- larg- *er*).

3. При образовании степеней сравнения если прилагательное оканчивается на **y** с предшествующей согласной, в сравнительной и превосходной ***y*** переходит в **i** (bus**y**-bus**i**er), то есть происходит замена или альтернация звуков.

4. При образовании множественного числа существительных в английском языке имеются случаи замены, если слова оканчивается на –y с предшествующей согласной, словоформа принимает аффикс [-es], вследствие чего словоформа заканчивается на [-ies], то есть ***y*** переходит в **i** (lad**y**-lad**i**es).

5. При образовании форм множественного числа некоторых существительных имеют место замены гласной в корне слов (man-men).

6. При образовании Причастия 1, то есть причастия настоящего времени е односложных глаголов английского языка, к основе

глагола прибавляется суффикс –ing, в следствие чего происходит редупликация конечного согласного (put-pu**tt**ing), если же глагол заканчивается на немое **e**, то происходит ее выпадение.

7. При образовании формы глагола прошедшего времени и причастия прошедшего времени имеет место образование супплетивных форм, когда изменятся вся основа(go-went). В этих категориях также наблюдаются случаи нулевой мены, то есть начальная форма остается не измененной (cut-cut-cut) альтернации гласных в основе (do-did).

8. В азербайджанском языке, в отличие от английского языка, действует закон сингармонизма. Сущность этого закона заключается в следующем:

– гласные азербайджанского языка делятся на мягкие (ə,ö,ü,i) и твердые (a,o,u,ı);

– согласные азербайджанского языка делятся на мягкие и твердые;

– если первый слог содержит гласный твердого ряда, то все остальные гласные также принадлежат твердому ряду;

– если в первом слоге содержит гласный мягкого ряда, то все остальные гласные в слове также принадлежат мягкому ряду;

*– мягкие гласные согласуются с мягкими согласными (**qa̲ra̲**-черный, **kül̲ək** – ветер);*

– если в первом слоге входит гласный твердого ряда, то гласные аффиксы, присоединяемые к слову, должны относиться к этому ряду;

– если словоформа относится к исключениям, то сингармонизм прибавляемого аффикса будет определяться по гласной в последнем слоге.

9. Закон сингармонизма применим также и для аффиксов (наличие двух вариантов как для гласных твердого и мягкого ряда). Описанную выше закономерность можно выразить следующим образом:

*– Если аффикс обозначен индексом, например, в виде **-lar²** ,это означает, что при присоединении этого аффикса учитывается парадигма сингармонизма: для гласных твердого ряда -lar, для гласных мягкого ряда – lər (an̲a̲**lar ,** kül̲ək**lər**)*

*– Если аффикс обозначен индексом, например, как **-in⁴** ,это означает, что при присоединении этого аффикса учитывается*

следующая парадигма сингармонизма ***-in, -ın, -un,-ün*** *...(məclis**in**, anan**ın**, qurum**un**)*

Следует отметить, что поскольку закон сингармонизма также действителен для всех аффиксов азербайджанского языка, то это значит, морфонологические и фонетические признаки имеют не второстепенную значимость при синтезе перевода.

Морфологические признаки составляет совокупность грамматических категорий той или иной части речи, способы образования форм данных категорий и пути объединения их в смысловое единство.

1. Морфологические признаки существительного в английском языке сводятся к категории числа и падежей в качестве морфологических признаков, тогда как в азербайджанском языке к указанным категориям можно добавить 6 падежей и категорию сказуемости и принадлежности. В английском языке категория сказуемости не имеет такую важность, как в азербайджанском языке из-за многочисленности и изменяемости флексий.

2. Категория принадлежности одушевленных существительных английского языка, как известно, образуется при помощи апострофа и аффикса 's, и выражает принадлежность объекта данному существительному (sister’s room). Данная категория также характерна и для азербайджанского языка. В продукции словосочетания «Существительное в категории принадлежности + Существительное» при переводе на азербайджанский язык сочетаемые слова претерпевают значительные грамматические изменения. Порядок словосочетания N’s +N для английского языка совпадает с порядком в азербайджанском языке с некоторыми элементами агглютинации.

3. Категория сказуемости в азербайджанском языке характерна не только для существительных, но и для всех именных частей речи. Английский глагол *to be* является как смысловым, так и вспомогательным глаголом, выполняющим соединительную и формообразовательную функции. В азербайджанском языке данному глаголу соответствуют аффиксы категории сказуемости, которые присоединяются к именной части речи. Последняя может быть представлена существительным, прилагательным, числительным, местоимением, причастием и даже конструкциями.

4. Некоторые местоимения английского языка имеют отдельные формы для единственного и множественного числа (указа-

тельные местоимения: this -этот, these – эти). Некоторые местоимения имеют одну и ту же форму для единственного и множественного числа: **all** – весь, все. Некоторые местоимения имеют значение только одного числа: единственного **each** – каждый, **somebody** -кто-то, или множественного числа **both** – оба.

Местоимения английского языка можно поделить на местоимения-существительные и местоимения-прилагательные. Представители первой группы имеют категории рода, числа, падежа, категорию принадлежности. Так, личные местоимения имеют формы именительного и объектного падежей. В объектном падеже личные местоимения соответствуют местоимениям в винительном, в дательном падеже, а иногда и в творительном падежах. Различение принадлежности переводного эквивалента на выходном языке зависит от синтаксического местоположения и функции местоимения в предложении. Притяжательные местоимения-существительные обычно стоят в конце предложения и не принимают после себя существительных. Практически все категории местоимений нашли свои эквиваленты в азербайджанском языке.

5. Указательные, неопределенные и вопросительные местоимения могут иметь функции, как прилагательного, так и существительного. Как и все местоимения-существительные они обладают не только грамматическими признаками существительного, но и в предложении могут выполнять функции подлежащего, дополнения и именной части сказуемого.

6. Местоимения – прилагательные употребляются в функции определения. К ним относятся притяжательные местоимения-прилагательные, указательные, некоторые вопросительные, относительные и неопределенные местоимения. Одни и те же слова в перечисленных категориях могут выполнять функции прилагательных и существительных, что зависит от их места в английском предложении. Почему акцентируется внимание именно на предложении английского языка, так как при изменении строго порядка слов предложения меняется его смысл.

7. Формы английского и азербайджанского глагола делятся на личные (изменяемые) и неличные (неизменяемые). Личные формы выражают лицо (1-ое, 2-ое и 3-е лицо), число (ед.ч. и мн.ч.), наклонения, время и залог. Неличные формы глагола выражают действие без указания лица, числа и наклонения.

8. В английском языке имеются 3 наклонения: изъявительное, сослагательное и повелительное. В азербайджанской морфологии к уже указанным наклонениям прибавляются долженствовательное и желательное. Необходимо отметить, что категория долженствовательного наклонения в английском языке выражается несколькими модальными глаголами, а желательное наклонение выражается отглагольными словосочетаниями модального характера.

9. Оба языка имеют действительный и страдательный залоги, однако в азербайджанском языке помимо вышеуказанных существуют также взаимно-совместный, возвратный и понудительный залоги. Взаимо-совместный залог обозначает действие, которое выполняется не только со стороны субъекта, в роли подлежащего, но и одновременно выполняется со стороны объекта, в роли дополнения. Например: Он встретился с другом. – O dostu ilə görüşdü. Взаимо-совместное действие образуется при помощи частицы *ilə* и аффикса iş[4]. В английском языке это наклонение выражается посредством сочетания обычных глаголов и предлога with или же сочетанием глаголов и взаимных местоимений each other / one another, а иногда просто при помощи глагола.

Возвратный залог обозначает действие, совершаемое субъектом в отношении самого себя. В английском языке этот залог выражается просто глаголом или сочетанием глагола с возвратными местоимениями. Понудительный залог обозначает, что действующее лицо совершает действие по инициативе другого лица. В английском языке значение понудительного залога передается при помощи сочетания инфинитива со словами make, а иногда ask, order.

10. В таблице 3 приведены результаты сравнительного анализа категорий английского и азербайджанского глагола, где даны категории и способы их образования и выражения в обоих языках. Для выражения времени совершения действия – настоящего, прошедшего и будущего – английский глагол имеет своеобразную систему глагольных времен, которые подразделяются на группы: простое/неопределенное, продолженное/прогрессивное, совершенное и совершенное длительное. Каждая группа склоняется в настоящем, прошедшем, будущем и в будущем в прошедшем. Итого в английской вербальной системе имеется 16 временных категорий. В азербайджанском языке имеются 5 вре-

менных категорий (настоящее время, прошедшее результативно-повествовательное, прошедшее категорическое время, будущее и будущее некатегорическое время).

Таблица 3. **Сравнительная характеристика категорий глагола в английском и азербайджанском языках**

Категория глагола Английского языка	Формальное представление	Категории глагола Азербайджанского языка	Формальное представление
Present Simple (настоящее простое) Present continuous (настоящее продолженное время)	$S + V_I / V_s$ S+ (am, is, are)+ Participle I	Настоящее время	S+(imperative) V^A+**ır**^{4}**(yir)**^{4} + personal affix
Present Perfect (настоящее завершенное время)	S + have + Participle II	Прошедшее категорическое время	S+(imperative) V^A+**dı**^{4} +personal Affix
Past Simple (прошедшее простое время)	$S+ V_{II}/V_{ed}$	Прошедшее результативно-повествовательное время	S+(imperative) V^A+**ib**^{4}**(mış)**^{4} + personal Affix
Future Simple (будущее простое)	S+will /shall + V_I	Будущее категорическое время	S+(imperative)V^A+ **acaq**^{2}+ +personal Affix
Past perfect (прошедшее завершенное время)	S+ have + Participle II	Прошедшее категорическое время в повествовании	S+(imperative) V^A+(**mış**)^4+**dı**^{4}+ personal Affix
Past continuous (Прошедшее продолженное время)	S+ (was, were)+ Participle I	Настоящее время в повествовании(в прошедшем)	S+(imperative) V^A+**ır**^{4}**(yir)**^{4}+ **dı**^{4} + personal affix

Modal verb (модальные глаголы)	Must/ ought to/ have to + V	Долженствовательное наклонение глагола	S+(imperative) V^{A}+**məli**^{2} +personal Affix
Modal verb (модальные глаголы)	Should/had to/ needed	Желательное наклонение в прошедшем	S+(imperative) V^{A}+**məli**^{2}**+idi**^{4} + personal Affix
Verbal phrase (глагольные сочетания)	Used to	Некатегоричное будущее время в прошедшем	S+(imperative) V^{A}+**ərdi**^{2} + personal Affix
Verbal phrase (глагольные сочетания)	To be going to, to be to	Желательное наклонение	S+(imperative) V^{A}+**ar**^{2}+ +personal Affix
Verbal phrase (глагольные сочетания)	To make smb. do smth	Понудительный залог	S+(imperative) V^{A}+**dır**^{4}+ +personal Affix

Эти группы в свою очередь имеют формы в прошедшем и условном (предполагаемом) повествовании. Не все временные категории английского языка нашли свой эквивалент в азербайджанском языке.

11. В формулах, показанных в таблице 4, имеются символы *Vı, Vıı, V_{ed}, $V_{imperative}$, Participle I, Participle II.* Глаголы английского языка имеют 4 формы: инфинитив – неопределенная форма глагола, выражаемая при помощи частицы **to** и без нее, обозначенная в таблице $\mathbf{V_1}$; вторая форма – прошедшее неопределенное время V_{ed}/ V_{ir}, третья форма – причастие прошедшего времени *Participle II*, четвертая – причастие настоящего времени *Participle I*- при помощи этих форм строятся временные категории действительного и страдательного залога и т.д. В азербайджанском языке существуют: инфинитив – неопределенная форма глагола, выражающаяся при помощи суффиксов -maq/mək и форма повелительного наклонения для 2 лица единственного числа – Vimperative, который образуется путем удаления суффиксов -maq/mək инфинитива, другими словами, основа слова, к которой присоединяются цепочки аффиксов, в свою очередь имеющие определенный строгий порядок. Каждая форма глагольной категории имеет определенный

состав аффиксов, при удалении и добавлении которых меняется грамматическое значение слова.

12. При образовании форм страдательного залога в английском языке, используются аналитические средства (вспомогательный глагол *to be* в формах соответствующих временных категорий и *Причастие II*). В азербайджанском языке образование данной категории в машинной морфологии вызывает определенные сложности. При формальном описании порождения этой категории на выходном, то есть азербайджанском языке предполагает склонение личных местоимений и существительных в ед.числе и мн.числе наряду с присоединением соответствующей цепочки аффиксов к императивной форме глагола. Если при переводе одного и того же глагола в разных значениях, переводной эквивалент объекта (существительного или местоимения), над которым осуществляется действие, выраженное глаголом в страдательном залоге на выходном (Азербайджанском) языке требует разные падежи, то при вводе данных этого глагола дается определенное указание на использование аффиксальных форм именно данного падежа.

Например: сочетание *I am called* может переводиться двояко: Mənə zəng edirlər или **Məni çağırırlar.** В обоих вариантах перевода наблюдается одинаковая аффиксальная цепочка образования страдательного залога настоящего простого времени -ırlar[4] однако, падежная форма личного местоимения Mən имеет различные окончания (дательного и винительного падежа).

Синтаксические признаки отражают способы сочетаемости различных частей речи друг с другом, определяют место и функцию словоформы в предложении. Синтаксические признаки несомненно тесно связаны и с морфологическими и семантическими признаками языка. Они взаимодействуют в системном единстве друг с другом и поэтому взаимозависимы. При описании функционирования частей речи друг с другом в правилах распознавания и порождения той или иной грамматической категории, применятся подход разделения предложения на все виды словосочетаний, составные элементы которых находятся в определенных парадигматических отношениях.

К синтаксическим признакам для автоматической обработки простого предложения в составе текста относятся:

– часть речи и ее функция в предложении;

– сочетаемость различных частей речи друг с другом;

– виды словосочетаний;

– порядок слов в простом предложении.

1. Существительные в обоих языках характеризуется правосторонней сочетаемостью с существительным в притяжательном падеже, глаголом и левосторонней сочетаемостью с существительным, предлогом, местоимением, прилагательным, числительным, причастием и с артиклем для английского языка. В левосторонней сочетаемости двух существительных в исследуемой паре языков происходит процесс адъективации существительного, и оно может служить определением стоящему после него существительному. Во многих случаях существительному могут предшествовать не одно, а несколько существительных. Например, life insurance – həyat sığortası (страхование жизни), stone wall – daş divar (каменная стена).

2. При сочетании существительного с глаголом основными отношениями являются взаимозависимость подлежащего и сказуемого в различных временных категориях в действительном и страдательном залоге. При описании этих отношений в правилах распознавания форм английского языка предлагаются различные сочетания существительных или местоимений во мн.ч и ед.ч с парадигматически измененными формами глагола для каждого лица, с последующим порождением эквивалентов данного грамматически связанного сочетания на выходном языке. Например: При сочетании местоимения I и исходной формы глагола V без частицы **to**, верно сочетание, переводного эквивалента (ПЭ) местоимения и эквивалента азербайджанского глагола, к которому прибавляются суффиксы $–ır^4 + – əm^2$, если основа ПЭ глагола оканчивается на согласную букву. Формальное представление образования времен глагола в таком виде способствует созданию облегченного варианта блоков синтаксического анализа словосочетаний в составе одного предложения.

3. Особое место в морфосинтаксических отношениях имеет перевод сочетаний предлогов с существительными, так как в английском языке они заменяют падежные окончания для соединения слов в предложении, тогда как в азербайджанском языке предлогов как таковых нет. Предлоги заменяются падежными окончаниями в сочетании с послелогами или же наречиями.

Таблица 4. **Соответствие предлогов английского языка падежным формам азербайджанского языка**

Косвенные падежи азербайджанского языка	**Предлоги английского языка,**
Родительный падеж	Across, before, into, of, off, around, past, under
Дательный падеж	To, for, towards
Винительный падеж	By
Местный падеж	Against, at, in, inside, on, during
Исходный падеж	About, after, except, from, out of, since, through, within

При сопоставлении предлогов, было установлено, что каждый предлог английского языка соответствует одному из шести падежей азербайджанского языка.

4. Предложные сочетания существительных играют важную роль при анализе предложения. О строгой фиксированности порядка слов в английском предложении мы уже упоминали. Дополнение английского языка имеет 3 вида, представленных на схеме 2: прямому дополнению в азербайджанском языке соответствует существительное в винительном падеже, которое при отсутствии других его видов стоит после сказуемого. *I have sent* ***a letter***. – Mən *məktubu* göndərdim. Косвенное дополнение без предлога соответствует существительному в дательном падеже, и всегда стоит перед прямым дополнением. *I have sent John a letter.* Порядок слов в простом предложении выглядит следующим образом:

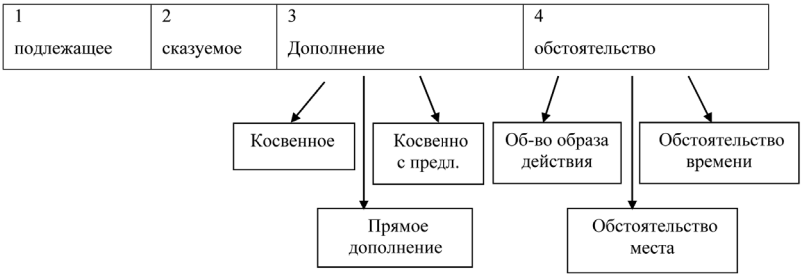


Схема 1. Порядок слов предложения в английском языке

Mən *məktubu* ***Cona*** göndərdim. Косвенное дополнения с предлогом может выражается различными сочетаниями, а иметь различные падежные окончания в Азербайджанском языке. Оно стоит всегда после прямого дополнения. *I have sent letter to John*. *Mən məktubu* ***Cona*** göndərdim.

Все три вида дополнений не могут одновременно находиться в составе одного предложения, либо прямое с косвенным без предлога, либо прямое дополнения в предложным косвенным в паре.

5. Обстоятельство места и времени помимо наречий, также может выражаться предложными сочетаниями (*in the morning* – **səhər çağı**, *at the corner*- **tində**). Структура азербайджанского предложения имеет более усложненную структуру, как во всех тюркских языках, с глаголом в конечной позиции, соответственно местоположение дополнений и обстоятельств также варьирует.

Таким образом, морфо-синтаксический анализ предложения посредством именных и глагольных словосочетаний значительно упрощает процесс перевода. Для моделирования подобных объектов и процессов в математике используется абстрактная фигура – дерево. Предложение естественного языка также имеет ветвящуюся иерархическую структуру: дерево. Подчинительные связи слов образуют дерево. Морфосинтаксический анализатор позволяет рассмотреть подчинительные связи и грамматические значения слов в предложении. Он реализует универсальные языковые зависимости, что позволяет применять его для разработки анализаторов на единой платформе.

5. База знаний ЭСПМП и ее взаимодействие с базой данных

База знаний содержит информацию, необходимую для решения задач требуемого типа, в виде правил и фактов. Механизм вывода представляет собой общий алгоритм решения задач, реализуемый, как правило, в виде интерпретатора. Применение его к базе знаний о конкретной предметной области, задаваемой экспертом, и к данным о текущей ситуации, задаваемой пользователем, дает решение требуемой задачи.

Интерфейс с пользователем предназначен для взаимодействия с ним во время решения задачи, и в зависимости от типа задачи, может использовать средства анализа фраз на естественном

языке, выбора меню, графического ввода и вывода. Задача лингвиста формулировка лингвистических правил – правил языка и речи. Моделирование лингвистических правил как базового набора предопределяет соответствующую архитектуру экспертной системы.

Функционирование АС в качестве базы данных во взаимодействии с базой знаний экспертной системы поддержки машинного перевода базируется на алгоритме автоматической обработки текста, этапы которой описаны ниже:

– Предложение вводится в программу на языке оригинала и после начального анализа поступает в базу данных для идентификации словоформ или словосочетаний, имеющихся в составе автоматического словаря.

– Далее предложение поступает в базу знаний, в которой посредством правил распознавания происходит морфологический анализ словоформ, установление грамматических форм слов и отделение суффиксов или других грамматических показателей от основы.

– Затем установленные основы возвращаются в базу данных, где находят свои переводные эквиваленты.

– Найденные эквиваленты посредством правил подстановки подставляются в соответствующие формализованные синтаксические структуры, зафиксированные в базе знаний экспертной системы в виде взаимозависимых кодовых цепочек грамматических форм слов, т.е. установленные грамматические формы словоформ поставляются в соответствующую контексту формализованную синтаксическую единицу(словосочетание иди предложение).

– На последнем этапе текстовой обработки с помощью правил порождения осуществляется синтез составляющих предложения на переводном языке. При этом порождающая грамматика с элементами грамматики непосредственно составляющих и грамматики конечных состояний позволяет решать задачу исчерпывающего описания множества правильных предложений языка, одновременно детально указав действующие в нем синтаксические закономерности.

В проведенном исследовании составление правил базы знаний основывается на методике сочетаемости слов друг с другом в процессе образования их грамматических форм, иными сло-

вами, распознавание тех или иных грамматических форм слов в предложении и выявление их морфологической и синтаксико-семантической принадлежности возможно осуществить в рамках сочетаний этих словоформ друг с другом. Синтаксическое разделение предложения на словосочетания с внутренними подчинительными связями, редко согласованием и примыканием, предопределяет работу синтаксического блока анализа.

При создании ЭСПМП необходимо учитывать следующие аспекты хранения и переработки знаний:

– содержание знаний;

– репрезентация знаний, форма хранения, предназначенная для эффективного поиска, переупорядочения и модифицирования;

– формализация объектного знания (в нашем случае – формализация знаний о языке-объекте.);

– переработка текстов как последовательности слов;

– интерпретация синтаксических структур и перевод их в лексическое представление при учете контекста;

– представление знаний внутри самой системы;

– разработка вспомогательных средств для формализации, хранения и поиска знаний при обработке показаний экспертов.

База знаний осуществляет работу механизма ЭСПМП на основе трансформационной (порождающей) грамматики, представляющей собой систему правил, экспериментальным образом приписывающую предложениям структурные описания.

Любая трансформационная грамматика имеет в своем составе морфосинтаксический, морфонологический и семантический компонент. Морфосинтаксический компонент определяет бесконечное множество абстрактных формальных объектов, каждый из которых включает в себя всю информацию, существенную для одной интерпретации конкретного предложения. Морфонологический компонент определяет фонетическую форму предложения, порождаемого синтаксическими правилами. Он соотносит структуру, порождаемую синтаксическим компонентом, с фонетически репрезентированным сигналом. Семантический компонент определяет семантическую интерпретацию предложения. Он соотносит структуру, порождаемую морфосинтаксическим компонентом, с определенной семантической репрезентацией.

Следовательно, морфосинтаксический компонент грамматики должен указывать для каждого предложения глубинную структу-

ру, обусловливающую его семантическую интерпретацию и поверхностную структуру, которая определяет его фонетическую интерпретацию.

Основополагающей идеей трансформационной грамматики является идея о том, что поверхностная структура задается неоднократным применением определенных формальных операций, называемых «грамматическими трансформациями», к объектам элементарного вида. База морфосинтаксического компонента – система правил, порождающая конечное множество базовых цепочек, каждое из которых имеет связанное о ней структурное описание, называется базовым показателем структуры составляющего. Эти базовые показатели являются элементарными единицами, составляющими глубинные структуры. В основе предложения лежит последовательность базовых показателей, каждый из которых порождается базой синтаксического компонента. Общий смысл предложения зависит не только от смысла его слов, но и от синтаксической структуры предложения. Синтаксическая структура предложения – это совокупность сведений о связях между его словами и словосочетаниями.

В нашем исследовании составление правил базы знаний основывается на методике сочетаемости слов друг с другом в процессе образования их грамматических форм. Иными словами, распознавание тех или иных грамматических форм слов в предложении и выявление их морфологической и синтаксико-семантической принадлежности возможно осуществить в рамках сочетаний этих словоформ друг с другом. Синтаксическое деление предложения на словосочетания с внутренними подчинительными связями, редко согласованием и примыканием предопределяет работу синтаксического блока анализа.

6. Виды правил базы знаний ЭСПМП

Знания ЭСПМП представлены набором правил, каждое из которых состоит из: антецедента (условия) и консеквента (результата) [53]. На простом языке пользователя правило состоит из правой и левой части. Знания ЭСПМП представляют собой комплекс правил унификационной грамматики, которая включает в свой состав элементы грамматик разных видов, таких как: контекстно-свободная грамматика (КСГ), обеспечивающая морфологический

анализ и синтез, и являющаяся основой анализаторов, цепочечная грамматика (ЦГ) и грамматика непосредственно составляющих (ГНС), обеспечивающие синтаксический анализ и синтез.

Так, элементы КСГ формализуют описание языковой модели как формальной грамматики с конечным числом состояний. Элементы ЦГ фиксируют порядок следования объектов цепочки формально языковой модели, то есть линейные структуры предложения формальной языковой модели, заданные в терминах грамматических классов слов. Применяемая в ЭСПМП стратегия анализа «слева направо»: перебор слов, проверка условий, наличие или отсутствие изменений по условиям и добавление недостающих элементов формально представляет собой реализацию на ПК грамматики с конечным числом состояний или КСГ, построенной на ЦГ. В базе знаний системы ЭСПМП синтаксическая структура предложения может быть представлена деревом синтаксического согласования или подчинения линейных узлов, т.е. слова в предложении находятся в несимметричных отношениях друг к другу: одни слова подчиняют себе другие. Формальное подчинение состоит в том, что одно слово определяет грамматическую форму другого. Синтаксическая структура предложения может быть представлена деревом синтаксического подчинения или просто деревом подчинения, заданным на множестве словоформ предложения.

С учетом морфосинтаксических и семантических признаков английского и азербайджанского языков правила в базе знаний представлены следующими видами сочетаемости:

– существительное в им.п.+ существительное в притяжательном падеже

(категория принадлежности);

– предлог+ артикль +существительное;

– предлог+существительное (падежные эквиваленты в азербайджанском языке);

– существительное + to be (категория сказуемости существительных);

– существительное + ед.ч/мн.ч + глагол (временные формы глагола страдательного и действительного залога)

Собственно-синтаксические правила разделяются на именную, предложную и глагольную группы. В формальном описании правил сочетания можно разделить на:

– правила распознавания;
– правила порождения;
– правила подстановки.

Например, к правилам распознавания можно отнести описание образования множественного числа существительных в английском языке, выявляющимися посредством обнаружения тех или иных аффиксальных изменений, при выполнении заданных условий.

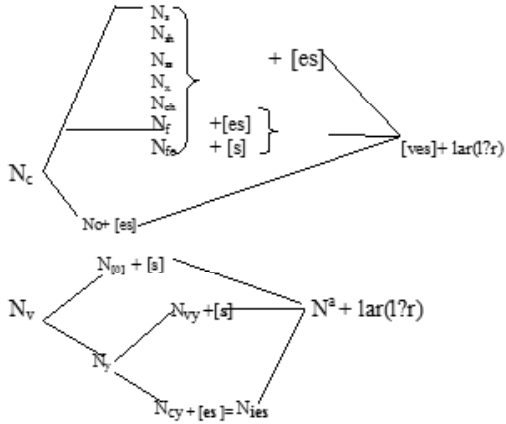


Рис. 2. Схема порождения правил распознавания словоформ множественного числа в англо-азербайджанской среде

– Рисунок 3 схематически демонстрирует особенности функционирования существительного при переводе с английского языка на азербайджанский. Так, расшифруем данную кодировку: существительное оканчивается на согласную/гласную фонему и принимает словоизменительный аффикс в следующих случаях:

– если существительное оканчивается на -s, – ss, -sh, -ch, -x , то оно принимает аффикс [-es];

– если существительное оканчивается на -f, то оно принимает аффикс [-es] вследвие чего, словоформа заканчивается на [ves];

– если существительное оканчивается на -fe, то оно принимает аффикс [-s] вследствие чего словоформа заканчивается на [ves];

– если существительное имеет нулевое окончание, то есть оно принимает аффикс

– [-s];

– если существительное оканчивается на –о, то оно принимает аффикс [-es] (список исключений вводится в базу данных);

– если существительное оканчивается на –у, то оно принимает аффикс [-s] при сочетании -у с гласной, а при сочетании –у с согласной, словоформа принимает аффикс [-es], вследствие чего словоформа заканчивается на [-ies].

– Для образования формы множественного числа существительного в азербайджанском языке к переводному соответствию N^a прибавляется аффикс **-lar²**. Это верно для всех указанных продукций, то есть как для слов, как с согласным окончанием, так и с гласным.

Категория принадлежности одушевленных существительных английского языка, как известно, образуется при помощи апострофа и аффикса *'s*, и выражает принадлежность объекта данному существительному *(sister's room)*. В продукции словосочетания *Существительное в категории принадлежности + Существительное* при переводе на азербайджанский язык сочетаемые слова претерпевают значительные грамматические изменения. На рисунке 2 дается раскладка следующих правил:

*– если после существительного англ. языка стоит **'s**, то это существительное в ед. числе и Р.д (в категории принадлежности)*

– после существительного в категории принадлежности должно стоять существительное в исходном состоянии.

– Порядок словосочетания N's +N для английского языка совпадает с порядком в азербайджанском языке, но:

*– если переводной эквивалент азербайджанского языка N^a оканчивается на согласную фонему, то в категории принадлежности принимает аффикс – **in⁴**;*

*– если переводной эквивалент азербайджанского языка N^a оканчивается на гласную фонему, то в категории принадлежности принимает аффикс – **n+ in⁴**;*

– в переводном соответствии для словосочетания двух существительных (1)N's +(2)N в азербайджанском языке первое существительное(1)N принимает окончания род.п., тогда второе (2)N принимает окончания винительного падежа;

*– если существительное аз. языка имеет форму вин.п. и ед.ч, то оно принимает аффикс – **i⁴**;*

– если существительное аз. языка имеет форму вин.п. и мн.ч, то оно принимает аффикс множественного числа – ***lar***^{2} *и аффикс вин.падежа –* ***i***^{4}*;*

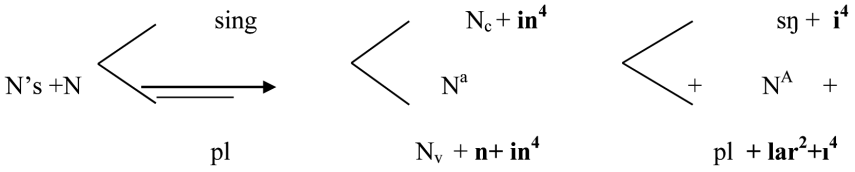


Рис. 3. Схема порождения правил распознавания категории принадлежности существительных в англо-азербайджанской среде

Глагол *to be* является как смысловым, так и вспомогательным глаголом, выполняющим соединительную и формообразовательную функции. В азербайджанском языке данному глаголу соответствуют аффиксы категории сказуемости. Кратко данное правило можно представить так:

Personal pronoun + to be + nominative POS → Personal pronoun + Nom POS^A+ affix of predicativeness,

где именная часть речи может быть представлена существительным, прилагательным, числительным, местоимением, причастием и даже конструкциями.

Правила порождения формальных грамматических форм входного языка с последующим порождением эквивалентов этих форм на выходном языке выполняются после реализации условий, заданных в правилах распознавания и подстановки. Если группы правил распознавания и порождения приурочены к функционированию, в основном, морфологического блока анализа и синтеза, то правила подстановки, в большинстве случаев, применяются для синтаксического блока:

– Сочетание подлежащего + сказуемого, где подлежащее может быть выражено существительным, личным/указательным местоимением, числительным;

– В сочетании определение + подлежащее, где определение может выражаться прилагательным, указательными или притяжательными местоимениями и т.д.

Правила подстановки могут быть и такого вида:

Сочетания существительных с другими частями речи

Правило 4. *Если сочетается местоимение [my] и существительное в единственном числе, то к эквиваленту азербайджанского языка данного существительного, который заканчивается на согласную букву, прибавляется [ım, im, um, üm]:* N^{a}_{o}*+[ım, im, um, üm].*

Правило 5. *Если сочетается местоимение [my] и существительное в единственном числе, то к эквиваленту азербайджанского языка данного существительного, который заканчивается на гласную букву, прибавляется [m]:* $N^{a}_{[-s]}$*+ [m].*

Правило 6. *Если сочетаются местоимение [my] и существительное во множественном числе, то к эквиваленту азербайджанского языка прибавляется lar+[ım, im, um, üm]:* N^{a} *+ lar+[ım, im, um, üm].*

Правило 7. *Если сочетаются местоимение [your] и существительное в единственном числе, то азербайджанскому эквиваленту, который заканчивается на согласную букву, прибавляется [un, ün, ın, in]:* N^{a}_{o}*+[un, ün, ın, in].*

Правило 8. *Если сочетаются местоимение [your] и существительное в единственном числе, то азербайджанскому эквиваленту, который заканчивается на гласную букву, прибавляется [n]:* N^{a}_{v}*+[n].*

Правило 9. *Если сочетаются местоимение [your] и существительное во множественном числе, то к азербайджанскому эквиваленту прибавляется lar+[ın, in, un, ün]:* N^{a}*+ lar+[ın, in, un, ün].*

Правило 10. *Если сочетаются местоимение [his/her/its/their] и существительное в единственном числе, то азербайджанскому эквиваленту* N^{a}_{ε}*, который заканчивается на согласную букву, прибавляется [ı, i, u, ü]:* N^{a}_{ε}*+[ı, i, u, ü].*

Правило 11. *Если сочетаются местоимение [his/her/its/their] и существительное в единственном числе, то азербайджанскому эквиваленту, который заканчивается на гласную букву, прибавляется окончание [sı, si, su, sü]:* N^{a}_{v}*+[sı, si, su, sü].*

Правило 12. *Если сочетаются местоимение [his/her/its/their] и существительное во множественном числе, то азербайджанскому эквиваленту прибавляется lar+[ı, i, u, ü]:* N^{a}*+lar+[ı, i, u, ü].*

Правило 13. *Если сочетаются местоимение [our] и существительное в единственном числе, то азербайджанскому эквиваленту, который заканчивается на согласную букву, прибавляется [ımız, imiz, ümüz, umuz]:* N^{a}_{ε}*+[ımız, imiz, ümüz, umuz].*

Правило 14. *Если сочетаются местоимение [our] и существительное в единственном числе, то азербайджанскому эквиваленту, который заканчивается на гласную букву, прибавляется [mız, miz, müz, muz]:* N^{a}*+[mız, miz, müz, muz].*

Правило 15. *Если сочетаются местоимение [our] и существительное во множественном числе, то азербайджанскому эквиваленту прибавляется lar+[ımız, imiz, ümüz, umuz]:* N^{a}*+ lar+[ımız, imiz, ümüz, umuz].*

В общем случае синтаксические (формальные и смысловые) связи в предложении не сводятся естественным образом к связям между отдельными словами. Поэтому способ представления синтаксической структуры предложения, при котором выделяются группировки слов, связанных друг с другом. Определенным образом устроенное множество отрезков предложения называется

его системой составляющих, а каждый отрезок в этой системе, составляющей предложения. Эта система составляющих предложения может быть представлена графом, который называется деревом составляющих.

Правило 28.

а) Если верно сочетание предлога [about]+сущ. N английского языка, то ему соответствует сочетание азербайджанского эквивалента существительного N^a+аффикс [dan, d?n];

б) Если верно сочетание предлога [about]+сущ. N, то ему соответствует сочетание азербайджанского эквивалента существительного N^a+?trafında.

Правило 29. При сочетании английского предлога [above]+существительное, верно сочетание азербайджанского эквивалента существительного+[ın, in, un, ün] для словоформы, окончившейся на согласную букву+üstünd?.

Правило 30. При сочетании английского предлога [across]+существительное N, верно сочетание азербайджанского эквивалента существительного N^A+[nın, nin, nun, nün] +o biri t?r?fin?.

Правило 31. При сочетании английского предлога [after]+существительное N, верно сочетание азербайджанского эквивалента существительного N^A+[dan, d?n]+sonra.

Правило 32. При сочетании английского предлога [against]+существительное N, верно сочетание азербайджанского эквивалента существительного $N^A{}_C$, который оканчивается на согласную букву + окончание [a, ?].

Правило 33. При сочетании английского предлога [against]+существительное N, верно сочетание азербайджанского эквивалента существительного $N^A{}_C$, который оканчивается на гласную букву, $N^A{}_C$ + аффикс [ya, y?].

Правило 34. При сочетании английского предлога [along]+существительное N, верно сочетание азербайджанского эквивалента существительного N^A + [boyu, uzunu].

Правило 35. При сочетании английского предлога [among/amongst]+существительное N, верно сочетание азербайджанского эквивалента существительного N^A +arasında.

Правило 36. При сочетании английского предлога [at]+существительное N, верно сочетание азербайджанского эквивалента существительного N_t+[da, d?].

Правило 37. При сочетании английского предлога [before]+существительное N, верно сочетание азербайджанского эквивалента существительного N^A + [ın, in, un, ün]+qabağında/?vv?lind?/qarşısında, если переводной эквивалент оканчивается на согласный звук.

Правило 38. При сочетании английского предлога [before]+существительное N, верно сочетание азербайджанского эквивалента существительного N^A + [nın, nin, nun, nün]+qabağında/?vv?lind?/qarşısında, если переводной эквивалент оканчивается на гласный звук.

Правило 39. При сочетании английского предлога [behind]+существительное N, верно сочетание азербайджанского эквивалента существительного, который заканчивается на соглас. прибавляется [ın]^4+arxasında;

Правило 40. При сочетании английского предлога [behind]+существительное N, верно сочетание азербайджанского эквивалента существительного, который заканчивается на глас. букву прибавляется [nın]^4+arxasında;

Правило 41. При сочетании английского предлога [below]+существительное N, верно сочетание азербайджанского эквивалента существительного+dan^2 +aşağı;

Формализация процесса распознавания глагольных форм в морфо синтаксическом анализаторе

Правило 94. *В английском сочетании существительного или личного местоимения в исходной форме с глаголом в форме инфинитива без to+[ ]n-ные суффиксы представляет собой сочетание подлежащего и сказуемого: I live, john? work[ed];*

Правило 95. *При сочетании местоимения I и глагола к его азербайджанскому эквиваленту, заканчивающуюся на согласную букву прибавляется ir+əm[2]: I go – mən gedirəm, I live – mən yaşayıram;*

Правило 96. *При сочетании местоимения I и глагола к его азербайджанскому эквиваленту, заканчивающуюся на гласную букву прибавляется yir[4]+əm[2]: I live – mən yaşayıram;*

Правило 97. *Глагол на английском языке представляет собой основу глагола V_I (məsdər) без to, тогда как на азерб. языке он представляется как основа глагола без окончания maq[2] (live→yaşa, go→get, pass→keç, run→qaç);*

Правило 98. *При сочетании местоимения You и глагола, то к его азербайджанскому эквиваленту, заканчивающуюся на гласную букву прибавляется yır[4]+san[2](sınız)[4]: you live →sən yaşayırsan;*

Правило 99. *При сочетании местоимения You и глагола к его азербайджанскому эквиваленту, заканчивающуюся на согласную букву прибавляется ir[4]+sən[2](sınız)[4]: you live →sən yaşayırsan;*

Правило 116. *Если к глаголу английского языка прибавляется аффикс –ed, то это правильный глагол V_r: worked, translated;*

Правило 117. *Если при поиске в словнике базы данных слово находится во втором столбике неправильных глаголов – это неправильный глагол V_{ir};*

Правило 118. *При сочетании местоимения I и правильного или неправильного глагола, то к азербайджанскому эквиваленту глагола прибавляется miş[4]+əm[2]: I worked (mən işləmişəm), I lived (mən yaşamışam);*

Правило 119. *При сочетании местоимения You и правильного или неправильного глагола к азербайджанскому эквиваленту глагола прибавляется mi[4]+sən[2]: you worked (sən işləmisən), you went (sən getmisən);*

Правило 120. *При сочетании местоимения He/She/İt и правильного или неправильного глагола к азербайджанскому эквиваленту глагола прибавляется miş^4+di^4: he worked (o işləmişdi), he went (o getmişdi);*

Правило 121. *При сочетании местоимения We и правильного или неправильного глагола к азербайджанскому эквиваленту глагола прибавляется miş^4+diq^4: we worked (biz işləmişdik), we went (biz getmişdik);*

Правило 122. *При сочетании местоимения They и правильного или неправильного глагола к азербайджанскому эквиваленту глагола прибавляется miş+di^4/dirlər: they worked (onlar işləmişdirlər), they went (onlar getmişdirlər);*

Правило 123. *При сочетании местоимения I+shall/will+глагол V к его азербайджанскому эквиваленту, заканчивающуюся на согласную букву, прибавляется аффикс acaq^2+am^2;*

Правило 124. *При сочетании местоимения I+shall/will+глагол V к его азербайджанскому эквиваленту, заканчивающуюся на гласную букву, прибавляется аффикс yacaq^2+am^2;*

Правило 125. *При сочетании местоимения You+will+глагол V к его азербайджанскому эквиваленту, заканчивающуюся на согласную букву, прибавляется аффикс acaq^2+san^2/sız^2;*

Правило 126. *При сочетании местоимения You+will+глагол V к его азербайджанскому эквиваленту, заканчивающуюся на гласную букву, прибавляется аффикс yacaq^2+san^2/sız^2;*

Правило 127. *При сочетании местоимения He/She/It+will+глагол V к его азербайджанскому эквиваленту, заканчивающуюся на согласную букву, прибавляется acaq^2;*

Правило 128. *При сочетании местоимения He/She/It+will+глагол V к его азербайджанскому эквиваленту, заканчивающуюся на гласную букву, прибавляется yacaq^2;*

Правило 129. *При сочетании местоимения We+shall/will+глагол V к его азербайджанскому эквиваленту, заканчивающуюся на согласную букву, прибавляется аффикс acağ^2+ık^4;*

Правило 130. *При сочетании местоимения We+shall/will+глагол V к его азербайджанскому эквиваленту, заканчивающуюся на гласную букву, прибавляется аффикс yacağ^2+ık^4;*

Правило 131. *При сочетании местоимения They+will+глагол V к его азербайджанскому эквиваленту, заканчивающуюся на согласную букву, прибавляется аффикс acaq²+lar²;*

Правило 132. *При сочетании местоимения They+will+глагол V к его азербайджанскому эквиваленту, заканчивающуюся на гласную букву, прибавляется аффикс yacaq²+lar²;*

Правило 133. *При сочетании местоимения I+shall+not/ will+not+глагол V к его азербайджанскому эквиваленту прибавляется ma² +yacağ²+am²;*

Правило 134. *При сочетании местоимения You+will+ not+глагол V к его азербайджанскому эквиваленту прибавляется ma² +yacaq²+san²;*

Правило 135. *При сочетании местоимения He/She/It+will+ not+глагол V к его азербайджанскому эквиваленту прибавляется ma²+yacaq²;*

Правило 136. *При сочетании местоимения We+shall+not/ will+not+глагол V к его азербайджанскому эквиваленту прибавляется ma²+yacağ²+ıq²;*

Правило 137. *При сочетании местоимения They+will+ not+глагол V к его азербайджанскому эквиваленту прибавляется ma²+ yacaq²+lar²;*

Правило 138. *При сочетании существительных в един. числе +will+глагол V данное правило верно правилу 127;*

Таким образом, необходимость каталогизировать все морфосинтаксические элементы языка привела к созданию базы знаний на основе грамматики двух языков. Сегодня становится актуальным построение системы более широкого круга морфологических и синтаксических знаний как инструмента для обобщения, применения и оценки разнообразных грамматических концепций.

Анализатор лингвистических знаний задает метаязык для метаязыков морфологических теорий, настраивается на некоторую авторскую морфологическую концепцию и на параметры языков-объектов. С помощью такой системы исследователь может – на основе своей оригинальной концепции – описывать материал, а в режиме интерпретации конкретных примеров проверять, адекват-

на ли его концепция языковому материалу, словоизменительным типам в данном языке, чередованиям и отношениям супплетивизма.

В отличие от баз данных, созданное представление знаний – не простое хранилище сведений: на основе обновляемых данных она является частью системы, анализирующей и синтезирующей словоформы языка-объекта.

ЛИТЕРАТУРА

1. Бреус Е.В. Основы теории и практики перевода с русского языка на английский. – М.: УРАО, 1998.

2. Ванников Ю.В. Языковая сложность текста как фактор трудности перевода (Методическое пособие). – М.: Всесоюзный центр переводов,1998.

3. Васильев А. (Компьютер на месте переводчика). // Подводная лодка – 1998, № 6.

4. Велиева К.А. Автоматическая расстановка и огласовка морфем азербайджанского слова (при машинном переводе). // Статис. и инф. изучение тюркских языков. Алма-Ата, 1969.

5. Велиева К.А. Алгоритмическое описание правил порождения азербайджанских словоформ.// Ученые записки, серия языка и литер. Б., 1968, 5.

6. Велиева К.А. Формальное описание расстановки морфем азербайджанского слова.// Учен. зап., АГУ, серия языка и литер. Б., 1968.

7. Велиева К.А. Формальное описание синтеза азербайджанского слова. АКД; М., 1971.

8. Велиева К.А., Мамедова М.Г., Махмудов М.А., Пинес В.Я. Проблемы автоматической переработки текстов на азербайджанском языке. Материалы семинара «Республиканская система научно-технической информации и опыт организации информационного обеспечения народного хозяйства» – Баку, АзНИИНТИ, 1985.

9. Велиева К.А., Мамедова М.Г., Махмудов М.А., Пинес В.Я., Рахманов Дж.А. Принципы построения систем лексико-грамматического машинного перевода с тюркских языков. – Материалы международной конференции «Теория и практика научно-технического перевода». Москва, 1985.

10. Винокуров А.А., Чуканов В.О. Новый метод оценки машинного перевода. // Информационные технологии и системы. Hardware Software Security. Тенденции и перспективы – Сборник статей: М., Международная академия информатизации, 1997.

11. ВЦП (Всесоюзный центр переводов), «Международная конференция по теории и практике научно-технического перевода» Москва. 1985. Тезисы докладов.

12. Гаджиева Н.З., Серебреников Б.А. Сравнительно-историческая грамматика тюркских языков. Синтаксис. М., «Наука»,1986. 384 стр.

13. Кулиева З.Ю. Проблемы разрешения неоднозначности в СМП (на примере английского и азербайджанского языков). Известия НАНА, сер. гуман. наук, языкознание, 2005, № 4.

14. Кулиева З.Ю. Автоматическое разрешение смысловой неоднозначности в СМП (на примере английского и азербайджанского языков) Журнал Искусственный Интеллект. Киев, 2005. стр. 578–589

15. Кулиева З.Ю. Определение эквивалентов категории Present Perfect для ввода базы знаний морфосинтаксического анализатора. Tədqiqlər, 2007. №4 стр. 61–70.

16. Кулиева З.Ю. Построение базы знаний морфосинтаксического анализатора. Тюркология, 2007. 65–72

17. Кулиева З.Ю. Применение формальных признаков языка для построения базы знаний морфосинтаксического анализатора. Донецк. 2007. №1 http://iai.dn.ua

18. Кулиева З.Ю. Создание базы знаний для обучающей системы перевода(TTC-Tədris Tərcümə Sistemi). Материалы Республиканской конференции Informatika, informasiya Texnologiyalarının Təhsildə Tətbiqi məsələləri. Б., 2007 146–149. Махмудов М.А. Лексико-морфологический МП азербайджанского текста на русский язык. Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика» – Чимкент, 1980.

19. Махмудов М.А. Разработка системы формального морфологического анализа тюркской словоформы (на материале азербайджанского языка). – АКД, Баку, 1982.

20. Махмудов М.А. Система автоматической переработки тюркского текста на лексико-морфологическом уровне. Б., 1991.

УДК 81’33

DYNAMIC LEXICON: COMPUTER PRESENTATION OF MODELS OF WORD-FORMATION

A. Chemyshev
V.M. Vasilyev Mari Research Institute of Language, Literature and History, Yoshkar-ol, Mari El, Russia
chemyshev.andrey@gmail.com

In a real language, neologisms, occasionalisms, authorial expressions are formed daily, which can not be adequately analyzed by existing automatic parsing tools, often built on the principle of synchronism and spelling of word forms. It is proposed to build and test a universal model of the computer representation of the morphemic composition of the difference-structured languages and create software tools for the dynamic analysis of verbal facts, taking into account the association of related words. The computer linguistic model should be able to describe the free combinatorics of morphemes taking into account their linear positioning, activated phonemic interlaces and semantics.

Keywords: deep morphemes, postulated morphemes, base morphemes, word-forming nests, allomorphs, dynamic analysis, linguistic ontology.

ДИНАМИЧЕСКИЙ ЛЕКСИКОН: КОМПЬЮТЕРНОЕ ПРЕДСТАВЛЕНИЕ МОДЕЛЕЙ СЛОВООБРАЗОВАНИЯ

А.В. Чемышев
Марийский научно-исследовательский институт языка, литературы и истории им. В.М. Васильева, г. Йошкар-Ола
chemyshev.andrey@gmail.com

В реальном языке ежедневно формируются неологизмы, окказионализмы, авторские выражения, которые невозможно адекватно анализировать существующими средствами автоматического разбора, зачастую построенные на принципе синхронности и орфографического членения словоформ. Предлагается отработать универсальную модель компьютерного представления морфемного состава разноструктурных языков и создать программные средства динамического анализа речевых фактов, учитывающее гнездовое объединение родственных слов. Компьютерная лингвистическая модель должна быть способна описать свободную комбинаторику морфем с учётом их линейного позиционирования, активируемых фонемных чередований и семантики.

Ключевые слова: глубинные морфемы, постулируемые морфемы, опорные морфемы, словообразовательные гнёзда, алломорфы, динамический анализ, лингвистическая онтология.

Введение

Современные средства автоматического морфологического анализа практически игнорируют внутреннее морфемное членение анализируемых языковых форм. Большинство программных решений включают так называемые «основы» в постоянный словарь своих систем и используют их в качестве готовых и неделимых лексических единиц.

Однако реальный язык динамичен – в письменной и устной речи ежедневно порождается множество фактов, которые нельзя объяснить заранее заготовленным списком словоформ или даже основ. Появляются новые живые комбинации морфем, формируются неологизмы, окказионализмы, авторские выражения. Для адекватного анализа этих языковых фактов средствами автоматического разбора необходимо использовать компьютерные лингвистические модели, способные описать свободную комбинаторику морфем с учётом их линейного позиционирования, активируемых фонемных чередований и, разумеется, семантики.

Созданный «Морфемный словарь марийского языка» позволить создать основу для отработки универсальной модели компьютерного представления морфемного состава марийского языка и создать программные средства динамического анализа речевых фактов. Ключевым требованием является необходимость сведения репертуара морфемных единиц к минимально достаточному набору, обладающему большой объяснительной силой. Абстрактно постулируемая морфема («глубинная морфема») должна порождать все необходимые алломорфы на поверхностном орфографическом уровне. Это позволит активно привлекать в процесс моделирования исторические сведения о языке.

На первом этапе в «Морфемный словарь» будут включаться морфемы из «Марийского орфографического словаря» 2011 года (40 тыс. слов).

На втором этапе в словарь будут включаться слова из других словарей и текстов, в том числе ненормированные слова с соответствующими пометами, как разг. – разговорное слово; диал. – диалектное слово (с указанием диалекта); ст.орфогр. – старая орфография (орфография 1994 года, 1972 года и т.д.) и др. Также будут созданы словообразовательные гнёзда.

1. Разметка слов

Все слова из «Марийского орфографического словаря» 2011 года (40 тыс. слов):

а-а	ава́	авалткала́ш, -ем
абажу́р	ававаре́нге	авалтма́ш
абажура́н	ававеэ́л	ава́лтме
аббревиату́р	ававуэ́н	ава́лтше
абза́ц	авагалта́	авалтыдыма́ш
абисси́н (калык)	авагара́ш	авалтыма́ш
абитуриэ́нт	авагашта́	ава́лтыме
абонеме́нт	авагу́до	ава́лтыш
абоне́нт	авагу́мдыр	ава́лтыше
абориге́н	ава́дыме	ава́лык
абрико́с	аважэ́	авалыма́ш
абсолюти́зм	ава́й	ава́лыме
абстра́ктнын	авала́ш, -ем	ава́лыше
абстра́кций	авали́йше	авамла́нде
абсу́рд	авалта́ш, -ам	авангА́рд
абха́з (калык)	авалта́ш, -ем	...

разметим следующим образом:

а-а	Interj	{а-а}
абажур	N	{абажу!р}
абажуран	A	{абажура!н}
аббревиатур	N	{аббревиату!р}
абзац	N	{абза!ц}
абиссин	N	{абисси!н}
абитуриент	N	{абитурие!нт}
абонемент	N	{абонеме!нт}
абонент	N	{абоне!нт}
абориген	N	{абориге!н}
абрикос	N	{абрико!с}
абсолютизм	N	{абсолюти!зм}
абстрактнын	Adv	{абстра!ктнын}
абстракций	N	{абстра!кций}
абсурд	N	{абсу!рд}
абхаз	N	{абха!з}

ава	N	{ава!}
ававаренге	N	{ававаре!нге}
ававел	N	{ававе!л}
ававун	N	{ававу!н}
авагалта	N	{авагалта!}
авагараш	N	{авагара!ш}
авагашта	N	{авагашта!}
авагудо	N	{авагу!до}
авагумдыр	N	{авагу!мдыр}
авадыме	A	{ава!дыме}
аваже	N	{аваже!}
авай	N	{ава!й}
авалаш	V2	{авала!ш} (-ем)
авалийше	N	{авали!йше}
авалташ	V1	{авалта!ш} (-ам)
авалташ	V2	{авалта!ш} (-ем)
авалткалаш	V2	{авалткала!ш} (-ем)
авалтмаш	N	{авалтма!ш}
авалтме	Prc	{ава!лтме}
авалтше	Prc	{ава!лтше}
авалтыдымаш	N	{авалтыдыма!ш}
авалтымаш	N	{авалтыма!ш}
авалтыме	Prc	{ава!лтыме}
авалтыш	N	{ава!лтыш}
авалтыше	Prc	{ава!лтыше}
авалык	N	{ава!лык}
авалымаш	N	{авалыма!ш}
авалыме	Prc	{ава!лыме}
авалыше	Prc	{ава!лыше}
авамланде	N	{авамла!нде}
авангард	N	{аванга!рд}

где: в 1-ой колонке – слова без помет и ударений;

во 2-ой колонке – грамматические теги, обозначающие части речи: **N** – имя существительное, **A** – имя прилагательное, **Num** – имя числительное, **Pron** – местоимение, **V1** – глагол 1-го спряжения (-ам), **V2** – глагол 2-го спряжения (-ем), **Prc** – причастие, **Ger** – деепричастие, **Adv** – наречие, **Po** – послелог, **Conj** – союз,

Pcle – частица, **Interj** – междометие, **Imitat** – подражательное слово;

в 3-ей колонке – слова, заключённые между знаками «{» и «}» с ударениями, которые обозначены как «!»

Окончательный вариант разметки всех 40 тыс. слов «Марийского орфографического словаря» 2011 года должен выглядеть следующим образом:

Слово	Часть речи	Разметка слова	Опорная морфема	Словообразовательный аффикс	Неодушевл./ Одушевл./ Состояние/ Процесс
а-а	Interj	{а-а}	а	-	Signal
абажур	N	{аба|жу!р}	абажур	-	неодушевл.
абажуран	A	{аба|жу|ра!н}	абажур	ан	ATTR
аббревиатур	N	{аб|бре|ви|ат|у!р}	аббревиатур	-	неодушевл.
абзац	N	{аб|за!ц}	абзац	-	неодушевл.
абиссин	N	{аб|ис|си!н}	абиссин	-	одушевл.
абитуриент	N	{аби|ту|ри|е!нт}	абитуриент	-	одушевл.
абонемент	N	{або|не|ме!нт}	абонемент	-	неодушевл.
абонент	N	{або|не!нт}	абонент	-	неодушевл.
абориген	N	{або|ри|ге!н}	абориген	-	одушевл.
абрикос	N	{аб|ри|ко!с}	абрикос	-	неодушевл.
абсолютизм	N	{аб|со|лю|ти!зм}	абсолютизм	-	неодушевл.
абстрактнын	Adv	{аб|стра!кт|нын}	абстрактн	ын	состояние
абстракций	N	{аб|стра!к|ций}	абстракций	-	неодушевл.
абсурд	N	{аб|су!рд}	абсурд	-	неодушевл.
абхаз	N	{аб|ха!з}	абхаз	-	одушевл.
ава	N	{ава!}	ава	-	одушевл.
ававаренгге	N	{ава|ва|ре!н|ге}	ава+паренг	-	неодушевл.
ававел	N	{ава|ве!л}	ава+вел	-	одушевл.
ававун	N	{ава|ву!н}	ава+пун	-	неодушевл.
авагалта	N	{ава|гал|та!}	ава+калта	-	неодушевл.
авагараш	N	{ава|га|ра!ш}	ава+караш	-	неодушевл.
авагашта	N	{ава|гаш|та!}	ава+кашта	-	неодушевл.
авагудо	N	{ава|гу!|до}	ава+куд	-	неодушевл.
авагумдыр	N	{ава|гу!м|дыр}	ава+гум+тар	-	неодушевл.
авадыме	A	{ава!|ды|ме}	ава	дыме	ATTR
аваже	N	{ава|же!}	ава	же	одушевл.
авай	N	{ав|а!й}	ава	й	одушевл.
авалаш	V2	{ава|ла!ш} (-ем)	ава	л	процесс
авалийше	N	{ава|ли!й|ше}	ава+лий	;ше	одушевл.
авалташ	V1	{ав|ал|та!ш} (-ам)	ава	л+т	процесс

авалташ	V2	{ав|ал|та!ш} (-ем)	ава	л+т	процесс
авалткалаш	V2	{ав|алт|ка|ла!ш} (-ем)	ава	л+т+кал	процесс
авалтмаш	N	{ав|алт|ма!ш}	ава	л+т+маш	неодушевл.
авалтме	Prc	{ав|а!лт|ме}	ава	л+т+ме	ATTR
авалтше	Prc	{ав|а!лт|ше}	ава	л+т+ше	ATTR
авалтыдымаш	N	{ав|ал|ты|ды|ма!ш}	ава	л+ты+дым+аш	неодушевл.
авалтымаш	N	{ав|ал|ты|ма!ш}	ава	л+ты+маш	неодушевл.
авалтыме	Prc	{ав|а!л|ты|ме}	ава	л+ты+ме	ATTR
авалтыш	N	{ав|а!л|тыш}	ава	л+ты+ш	неодушевл.
авалтыше	Prc	{ав|а!л|ты|ше}	ава	л+ты+ше	ATTR
авалык	N	{ава!|лык}	ава	лык	неодушевл.
авалымаш	N	{ава|лы|ма!ш}	ава	лы+маш	неодушевл.
авалыме	Prc	{ава!|лы|ме}	ава	лы+ме	ATTR
авалыше	Prc	{ава!|лы|ше}	ава	лы+ше	ATTR
авамланде	N	{ава|мла!н|де}	ава+мланд	-	неодушевл.
авангард	N	{ав|ан|га!рд}	авангард	-	неодушевл.

где: **ATTR** – атрибут, **Signal** – сигнал;

«|» – деление на слоги (используется в мягких переносах и синтезаторах речи);

«**!**» – ударение.

Примечание: в слове «авагумдыр» (синоним «ававундыр») морфема «гум», заимствованное из других языков (<тат. *кунь, куен*), – «пазуха»; в морфеме «тар» (плата), «т» → «д» (грамматические правила) и «а» → «ы» (исторический аспект).

В данном орфографическом словаре рядом с глаголом расположены действительные, страдательные, реже отрицательные причастия этого глагола; существительные, образованные от этого глагола с помощью суффикса «маш»; также видовые классы, уменьшительно-ласкательные формы, формы иронично-насмешливого высказывания и другие глаголы, образованные от этих глаголов с помощью словообразовательных суффиксов. Количество слов в данных группах от 4 до 20:

авала́ш, -ем	кандалга́ш, -ем	толкыналта́ш, -ам
...	…	толкыналтма́ш
авалта́ш, -ам	кандалгыма́ш	толкына́лтме
авалта́ш, -ем	канда́лгыме	толкына́лтше
авалткала́ш, -ем	…	...
авалтма́ш	канда́лгыше	толкынаңа́ш, -ам

ава́лтме	…	толкынанґда́ш, -ем
ава́лтше	кандаґа́ш, -ам	толкынанґдыма́ш
авалтыдыма́ш	кандаґда́ш, -ем	толкына́ндыме
авалтыма́ш	кандаґдыма́ш	толкына́ндыше
ава́лтыме	канда́ґдыме	толкынаґма́ш
ава́лтыш	канда́ґдыше	толкына́ґме
ава́лтыше	кандаґма́ш	толкына́ґше
...	канда́ґме	толкындара́ш, -ем
авалыма́ш	канда́ґше	толкындарыма́ш
ава́лыме		толкында́рыме
ава́лыше		толкында́рыше
		толкынлана́ш, -ем
		толкынланыма́ш
		толкынла́ныме
		толкынла́ныше

Для облегчения работы по разметке слов данного орфографического словаря выделим в отдельный список глаголы. Получилось 6462 глагола. Из этого количества глаголов образуются свыше 20 тыс. причастий, существительных, которые включены в орфографический словарь. Общее количество слов с учётом с глаголов – около 30 тыс.

В итоге необходимо разметить около 16 тыс. слов, остальные слова будут сгенерированы из глаголов.

Далее выделим в отдельный список опорные морфемы и словообразовательные аффиксы. По нашим предположениям, количество марийских опорных морфем – около 3000. Из этих морфем образовываются все марийские слова.

2. Описание морфем

Описание морфем должно включать в себя следующие части: морфонологический состав атомарных единиц + описания синтактики (= правила связывания с другими единицами) + ссылки на концептуальную онтологию. Например:

<code n=»БАНК»>
<seq s=»Б+А|Н.|К#»/>
<denots>
<denot ref=»Obj Core -БАНКЕ-«/>

<denot ref=»Obj Core -БАНК-»/>
</denots>
</code>

<code n=»ПОНГ»>
<seq s=»П+О|Н.|Г#»/>
<denots>
<denot ref=»Obj Core -ПОНГ-»/>
</denots>
</code>

Это описание морфонологической последовательности слогов, образующих фонокод (атомарную языковую морфему). Поле «seq» (sequence, т.е. последовательность, цепочка) содержит запись слоговой последовательности, в которой слоги отделены вертикальной чертой. В каждом слоге могут иметь место следующие условные обозначения. Знак «+» разделяет согласный и гласный элементы слога. Если в слоге нет начального согласного, то ставим вместо него знак «?», например: <seq s=”?A|B+A”/> для описания морфемы «ава».

Если перед нами скопление согласных, то каждый отсутствующий гласный заменяем на знак «.» и всё равно выделяем отдельный слог (=псевдослог). Если в последнем слоге морфемы присутствует лишь согласный, ставим после него знак «#», что означает возможность заполнения этой позиции гласным от последующей морфемы.

Поле «denots» (=denotations, обозначаемое) содержит перечень возможных синтактико-семантических единиц, которые используют данный фонокод для своего выражения. Так разграничиваются омонимичные морфемы. В каждом «denot» содержится ссылка «ref» на единицу более высокого уровня (переходим от линейной фонологии к синтактике).

Заключение

«Морфемный словарь марийского языка» позволит решить задачу по подготовке основы для создания средства автоматического анализа марийских текстов, включающих морфологический, словообразовательный, фонетический и др. анализы словоформ, в том числе с учётом исторического аспекта.

УДК 81’33

THE UNIFORM MORPHOLOGICAL ANALYZER FOR THE KAZAKH AND TURKISH LANGUAGES

[1]A. Sharipbay, [1]G. Bekmanova, [2]G. Altnbek, [3]E. Adalı, [1]L. Zhetkenbay ,[1] U. Kamanur

[1]L.N.Gumilyov Eurasian National University, Astana Kazakhstan
[2]College of Information Science and Engineering, Xinjiang University, P.R. China
[3]Istanbul Technical University, Istanbul, Turkey
{gulmira-r@yandex.ru, sharalt@mail.ru,
jetlen_7@mail.ru unzila.88@mail.ru }[1]
glaxd2014@163.com[2] esrefadali@gmail.com [3]

The Kazakh and Turkish languages belong to the group of the Turkic languages and have much in common. The detailed comparison of the ontologies on the example of the Kazakh and Turkish nouns allowed entering the analysis of morphological rules of these languages and the unified system of designations to create the uniform morphological analyzer based on the general algorithm of the morphological analysis.

Keywords: morphological analysis of the Kazakh and Turkish languages, ontology, analysis of morphological rules.

ЕДИНЫЙ МОРФОЛОГИЧЕСКИЙ АНАЛИЗАТОР ДЛЯ КАЗАХСКОГО И ТУРЕЦКОГО ЯЗЫКОВ

[1]А.А. Шарипбаев, [1]Г.Т. Бекманова, [2]Г. Алтынбек, [3]Е. Адалы, [1]Л. Жеткенбай, [1]У. Каманур

[1]Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан
[2]Колледж информатики и инженерии, Синьзянский куниверситет, Урумчи, Китай
[3]Стамбульский Технический университет, Стамбул, Турция
{gulmira-r@yandex.ru, sharalt@mail.ru,
jetlen_7@mail.ru unzila.88@mail.ru }[1]
glaxd2014@163.com[2] esrefadali@gmail.com [3]

Казахский и турецкий языки относятся к группе тюркских языков и имеют много общего. Детальное сравнение онтологий на примере казах-

ских и турецких существительных позволило приступить к анализу морфологических правил этих языков и созданию единой системы обозначений для единого морфологического анализатора на основе общего алгоритма морфологического анализа.

Ключевые слова: морфологический анализ казахского и турецкого языков, онтология, анализ морфологических правил.

1. Introduction

One of the methods to reduce the semantic barrier between the human and the computer is searching new methods of a natural language processing. Nowadays it is obvious that in order to implement the human-computer interaction in a natural language and to create a linguistic support of the information processes the study of the language itself is required. Besides the resources consumed could be decreased due to formalization of language rules providing the storage of information in procedural but not in declarative form. For the Kazakh and the Turkish languages which morphological regularities are quite well yielded to formalization, it would produce an excellent result.

All language levels are characterized by existence of basic elements. A language studying can take place from two positions – the analysis and synthesis because the revealed rules of synthesis can assist to carry out the analysis and vice versa. In this case the Kazakh and Turkish languages are studied from both positions the analysis and synthesis. This very integrated approach allows to study in details all regularities and to reveal such nuances which when using only of one of approaches would remain outside our attention. For researching and the maximum formalizing of each language subsystem it is necessary to create the program tools implementing the studying process by identifying and verifing the analysis and synthesis rules. Therewith it will greatly automate the research process and a researcher doesn't need to accumulate and collect information. And the labor intensity is very low.

The morphology modeling is related to all applications such as natural language and tasks processing and includes information search, moods analysis, spelling correction, detection of the generated texts, parts of speech marking and entity extraction. The morphology is used in linguistics to refer to the study of structure and formation of words. The Agglutinative languages (agglutinare from Latin means "to stick together") are languages which morphological system is characterized

by agglutination (“pasting”) of various formant. As a formant either prefixes or suffixes act and each of them makes its own sense.

As the Kazakh and Turkish languages belong to the group of Turkic languages and the languages of this group can be classified as agglutinative languages. These languages are full of word forms (inflections). Inflections are formed by addition of suffixes. The suffixes are attached in the strict sequence and the resulting new words can belong to the other part of speech. The possessive form in Kazakh is similar to a possessive form in English [1, 2]. Plenty of researches covering formalization of morphological rules and morphological analyses of [3-6] the Turkic languages are available. The first morphological analyzer of Kazakh was developed in 2009 and based on the procedural method. The procedural method implies the preliminary systematization of morphological knowledge about a natural language and development of morphological information assignment algorithms to a separate word form [7, 8].The procedural morphological analyzer of Kazakh consisted of the following stages: marking the stem in the current word form, its identification, assigning to a word form the corresponding list of morphological information. The disadvantage of this method is high labor intensity while compiling the dictionaries of compatibility. This challenge is difficult to be settled and cannot be automated completely for languages which are characterized by a large number of counterexamples. The implementation of this method occupies considerably smaller memory size, but at the same time the morphological analysis period due to splitting a word form into components and applying the procedures of compatibility increases [8]. The second version of the morphological analyzer was developed in 2012 and based on the formal morphological rules [9]. Later versions were based on using the ontological models and the hyper graphs [10-13]. The other research groups developed their own morphological analyzers [15-17].

The works on creation of the morphological analysis for the Turkish language are carried out for a long period of time and presented in papers [18-26]. In this paper the results received in [18] were used. The peculiarity of this morphological analyzer is the methodology for carrying out the analysis. The Turkish words with affixation were used without any lexicon. This morphological analyzer is completely based on the rules and implies using only the dictionary of counterexamples. The analyzer is based on the final automatic model.

2 The generalized ontologic models of parts of speech of the Kazakh and Turkish languages

Ontology is a powerful and widely used tool for modeling relationships between objects which belong to the different subject area. Ontology should be classified based on the degree of dependence on the task or application area, ontological model for knowledge representation and expression as well as other criteria [26, 27].

We used the ontology editor Protégé (http://protege.stanford.edu) to build the ontology. It is a free open source ontology editor and a framework for building knowledge bases. It was developed at Stanford University in collaboration with the University of Manchester.

The morphological features of initial forms of nouns (N) are as follows. A noun can be either animate (anim) or inanimate (inanim); this feature determines the trajectory of the inflection of a noun. Nouns in the Kazakh language can be conjugated (pers_end) and vary for case (cases) and number (number), as well as have a possessive form (poss_end) [10].

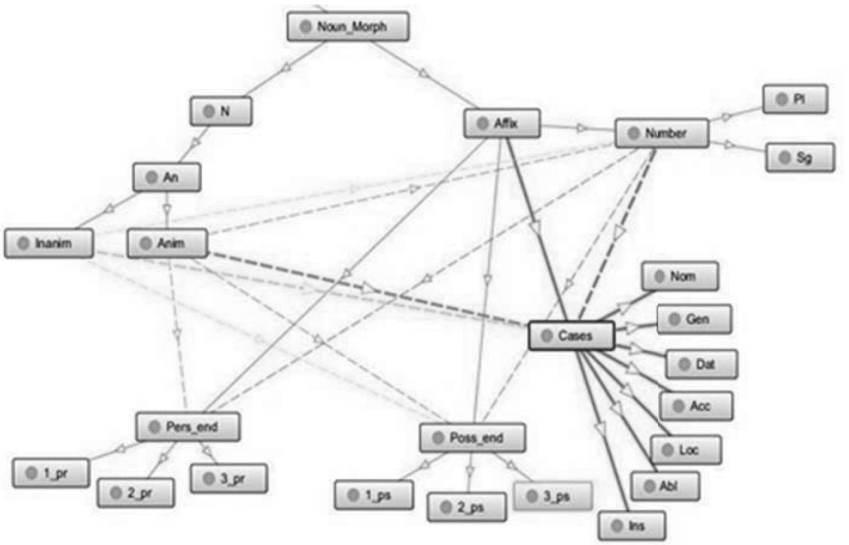


Fig. 1. Ontological model of the Kazakh noun [10]

Figure 1 shows the ontological model of the Kazakh noun with its morphological features. Concepts and relationships used in this ontological model are explained in Table 1.

Table 1. – **Concepts and relationships**

Notation	Description	Notation	Description
N	Noun	2 pr	2 personal/
Part of speech		3 pr	3 personal /
Item	Item	Poss_end	Possessive endings
Anim	Animate	1 ps	1 personal /
Sign of animacy	v16	2 personal	
Inanim	Inanimate	3 ps	3 personal
Sign of inanimacy		Number	Number
Cases	Cases	Pl	Plural
Nom	Nominative case	Sg	Singular
Gen	Genitive case	is_a	
Dat	Direction– dative case	Denotes	
Acc	Accusative		
e3, e4	has_feature		
Loc	Locative case	Has .	
Abl	Ablative case	Devided	
Ins	Instrumental case	Change	
Pers_end	Personal endings	Add	
1 pr	1 personal		

The ontology model of the Kazakh parts of speech allows us to completely describe the morphological rules and their relationships. On the basis of this ontological model we developed generalized ontological models of the Kazakh and Turkish language parts of speech.

The developed ontological model of nouns of the Kazakh language in Protege environment is displayed in the Figure 2, and the ontological model of nouns of the Turkish language is shown below in the Figure 3.

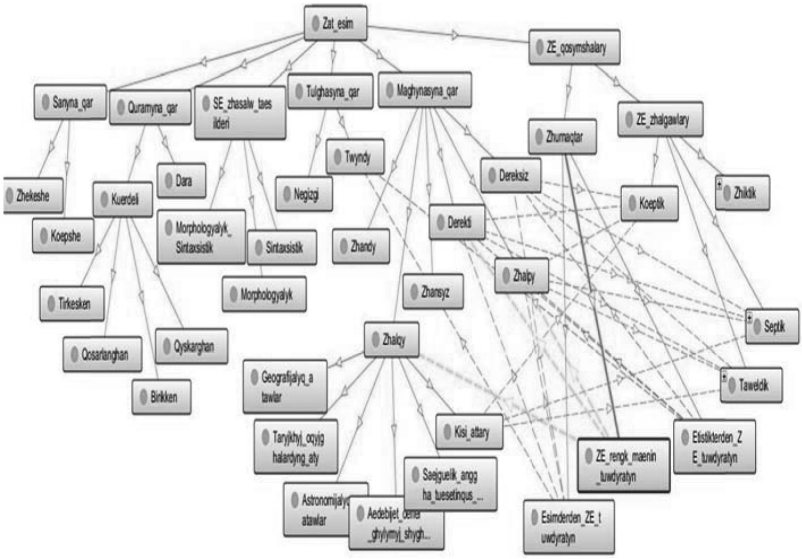


Fig. 2. The ontological models of nouns for the Kazakh language

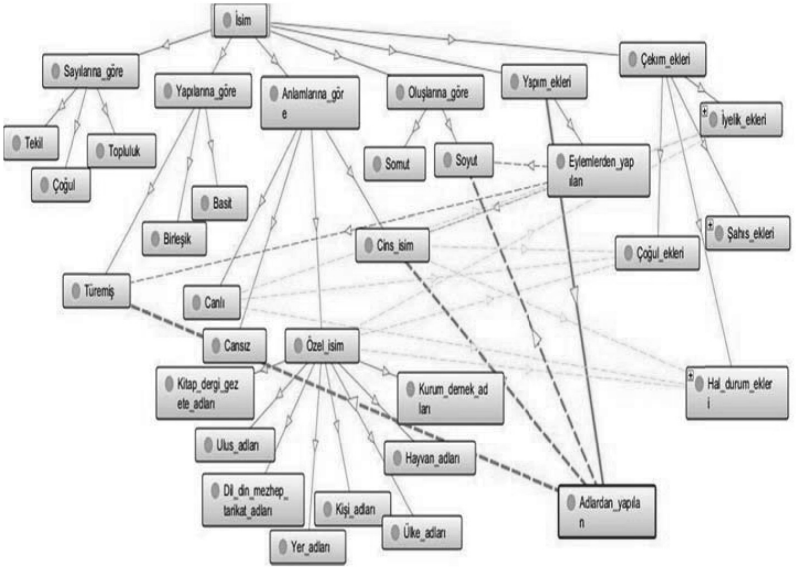


Fig. 3. The ontological models of nouns for the Turkish language

In this way the comparative ontological models of noun for machine translation system include all the categories of morphological features, for instance, noun is divided as stem and complex according to the structure of noun in the Kazakh language whereas in the Turkish language there is not such division, furthermore, a noun can be common, proper, concrete, abstract, animated, inanimate according to meaning in the Kazakh language, while in the Turkish language a noun can be common, proper, animated, inanimated. In both languages the divisions of affixation are similar, e.g., the forms of cases, number, possessives and conjugations. There are seven cases in Kazakh whereas in Turkish there are five.

3 The uniform morphological analyzer for the Kazakh and Turkish languages

The comparison of the ontological models allowed creating the general symbol system of morphological markers which are used in morphological analyzer.

Table 2. **The comparison of morphological markers of the Kazakh and Turkish languages nouns**

№	Abbreviation	Name in English	Name in Kazakh	Name in Turkish	Unified Tag
1.	+Noun	Noun	Zat esim	İsim	Noun
2.	+A1sg	Personal 1 singular	Zhiktik 1 zhaq, zhekeshe	1. Tekil Şahıs Uyum Özelliği	PERS.1SG
3.	+A2sg	Personal 2 singular	Zhiktik 2 zhaq, zhekeshe	2. Tekil Şahıs Uyum Özelliği	PERS.2SG
4.					PERS.2SG.POL
5.	+A3sg	Personal 3 singular	Zhiktik 3 zhaq, zhekeshe	3. Tekil Şahıs Uyum Özelliği	PERS.3SG
6.	+A1pl	Personal 1 plural	Zhiktik 1 zhaq, koepshe	1. Çoğul Şahıs Uyum Özelliği	PERS.1PL
7.	+A2pl	Personal 2 plural	Zhiktik 2 zhaq, koepshe	2. Çoğul Şahıs Uyum Özelliği	PERS.2PL

8.					PERS.2PL.POL
9.	+A3pl	Personal 3 plural	Zhiktik 3 zhaq, koepshe	3. Çoğul Şahıs Uyum Özelliği	PERS.3PL
10.	+P1sg	Possessive 1 singular	Zhiktik 1 zhaq, Taweldik 1 zhaq, zhekeshe	1. Tekil Şahıs İyelik Eki	POSS.1SG
11.	+P2sg	Possessive 2 singular	Taweldik 2 zhaq, zhekeshe	2. Tekil Şahıs İyelik Eki	POSS.2SG
12.	+P2sgpol	Possessive 2 singular (formal)	Taweldik 2 zhaq, zhekeshe, resmi tueri		POSS.2SG.POL
13.	+P3sg	Possessive 3 singular	Taweldik 3 zhaq, zhekeshe	3. Tekil Şahıs İyelik Eki	POSS.3SG
14.	+P1pl	Possessive 1 plural	Taweldik 1 zhaq, koepshe	1. Çoğul Şahıs İyelik Eki	POSS.1PL
15.	+P2pl	Possessive 2 plural	Taweldik 2 zhaq, koepshe	2. Çoğul Şahıs İyelik Eki	POSS.2PL
16.	+P2plpol	Possessive 2 plural (formal)	Taweldik 2 zhaq, koepshe, resmi tueri		POSS.2PL.POL
17.	+P3pl	Possessive 3 plural	Taweldik 3 zhaq, koepshe	3. Çoğul Şahıs İyelik Eki	POSS.3PL
18.	+Pnon	Non Possessive	Taweldenbegen	Belirsiz İyelik	NON. POSS
19.	+Nom	Nominative	Atau	Yalın Durum	NOM
20.	+Acc	Accusative (whom?)	Tabys	Belirtme Durumu	ACC
21.	+Dat	Dative	Barys	Yönelme Durumu	DAT
22.	+Abl	Ablative	Shyghys	Çıkma Durumu	ABL
23.	+Loc	Locative (where?)	Zhatys	Kalma Durumu	LOC

24.	+Gen	Genitive (whose?)	Ilik	Tamlayan Durumu	GEN
25.	+Ins	Instrumental	Koemektes	Aracılık Durumu	INS
26.	+Pos	+Positive	Bolymdy	Olumlu	POSIT
27.	+Neg	+Negative	Bolymsyz	Olumsuz	NEGAT

For example, the line 4 of the above-mentioned table does not have any meanings in the Kazakh and Turkish columns, there is no analogue in English, but preserved name means that for the other Turkic languages such morphological marker for noun exists. In the lines 12 and 16 the blank value in the Turkish language means that this morphological marker exists only for the Kazakh language.

Metalanguage is one of key concepts of system of the description of an object of science and is defined as artificial language of «the second order» in relation to which natural human language acts as «language object», that is as a subject of a linguistic research. In our case a natural language are the Kazakh and Turkish languages entering into the Turkic group of languages.

The unified symbol system (UNIFIED TAG) was developed based on the idea of creating unified metalanguage for Turkic Languages.

Firstly, the idea of creating metalanguage was proposed at the 1st so-called International Conference on Computer processing of the Turkic Languages (TurkLang-2013) which was held in Astana on 3-4 October, 2013. A group of famous professors of technical sciences A.A. Sharipbay (Astana, Kazakhstan), D.SH. Suleimenov (Kazan, Tatarstan, Russia), Eşref Adalı (Istanbul, Turkey) is working on the creation of metalanguage.

At the UniTurk scientific-practical seminar of the conference the discussion of problems related to the unification of grammatical categories for the corpuses of the Turkic languages raised a great interest and held successfully.

For computerizing the Kazakh language it is very important step to research the computational linguistics of the other Turkic-speaking countries. From this point studying the structures of agglutinative languages that are similar to Kazakh and make comparisons between them leads to a successful computer transforming of all languages belonging

to the Turkic languages group. We are very confident that it will bring great success in development of the Kazakh language computerizing.

Our goal is to use correctly these similarities and differences in the language automating direction. While entering to a computer the similarities between languages help to solve the unsolved problems in one language by supplementing the achievements of another language, moreover, studying the differences of languages according to its features in cooperation gives us an opportunity to implement a method in one language which didn't give any results in another language.

The analysis made revealed that the Kazakh and Turkish languages have much in common. The Table 3 shows the comparison of the rules for a noun window in the Kazakh and Turkish languages.

Table 3. Example of inflection a noun window

English	Kazakh	Turkish
Case endings of Noun (singular form)		
Window	tereze: tereze+Noun+A3sg+Pnon+ Nom	pencere: pencere+Noun+A3sg+ Pnon+Nom
window 's	terezening: tereze+Noun+A3sg+Pnon+ Gen	pencerenin: pencere+Noun+A3sg+Pnon+Gen
to window	terezege: tereze+Noun+A3sg+Pnon+Dat	pencereye: pencere+Noun+A3sg+Pnon+Dat
Window	terezeni: tereze+Noun+A3sg+Pnon+ Acc	pencereyi: pencere+Noun+A3sg+Pnon+ Acc
Window	terezede: tereze+Noun+A3sg+Pnon+ Loc	pencerede: pencere+Noun+A3sg+Pnon+ Loc
from window	terezeden: tereze+Noun+A3sg+Pnon+ Abl	pencereden: pencere+Noun+A3sg+Pnon+ Abl
with window	terezemen: tereze+Noun+A3sg+Pnon+Ins terezemen: tereze+Noun+A3sg+ P1sg +Ins	pencerele: pencere+Noun+A3sg+Pnon+ Ins pencerele: pencere+Noun+A3sg+P1sg+ Ins

Case endings of Noun (plural form)		
Windows	terezeler: tereze+Noun+A3pl+Pnon+ Nom	pencereler: pencere+Noun+A3pl+Pnon+ Nom
windows'	terezelerding: tereze+Noun+ A3pl +Pnon+Gen	pencerelerin: pencere+Noun+ A3pl +Pnon+Gen
to windows	terezelerge: tereze+Noun+ A3pl +Pnon+Dat	pencerelere: pencere+Noun+ A3pl +Pnon+Dat
windows	terezelerdi: tereze+Noun+ A3pl +Pnon+ Acc	pencereleri: pencere+Noun+ A3pl +Pnon+ Acc
windows	terezelerde: tereze+Noun+A3pl +Pnon+ Loc	pencerelerde: pencere+Noun+ A3pl+Pnon+ Loc
from windows	terezelerden: tereze+Noun+A3pl +Pnon+ Abl	pencerelerden: pencere+Noun+ A3pl+Pnon+ Abl
with windows	terezelermen: tereze+Noun+A3pl +Pnon+Ins	pencerelerle: pencere+Noun+ A3pl+Pnon+ Ins

The record of morphological rules in the unified form allowed to create the uniform rule-based algorithm of morphological analysis for the Kazakh and Turkish languages in the papers [9, 10, 18].

4 Conclusion

In the present scientific paper the morphological features of the Kazakh and Turkish languages are analyzed. The ontologies comparison is made, the uniform symbol system of morphological features is developed and the morphological rules of the Kazakh and Turkish languages are written over via new symbol system. The unified morphological analyzer is developed based on the general morphological analysis algorithm.

In the future it is supposed to create the unified metalanguage of the Turkic languages that will allow reaching the new level the Turkic languages processing.

REFERENCES

1. Batayeva, Z. (2012). Colloquial Kazakh, Routledge.

2. Kazakh grammar. (2002). Phonetics, word formation, morphology, syntax (in Kazakh). Astana.

3. Sharipbay A., Bekmanova G. The synthesis of word forms of Turkic language using semantic neural networks. Modern problems of applied mathematics and information technologies: abstracts – Al Khorezmy, 2009. – P.145.

4. Sharipbayev A. A., Bekmanova G. T. The building of logical semantics of the Kazakh words // The materials of the all-Russian conference with International participations «Knowledge-Ontology-Theory (ZONT-09)». – Novosibirsk, 2009 – C. 246–249.

5. Tantuğ, A. C., Adalı, E., Oflazer, K.: Computer Analysis of the Turkmen Language Morphology, Lecture Notes in Computer Science, vol. 4139, pp. 186–193. Springer, (2006).

6. Orhun, M., Tantuğ, A. C., Adalı, E.: Rule Based Analysis of the Uyghur Nouns. Proceedings of the International Conference on Asian Language Processing (IALP). Chiang Mai, Thailand (2008).

7. Bekmanova G. T. Some approaches to the problems of automatic inflection and morphological analysis in the Kazakh language // the newsletter of D. Serikbayev East Kazakhstan state technical university – Ust-Kamenogorsk, 2009. – C. 192–197.

8. Dobrushina E. P., Savina G. B., Gelbukh A. G., The system of an accurate morphological analysis and synthesis. The software of new information technology. – Kalinin, 1989. – 222 c.

9. Sharipbayev A., Bekmanova G., Mukanova A., Buribayeva A., Yergesh B., Kaliyev A. Semantic neural network model of morphological rules of the agglutinative languages. The 6th International Conference on Soft Computing and Intelligent Systems The 13th International Symposium on Advanced Intelligent Systems. – Kobe, Japan, 20–24 November 2012, P. 1094–1099.

10. Yergesh B.,Mukanova, A., Bekmanova G, Sharipbay, A., Razakhova, B. Semantic hyper-graph based representation of nouns in the Kazakh language. Computacion y Sistemas; Volume 18, Issue 3, 1 July 2014. – P. 627–635.

11. Mukanova, A., Yergesh, Bekmanova G, B. Razakhova, B., Sharipbay, A. Formal models of nouns in the Kazakh language. Leonardo Electronic Journal of Practices and Technologies; Issue 25 (July-December), 2014 (13). – P. 264–273.

12. A.Sharipbay, L. Zetkenbay, U.Kamanur. Ontological modeling of morphological rules for the adjectives in Kazakh and Turkish languages. Journal of Theoretical and Applied Information Technology, Vol. 91. No.2, 2016, P. 257–263. ISSN: 1992-8645, E-ISSN: 1817-3195.

13. U.Kamanur, A. Sharipbay, G. Altenbek, L.Zhetkenbay, G. Bekmanova Investigation and use of methods for defining the extends of

similarity of Kazakh language sentences. 15th China National Conference, CCL 2016, and 4th International Symposium, NLP-NABD 2016, Yantai, China, October 15-16, 2016, Proceedings, P. 153–161.

14. Tukeyev, U., Zhumanov, Zh., Rakhimova, D., Kartbayev, A. Combinational Circuits Model of Kazakh and Russian Languages Morphology. Abstracts of International Conference "Computational and Informational Technologies in Science, Engineering and Education", P. 241–242. Al-Farabi KazNU Press, Almaty (2015).

15. Gulshat Kessikbayeva, Ilyas Cicekli. Rule Based Morphological Analyzer of Kazakh Language. Linguistics and Literature Studies 4(1): 96-104, 2016.

16. Gulshat Kessikbayeva Hacettepe University, Department of Computer Engineering, Ankara,Turkey shatik2030@gmail.com Ilyas Cicekli Hacettepe University, Department of Computer Engineering, Ankara,Turkey ilyas@cs.hacettepe.edu.tr.

17. Olzhas Makhambetov, Aibek Makazhanov, Zhandos Yessenbayev, Islam Sabyrgaliyev, and Anuar Sharafudinov. Towards a data-driven morphological analysis of Kazakh language. Türkiye bilişim vakfi Bilgisayar bilimleri ve mühendisliği dergisi.

18. Gülşen Eryiğit, Eşref Adalı. An Affix Stripping Morphological Analyzer For Turkish. In Proceedings of the International Conference on Artificial Intelligence and Applications, pages 299–304, Innsbruck, 16–18 February.

19. Gülşen Eryiğit, Gülşen Eryiğit, Eşref Adalı. Synthetic Turkish Word Root Generation. Proceedings of the Turkish Artificial Intelligence and Neural Networks, TAINN 2003, Canakkale,Turkey.

20. V´ıt Baisa, V´ıt Suchomel. Large Corpora for Turkic Languages and Unsupervised Morphological Analysis.

21. Ahmet Af¸sin Akın and Mehmet Dündar Akın. 2007. Zemberek, an open source nlp framework for Turkic languages. Available at http://zemberek.googlecode.com/.

22. Çagrı Çöltekin and Cem Bozsahin. 2007. Syllables, mor- ˇ phemes and Bayesian computational models of acquiring a word grammar. In Proceedings of 29th Annual Meeting of Cognitive Science Society, pages 887–892, Nashville.

23. Aslı Göksel and Celia Kerslake. 2005. Turkish: A Comprehensive Grammar. London: Routledge.

24. Dilek Z. Hakkani-Tür, Kemal Oflazer, and Gökhan Tür. 2002. Statistical morphological disambiguation for agglutinative languages. Computers and the Humanities, 36(4):381–410.

25. Jorge Hankamer. 1986. Finite state morphology and left to right phonology. In Proceedings of the West Coast Conference on Formal Linguistics, volume 5. Stanford Linguistic Association.

26. Gruber, T.R. (1995). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Human-Computer Studies Vol. 43, Issues 5-6, 907–928

27. Khakhalin, G. (2009). Applied Ontology in the language of hypergraphs (in Russian). Proceedings of 2nd All–Russian Conference "Knowledge–Ontology–Theory" (KONT-09). Novosibirsk, 223–231.

СЕКЦИЯ 5 | **МАШИННЫЙ ПЕРЕВОД**

УДК 81'33

MODELING WORD COMBINATIONS IN TERMS OF PARTS OF SPEECH IN THE PROCESS OF ENGLISH-UZBEK MACHINE TRANSLATION

N. Abdurakhmonova
Doctoral student (PhD) Tashkent state university of the Uzbek language and literature named after Alisher Navoi
abdurahmonova.1987@mail.ru

The article is devoted to the modelling of word combination and classification of parts of speech in the process of English-Uzbek machine translation. The article shows the importance of using models of words combinations and the order of parts of speech in the text for machine translation system.

Keywords: linguistic database, word combination, machine translation, syntactic structures.

МОДЕЛИРОВАНИЕ СЛОВОСОЧЕТАНИЙ ПО ЧАСТЯМ РЕЧИ В ПРОЦЕССЕ АНГЛО-УЗБЕКСКОГО МАШИННОГО ПЕРЕВОДА

Н. Абдурахмонова
Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои
abdurahmonova.1987@mail.ru

Статья посвящена моделированию словосочетания и классификации по частям речи в процессе англо-узбекского машинного перевода

Ключевые слова: Лингвистическая база данных, словосочетание, машинный перевод, синтактические структуры.

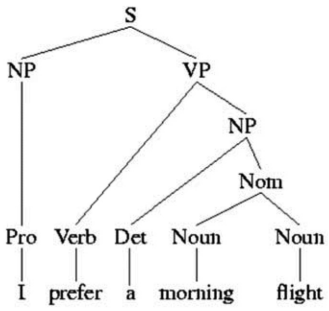
Obviously, there are about more than 3000 types of languages in the world. (7000 familiy types of the language in some literatures). Uzbek language belongs to the Karluk- Uigur group of Altai family of the languages. The Uzbek language is becoming more important in consequence of economic and political issues in globalization process. Furthermore more than 25 mln. (40 mln. in some statistics) people speak in Uzbek. Therefore, the importance of Uzbek language is crucial in process of translation studies for intercultural communication and diplomatic attitudes. As we live in cyber age, the process of information exchange is picking the pace. There are various approaches to create the system of machine translation. Machine translation is one of the type of the process to create database for Uzbek to translate into other languages. Each application based on linguistic database of natural language processing. Nevertheless, neuro machine translation, statistical machine translation, syntactical based translation methods are still being investigated in this field since it began to study last century, rule-based translation dominates in Turkic languages owing to multitude of suffix combinations in word forms by grammatical approach.

The article devoted to comparing English and Uzbek parts of speech and word combination in terms of syntactic models for machine translation. Word combination considered main material of syntactic construction of the text as unit of speech. However, the term of word combination used such “word collocations”, “words that go to together” these mean more than two relative syntactic attitudes of words. As an example, English dictionary has 14 thousand lexemes in word combinations and 75 thousand grammatical and lexical units in content[1]. Word combinations were not on focus from 1954 until 1984 years thanks to negative attitudes to it, on the other hand machine translation based on word-to word strategy. After those years, time showed how it would be important free word combination in the text to translate from source language into target language properly. Because each lexeme gives a real meaning only in the communicative act or written discourse. Therefore, a word never exists lonely separated from the speech to give full information, its place designating by contacting with other parts of sentences. Now we can see RETRANS (English-Russian polythematic phraseological dictionary) machine

[1] Benson M., Benson E., Ilson R. The BBI Combinatory Dictionary of English.1990.

translation system has 2,6 mln. words and 250 thousand Russian word combination of the synonyms[1]. Syntactic structures in English grounded on context-free grammar CFG (Jurafsky). Firstly, it is essential to analyze main parts of speech in the stage of syntax. CFG includes the following components: terminal symbols, nonterminal symbols, grammar rules, initial symbols.

S → NP VP I + want a morning flight
NP → Pronoun I
| *Proper-Noun* Los Angeles
| *Det Nominal* a + flight
Nominal → Noun Nominal morning + flight
| *Noun* flights
VP → Verb do
| *Verb NP* want + a flight
| *Verb NP PP* leave + Boston + in the morning
| *Verb PP* leaving + on Thursday
PP → Preposition NP from + Los Angeles



There are three types of sentence: Declarative structure, imperative structure, interrogative structure (yes-no-question structure, and wh-question structure **Declarative structure** *->NP VP (I attend to English course)*

Imperative structure *-> VP (Show me the meaning to live lonely)*

Interrogative structure *(*yes-no-question structure*) -> S → Aux NP VP (Do you go to the station?)* ***wh-subject-question*** *structure S → Wh-NP VP (Who is interviewing now ?) The* ***wh-non-subject-question*** *structure S → Wh-NP Aux NP VP (Where do you want to stay on your holiday?)*

We have to analyze what word phrases surrounds each components. If determiners come before noun, then it should be clarified as D-> Article|Pronoun (possesive, demonsrative, Inclusive quantifiers) | Numeral (cardinal, ordinal)| Quantifiers|Adjective. Determiners match attributes in Uzbek in condition in spite of articles. Both of the

[1] Хорошилов А.А. Теоретические основы и методы построения систем фразеологического машинного перевода//дисс…техн. Наук. 2006.

languages natural qualities are to be considered. Besides there are uncountable nouns, proper noun, abstract noun, collective noun that do not accept determiners (articles). It should be taken into account that in order to construct the model of the order of components word combination. Adjective phrase (combinational adjective – CA). Adverbs come before Adjective: *the **least** expensive fare,* additionally here is some example of models:

NP → (Det)-> (Card) (Ord)-> (Quant) ->(AP)-> Nominal

– The -> first -> few -> hard-nosed ->ways

The units of parts of speech before subject (noun combination) constructs syntactic models. These translated into Uzbek as extended noun combinations. Such models in English translated like extending units in Uzbek. For instance, prepositional phrases: *all students in auditory*; non-finite forms of the verb: *some fruits buying from the shop; relative clauses: My student who studies at university.* Post modifier: *a reservation [on flight six oh six] [from Tampa] [to Montreal] Nominal PP (PP) (PP). Moreover, gerund (V+-ing), Infinitive (to Verb), Participle I ($V_{III,-ed}$)*: *Which is the aircraft used by this flight? Nominal clauses: Nominal → Nominal RelClause; RelClause → (who | that) VP. Parts of speech contacts each other through the conjunctions*: *NP → NP and NP; VP → VP and VP; S → S and S.*

The process of machine translation needs finding equivalents of parts of speech. That is why our analysis implemented by three basic stages:

1) Parts of speech
2) The elements of speech
3) Types of sentence

Here is parts of speech in Uzbek:

Main parts of speech		**Helping words**	**Separated group**
Noun	Adverb	Conjunction	Exclamation
Verb	Numeral	Particle	Imitation word
Adjective	Pronoun	Auxiliary word	Modal words

As we see above there are differences between English and Uzbek. Building database of machine translation we have to direct our attention how to join common and differential categories. However,

there are dissimilar approaches for classification of parts of speech in English. In some literatures, we can see the following classification [3]:

I. Main parts of speech: 1) noun; 2) adjective; 3) verb; 4) stative verb (want, know); 5) pronoun; 6) numeral; 7) adverb; 8) modal words; 9) particle (as the part of phrasal verb); 10) exclamation

II. Secondary parts of speech: 1) preposition; 2) conjunction

The other literature shows [4]:

I. Main parts of speech: 1) noun; 2) adjective; 3) verb; 4) stative verb (want, know); 5) pronoun; 6) numeral; 7) adverb; 8) modal word; 9) exclamation

II. Secondary parts of speech: 1) preposition; 2) conjunction; 3) particle (as the part of phrasal verb); 4) article

But another one source indicates the following classification of parts of speech [5]:

I. Main word group: 1) verb; 2) noun; 3) adjective; 4) adverb; 5) preposition; 6) determinant, 7) pronoun; 8) conjunction.

Verb, noun, adjective, adverb –*vocabulary words*; preposition, pronoun, determinant, conjunction –*grammatical words*. Determinants also divides article, possessive pronoun, demonstrative pronoun, numerative words

In addition to those classifications, Jurafsky divides into *open and close group of parts of speech.* If main parts of speech included as open group, the second type may be considered closed type [6].

The difference between classifications of parts of speech affects to the quality of translation. After adjusting valences of categories, there will be not any hindrance to separate word combinations. There are two types of syntactic joining: equal joining and dependent linking.

Equal joining made by the following grammatical categories:

1) *copulative conjunction:* va, hamda, -u, -yu, ham (asr va yoshlar – age and youth); 2) *adversative conjunction*: ammo, lekin, biroq (oq lekin rang emas – white but it is not colour); 3) *disjunctive conjunction*: goh…goh, dam…dam, ba'zan…ba'zan, yoki, yo…yo (yo hamma, yo hech kim – either everybody or nobody)

Relative joining has three types:

1) Subordinating; 2) Coordinative; 3) Contiguity

Dependent linking of the word combinations has two parts: notional and helping word. What part of speech is in the notional part: 1) nominal phrases; 2) verbal phrases.

– *Nominal phrases* =dependent word+ main word (noun, adjective, numeral, pronoun, adverb, infinitive)

– *Verbal phrases*=dependent word +main word (verb, gerund)

Taking into account above- mentioned syntactic relation we could analyze the models of word combination comparing Uzbek-English translation according to parts of speech of both of the languages.

Parataxis word combination with noun

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.+ravishdosh+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 (but *ism qo‘yish, nonushta qilish* these are not word combinations, these are compound verb) <=> reading a book

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

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

Parataxis word combination with verb

1. Adj.+Verb=> yaxshi o‘qimoq<=>V+Adv=>read well

2. Adv.+Verb=> astoydil o‘qimoq<=>V+Adv=>study hard

3. Gerund+Verb=> kulib gapirmoq<=>V+Prep.+Gerund=>speak with smiling

Parataxis word combination with Noun

Parataxis word combinations are structured with the suffixes of accusative (-ni), locative (-da), dative (-dan) cases and auxiliary words in Uzbek language.

1. Noun+ dan+Noun=> Andijondan xat <=>Noun+Prep.+Noun=> the letter from Andijan

2. Noun+dan+↓ham|↓-da|↓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.}+Ge rund=>doubtful of seeing

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

7. ot|olmosh+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=>yozuvda 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

Direction word combination with verb

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

Auxiliary word –

A_1 (Auxiliary word comes after nominative case: bilan (qalam bilan, sening bilan, seningla), uchun, kabi, singari, yanglig‘, sayin,

sari, sababli, orqali, tufayli, chog‘li, osha, bo‘ylab, uzra, ichra, bo‘yi, chamasi, haqda//to‘g‘rida, haqida//to‘g‘risida, holda//yo‘sinda);

A_2 (Auxiliary word comes after the dative case: tomon, qadar, ko‘ra, qarshi, qarab, qaraganda, qaramasdan/qaramay, yarasha, doir, asosan, binoan, muvofiq, qarata);

A_3 (Auxiliary word comes after accusative case: ost, tag, ust, ust, old, orqa, yon, qarshi, bo‘yi, ich, o‘rta, bosh);

A_4 (Auxiliary word comes after prepositional case: so‘ng, keyin, boshqa, bo‘lak, tashqari, o‘zga, beri, buyon, nari/nariga, burun, ilgari, oldin, avval, boshlab, tortib)

8. Noun+A_1+Verb=> maktab sari ketmoq –to go to|along to school

9. Noun+A_2+Verb=>taklifga binoan kelmoq –coming in terms of invitation

10. Noun+A_3+Verb=>uyning orqasida turmoq–to stand at back of the house

11. Noun+A_4+Verb=>hammadan burun boshlamoq–to begin in all of them

Coordinative word combination

1. Noun|Pronoun+ning+Noun=>kitobning muqovasi <=> Noun+ +of+Noun=> the cover of the book

2. Noun+Noun=>talabalar soni<=>Noun+Noun=>the students number

In conclusion, we can estimate that word combination models and the place of parts of speech are not only for syntactic analysis but also to identify fit the model of the text and by order, they should be decoded into another one language.

REFERENCES

1. Benson M., Benson E., Ilson R. The BBI Combinatory Dictionary of English.1990

2. Хорошилов А.А. Теоретические основы и методы построения систем фразеологического машинного перевода//дисс…техн. Наук. 2006.

3. Кобрина Н.А., Корнеева Е.А. и др. Грамматика английского языка. Морфология. Синтаксис. Санкт-Петербург 2008. С. 7.

4. Каушанская В.Л., Ковнер Р.Л., Кожевникова О.Н., Прокофьева и др. A grammar of the English language, M., 2008 B. 14.

5. John Eastwood Oxford guide to English grammar. Oxford University 2002 P. 10.

6. Juravskiy Martin. Speech and language processing. 2007. P. 140.

7. Ҳакимов М.Х. Математические модели узбекского языка. ЎзМУ хабарлари, № 3, 2010, с. 185–188

УДК 81’33

ALGORITHM BASED ON LINGUISTIC MODELS IN MACHINE TRANSLATION BETWEEN ENGLISH AND UZBEK

N. Abdurakhmonova[1], X. Axmedova[2]
[1] Tashkent State University of the Uzbek Language and literature named after Alisher Navoi
abdurahmonova.1987@mail.ru
[2] University of the economy and diplomacy of the world

The article is devoted to the analysis of simple sentences’ structure of English and Uzbek languages. We propose an algorithm that solves crucial problem for machine translation of these unrelated languages, and the linguistic database that gives the possibility to implement the process of machine translation.

Keywords: database, machine translation, tokenization, programming and linguistic database, algorithm.

АЛГОРИТМ, ОСНОВАННЫЙ НА ЛИНГВИСТИЧЕСКОЙ МОДЕЛИ АНГЛО-УЗБЕКСКОГО МАШИННОГО ПЕРЕВОДА

Н. Абдурахмонова[1], Х. Ахмедова[2]
[1] Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои
abdurahmonova.1987@mail.ru
[2] Университет мировой экономики и дипломатии

В статье анализировано описание алгоритма программного языка Java, основанный на лингвистической модели машинного перевода.

Ключевые слова: база данных, машинный перевод, токенизация, программирование и лингвистические базы данных, алгоритм.

Computational linguistics is one of the complicated fields which crossroads of linguistics and computational technologies. Because it links directly with natural language processing, indeed it also depends on several factors that are psychological, cognitive, and cultural and so on. Nevertheless, translation is not only technical process but also creative activity that based on including both material and mental capability of human being. Therefore, for machine translation it is important to identify what kind of texts would be objects in the automatic process. We clarify the text in terms of genres like official or scientific texts that are more formal than others are. However, a lot of breakthrough in the field involving oral and written form of all genres texts since many attempts have been implemented over the world. Regarding progress, today as we mention some approaches of machine translation like neuro machine translation, statistical, phrasal-based etc. Owing to globalization and interactive communication between nations in Internet, translation tools have a pivotal role to ease and make the atmosphere that is necessary and so fast with quality to take daily information and transform them consumer as soon as possible. It is not even in social networking, but exchange academic background at any time at different parts of the world gives a great chance to analyze and criticize them wherever its needed. Therefore, in machine translation the Uzbek language is important as it one of Turkic language.

Our article is focus how to build up algorithm for machine translation from English into Uzbek and vice versa.

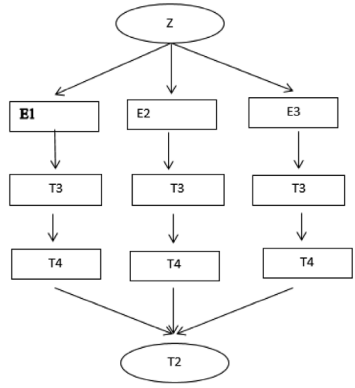
Firstly, it is applied morphological analysis in the first stage: tokenization (take apart word form) -> lemmatization (the analysis of morphemes) -> stemming (identify the roots of the words). Thereafter syntactic models of the text compared and checked each other.

Obviously, database is well structured systematically and by structure to keep data that are used in urgent time accurately and properly which are asked somehow. It is should be input symbols for environment of machine translation.

Data Name	Function
R_i	The database of phrase and terms of the scientific spheres.
Q1	The database of all of the words root in the language.
K1	The database of all derivational words
V2	Clause elements
V3	The database of parts of speech

The tables are created for each language. The environment translation services for scientific text. It is very important to address Grammar of the languages so that to identify the structure of the sentence and parts of speeches in the text. It could do this work through two directions: English-Uzbek, Uzbek-English.

Firstly, dividing into several parts of speech of input text (Z) and each words are taken the other term database; they are replaced in terms of grammar. We display the functional chart of translation algorithm:



The following symbols input in the entry part of language in order to model of natural language:

– $T3_{i1}$-translation into other language and the massive including the function in the sentence,$1 \le i1 \le m$;

– $T4_{j1}$- translation into other language,$1 \le j1 \le m1$;

– T2-translated text;

– E4-subject; G2 -predicate; E5-attribute;

– E6-object; E7- modifier.

There are two appropriate models of sentence in both of languages.

a) the different mathematical models of types of indicative mood in Uzbek:

I. 1. <E4>↓⊕<E5>↓⊕<E6>↓⊕<E7> ⊕<G2>.
2. ↓<E5> ⊕<E4>↓⊕<E6>↓⊕<E7> ⊕<G2>.
3. ↓<E5>↓⊕<E5> ⊕<E4>↓⊕<E6>↓⊕<E7> ⊕<G2>.
4. <E4>↓⊕<E5>↓⊕<E6> ⊕<G2>.
5. <E4> ⊕<G2>.
6. <E4>↓⊕<E7> ⊕<G2>.
7. <E4>↓⊕<E6> ⊕<G2>.

Thus we apply a bit change of mathematical models which presented at [1,3,4] types of component of sentence. Hence, some exact parts of speech could be appropriate clause elements in some cases that identified as models of the text. Afterwards it is taken from other translation in the second language and it is replaced in order by normal principles. In next stage algorithm takes function in order to the most optimal and meaningful translation. Above mentioned the forms Uzbek sentences are formed as English mathematical models:

II. 1. <E4>↓⊕<E5> ⊕<G2>↓⊕<E6>↓⊕<E7>.
2. ↓<E5>⊕<E4>⊕<G2>↓⊕<E6>↓⊕<E7>.
3. ↓<E5>↓⊕<E5>⊕<E4> ⊕<G2>↓⊕<E6>.
4. ↓<E5>⊕<E4> ⊕<G2>↓⊕<E6>.
5. <E4> ⊕<G2>.
6. <E4> ⊕<G2>↓⊕<E7>.
7. <E4> ⊕<G2>↓⊕<E6>.

b) Let's take the maphematic models of simple interrogative sentences of Uzbek language as an example:

1.<M4>↓⊕<E5>↓⊕<E5>↓⊕<E6>⊕<G2>
2. <M4>↓⊕<E6>↓⊕<E5>⊕<G2>
3. ↓<E6>⊕<M4>⊕<G2>
4. <M4>↓⊕<E5>↓⊕<E6>⊕<G2>
5. ↓<E5>↓⊕<E6>⊕<M4>⊕<G2>

These interrogative sentences suit in English such models as following examples:

1. < M4 >⊕<G2>↓⊕<E7>↓⊕<E6>
2. < M4 >⊕<G2>↓⊕<E6>
3. < M4 >⊕<G2>↓⊕<E5>↓⊕<E7>
4. < M4 >⊕<G2>↓⊕<E6>↓⊕<E7>

Using above mentioned database structure of sentences and terms, translation algorithm is given like this:

Q1_uz=» SELECT * FROM `Q1_uz`»-all stems in Uzbek;

K1_uz=» SELECT * FROM `K1_uz`»-all word forms in Uzbek;

Q1_eng=”SELECT * FROM `Q1_eng`”- all stems in English;

K1_eng=”SELECT * FROM `K1_eng`;”- all word forms in English;

E_i – sentence taken from text Z, $1 \le i \le n$;

$L1_J$ – words taken from E_i, $1 \le j \le n1$;

After doing algorithm [2], the following “search” algorithm divides into Z sentences, and after that it breaks apart words or word combinations, then each word formations is searched in the database of stem list, if there is not need words turning another one type of database. After finding words, taken translation form the target language. As we take one more example for Uzbek-English direction the 1st translation algorithm like this:

1. Search the words in $L1_j$ from Q1_uz. If find go 2nd step, otherwise 4th step;

2. Take the stem from Q1_uz in terms of English order (ID);

3. Take translation of sterom m fQ1_eng and go through the 7th step;

4. Search each word in $L1_j$ from K1_uz;

5. Take the order (ID) word formation in K1_eng form K1_uz;

6. Take translation of word formation from K1_eng;

7. Identify the function in the sentence and replace in the massive $T3_{i1}$;

8. Pass filled massive of $T3_{i1}$ to function UzbekIngliz ($T3_{i1}$);

Replace the results of function UzbekIngliz ($T3_{i1}$) to T2;

Here UzbekIngliz ($T3_{i1}$) [2] function which is written translation algorithm for Uzbek-English direction. UzbekIngliz($T3_{i1}$) function is written as following. So we used some signs to write function:

$ET3_{k1}$ –Uzbek and English the structures that are suited each other $1 \leq k1 \leq m2$;

1. Load the functions of words which are input $T3_{i1}$ to $E8_k$ massive;

2. Find appropriateness structure sentence to $E8_k$ form from $ET3_{k1}$;

3. Take found the fords as clause elements from $ET3_{k1}$ and load to T2;

This function is such a form in programming language (in Java):

```
private String UzbEng(String suz) throws ObjectNotFoundException {
        int engId=0;
        String engSuz ="";
        int gapBulagiId=0;
        UzakSuzlar us=uzakSuzUzbekDao.getUzakUzbekByWord(suz);
                if(us.getUzakSuzlar().equals(suz)){
                        engId=us.getUzakEnglishId();
        List<UzakEnglish> ueList=uzakSuzEnglishDao.getuzakSuzlarListByRId(engId);
                        for (UzakEnglish ue : ueList) {
                                engSuz=ue.getUzakEnglish();
                        }
                        }else{
        YasamaSuzlar ys=yasamaSuzUzbekDao.getYasamaUzbekBySuz(suz);
                if(suz.equals(ys.getYasamaSuzlar())){
```

```
                engId=ys.getYasamaEnglishId();
    YasamaEnglish ye=(YasamaEnglish) yasamaSuzEnglishDao.getY
asamaEnglishListByRId(engId);
                engSuz=ye.getYasamaEnglish();
                }else{
                engSuz=suz;
                }
                }
            return engSuz;
                }
```

The algorithm 2 is for English-Uzbek direction like this:

Search each word in $L1_j$ from Q1_eng. If it is found, go to the 2nd step, otherwise to the 4th ;

1. Take the order (ID)stem in English from Q1_eng;
2. Take translation stem from Q1_uz and go to the 7th step;
3. Search each word in $L1_j$ from K1_eng;
4. Take the order (ID) in word formation in K1_uz from K1_eng;
5. Take translation derivative word from K1_uz;
6. Identify the function of the word in the sentence and replace in the massive of $T3_{i1}$;
7. Pass filled massive $T3_{i1}$ to function InglizUzbek ($T3_{i1}$);

Replace the results of function InglizUzbek ($T3_{i1}$) to T2;

Here InglizUzbek ($T3_{i1}$) is the function written in [2] based on English-Uzbek translation direction algorithm. InglizUzbek ($T3_{i1}$) function is as following, accordingly used some signs to write function:

$ET4_{k1}$ – Uzbek and English the structures that are suited each other $1 \le k1 \le m2$;

1. Load the function in the sentence of the word input $T3_{i1}$ massive to $E8_k$;
2. Find proper the structure sentence to $E8_k$ from r $ET4_{k1}$;
3. Take clause elements of the words found in $ET4_{k1}$ and load to T2;

These tags represented in the following process:

```
    private String EngUzb(String suz) throws
ObjectNotFoundException {
                int uzakId=0;
                String uzbSuz ="";
                int gapBulagiId=0;
```

```
            UzakEnglish ue=uzakSuzEnglishDao.
getUzakEnglishByword(suz);
            if(ue.getUzakEnglish().equals(suz)){
                uzakId=ue.getUzakSuzlarId();
            List<UzakSuzlar> usList=uzakSuzUzbekDao.
getuzakSuzlarListByRId(uzakId)
                        for (UzakSuzlar us : usList) {
                            uzbSuz=us.getUzakSuzlar();
                            }
                        }else{
            YasamaEnglish ye=yasamaSuzEnglishDao.
getYasamaEnglishByWord(suz)
                        if(suz.equals(ye.getYasamaEnglish())){
                        uzakId=ye.getYasamaSuzlarId();
  YasamaSuzlar yu=(YasamaSuzlar) yasamaSuzUzbekDao.getYasa
maSuzlarListByRId(uzakId);
            uzbSuz=yu.getYasamaSuzlar();
                        }else {
                        uzbSuz=suz;
                        }
                        }
                    return uzbSuz;
                        }
```

In conclusion we may say that although our investigation on machine translation system seems a bit a simple, there are very pivotal issues should be done in terms of linguistic models. According to this rule based translation is important for non familiar and relative languages like English and Uzbek. In the future, our research will be directed multilingual machine translation system for the Uzbek language.

REFERENCES:

1. Abdurakhmonova N. Z. Grammatical analyze in machine translation 1-я Международная конференция “Компьютерная обработка тюркиских языков. Латинизация письменности” Казахстан, Астана, 2013.

2. R.Delmonte. Computational Linguistic Text Processing: Lexicon, grammar, Parsing and Anaphora Resolution, Nova Science Publishers, Inc. New York, 2008, 4-5 Ps.

3. Абдураҳмонова Н., Ҳакимов М.Х. Логико-лингвистические модели слов ипредложений английского языка для многоязычных ситуаций компьютернеого перевода. 1-я Международная конференция “Компьютерная обработка тюркиских языков. Латинизация письменности” Казахстан, Астана, 2013, С. 302–306.

4. Abdurakhmonova N. Z. Automatic morphological analyze for English-Uzbek system // Известия Кыргызский государственный технический университет им. И.Раззакова. Теоретический и прикладной научно-технический журнал № 2 (38) 2016, С. 12–18.

5. Абдурахмонова Н. Ҳакимов 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.

6. Ахмедова Х.И. «Моделлашган компютер таржимаси технологиясининг алгоритмлари» Амалий математика ва информацион технологияларнинг долзарб муаммолари Ал-Хоразмий 2014. Самарқанд 15-17-сентабр 2014 йил.

7. Ҳакимов М.Х. Математические модели узбекского языка. ЎзМУ хабарлари, № 3, 2010, с. 185–188.

8. Ҳакимов М.Х. «Семантическые базы и математические модели русского языка для многоязычных ситуаций компьютерного перевода» ЎзМУ хабарлари, № 2, 2011, с. 57–64.

9. Шаляпина З.М.Текст как объект автоматического перевода. – В кн.: Текст и перевод. – М.: Наука, 1988, с. 113–129.

10. Марчук Ю.Н. Компютерная лингвистика (учеб. пособ.) Москва, Восток Запад, 2007, 61–б.

УДК 81'33

EVALUATION OF THE QUALITY OF MACHINE TRANSLATION IN THE ASSIMILATION SCENARIO FOR ENGLISH-KAZAKH AND KAZAKH-RUSSIAN LANGUAGE PAIRS

Zh. Zhumanov[1], D. Amirova[2]
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
z.zhake@gmail.com
[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan
amirovatdina@gmail.com

This paper is devoted to the problem of evaluating quality of machine translation systems in the scenario of understanding general meaning of translated text (assimilation). Some possible ways to assess understanding are provided. We describe an experiment conducted on the basis of gap filling method for English-Kazakh and Kazakh-Russian language pairs from Apertium machine translation platform. The results of the experiment are presented.

Keywords: evaluation of machine translation quality; gap filling method.

ОЦЕНКА КАЧЕСТВА МАШИННОГО ПЕРЕВОДА В СЦЕНАРИИ АССИМИЛЯЦИИ ДЛЯ АНГЛО-КАЗАХСКОЙ И КАЗАХСКО-РУССКОЙ ЯЗЫКОВЫХ ПАР

Ж.М. Жуманов[1], Д.Т. Амирова[2]
[1]КазНУ им. аль-Фараби, Алматы, Казахстан
z.zhake@gmail.com
[2]КазНУ им. аль-Фараби, Алматы, Казахстан
amidi_92@mail.ru

Данная статья посвящена проблеме оценки качества систем машинного перевода в сценарии понимания общего смысла переводимого текста (ассимиляции). Приводятся некоторые возможные способы оценки понимания. Описывается эксперимент на основе метода восполнения пропусков, выполненный для англо-казахской и казахско-русской языковых пар платформы машинного перевода Apertium. Приводятся результаты эксперимента.

Keywords: оценка качества машинного перевода; метод восполнения пропусков.

1. Введение

Текущие версии различных систем машинного перевода не обеспечивают пользователей полностью автоматическими, высококачественными переводами. Качество получаемого перевода, как правило, значительно ниже, чем у перевода, выполненного человеком, и специфические ошибки встречаются чаще. Практика показывает, что постоянное участие человека на различных этапах процесса перевода имеет важное значение для обеспечения надлежащего качества перевода.

Системы машинного перевода, как правило, используются в одном из двух сценариев:

1) получение такого перевода текста, который (с возможным постредактированием) подходит для публикации;

2) получение такого перевода текста, который, не являясь грамматически или стилистически корректным, все же позволяет пользователю понять суть оригинала.

Первый сценарий называют «распространение информации» (диссеминация). Переводы такого типа обычно используются для публикации или массового распространения. Требования к качеству, как правило, высоки во всех отношениях: точность, читаемость, стиль. В этом случае важное значение приобретают изменения или корректировки, применяемые к машинному переводу. Процесс корректировки называют постредактированием.

Второй сценарий называют «понимание сути информации» (ассимиляция). Переводы такого типа в основном используются для усвоения основного смысла информации в текстах, например, «беглые» переводы научно-технических документов позволяют исследователям и инженерам отслеживать тенденции в изучаемых ими областях. Для такого вида перевода, требования к качеству могут быть невысокими. Однако перевод должен быть, по крайней мере, понятным с минимальным объемом постредактирования или, предпочтительно, без какого-либо постредактирования. Особенность ассимиляции заключается в том, что пользователь не знает иностранного текста, в то время как при диссеминации, пользователь является переводчиком, который знает иностранный/исходный язык (и часто будет читать как исходный текст, так и перевод).

2. Обзор методов оценки машинного перевода в сценарии ассимиляции

В результате широкого распространения систем машинного перевода, доступных онлайн, понимание сути переводимого текста (ассимиляция) стала наиболее часто используемым сценарием для машинного перевода. Существует несколько методологий оценки ассимиляции при использовании машинного перевода. Они включают в себя как постредактирование и сравнение экспертами-билингвами (Ginestı-Rosell et al., 2009), так и тесты со множественным выбором (Jones et al., 2007) (Trosterud, Unhammer, 2012). Но данные подходы являются затратными и имеют склонность к субъективности. В качестве альтернативы, в (Taylor, 1953) для оценки ассимиляции была представлена модификация обучающего теста, сначала в качестве дополнения к методу, описанному в (Trosterud, Unhammer, 2012), а затем уже в качестве отдельного метода (O'Regan, Forcada, 2013).

Ранее тесты использовались для оценки качества «сырого» (неотредактированного) машинного перевода (Van Slype, 1979), (Somers, Wild, 2000). Авторы просили участников заполнить пропуски в машинном переводе. Позже другие исследователи в своих экспериментах просили участников заполнить пропуски в отредактированных предложениях (Trosterud, Unhammer, 2012), (O'Regan, Forcada, 2013). Основная идея заключается в том, что определенное количество ключевых слов удаляется из перевода предложения, выполненного человеком. Участников теста просят заполнить пропуски подходящими словами с помощью или без помощи машинного перевода. Восполнения пропусков имеет соей целью оценить, насколько хорошо пользователи понимают ключевые точки в тексте, это эквивалентно ответам на вопросы по тексту. Таким образом, такой подход не оценивает напрямую качество перевода, выполненного системой машинного перевода. Он оценивает его полезность в понимании смысла исходного текста.

В одном из методов авторы для оценки предлагают, кроме заполнения пропусков, еще два теста (Trosterud, Unhammer, 2012). В первом тесте участникам необходимо выбрать один из 3 вариантов перефразированных переводов, составленных человеком,

имея при этом текст на исходном языке и машинный перевод. Во втором тесте участникам необходимо ответить на открытый вопрос, ориентируясь на машинный перевод.

3. Описание метода восполнения пропусков

Метод восполнения пропусков для оценки задачи ассимиляции в машинном переводе основан на следующем предположении: понимание читателями текста соотносится с количеством слов, которые они могут корректно поставить вместо пропусков. Следовательно, основа для оценки понимания сути текста (ассимиляции) это предложение на целевом языке, в котором отсутствуют (удалены) некоторые слова. Предложение составлено человеком (в противовес машинному переводу) на языке, который понятен оценивающему лицу (в системах машинного перевода он называется целевым языком). Дополнительные элементы задания это вспомогательные данные, или дополнительные предложения, которые помогают участнику понять главное предложение. Существует два вида вспомогательных данных. Первый – исходное предложение, которое семантически эквивалентно предложению на целевом языке, также составленное человеком, но на исходном языке. Второй вид вспомогательных данных – это машинный перевод исходного предложения.

Вспомогательные данные можно комбинировать следующим образом (Ageeva, et al., 2015):

- Только предложение на целевом языке. В этом случае пропуски заполняются, без какого-либо контекста. Эта задача служит базовым уровнем оценки и индикатором пропусков, которые могут быть заполнены, с помощью общеупотребительных слов или основываясь на интуиции (как в случае идиом или устойчивых выражений). Например, в английской фразе «Jack ordered _____ and chips» правильным ответом является «fish». Такой ответ, однако, может быть не связан со значением исходного текста, и может быть дан только на основе знания устойчивых словосочетаний.
- Предложение на целевом языке и предложение на исходном языке. Это может помочь при заполнении имен собственных или заимствованных слов.

• Предложение на целевом языке и машинный перевод. Вдобавок к предложению на целевом языке также предоставляют машинный перевод исходного предложения. Такое сочетание используется для оценки вклада машинного перевода в понимание сути текста.

• Предложение на целевом языке и два вида вспомогательных данных. Эта комбинация позволяет проверить, насколько достаточно, машинный перевод и предложение на исходном языке, определяют дополнительные данные.

Для подготовки тестовых заданий, проверяются и удаляются ключевые слова из предложения на целевом языке. Рассматриваются два параметра: список, разрешенных частей речи, и количество пропусков, которое зависит от длины предложения («плотность пропусков»). Для оценки может применяться плотность пропусков в 10%, 20% или 30%, и следующие части речи: имена существительные (включая имена собственные), прилагательные, наречия, глаголы.

Для каждого предложения подготавливается список ключевых слов. Он состоит из всех слова, которые входят в список разрешенных частей речи. Количество пропусков высчитывается на основе длины предложения и плотности пропусков. Все предложения на целевом языке имеют длину более 10 слов. Требуемое количество ключевых слов выбирается из списка кандидатов таким образом, чтобы количество пропусков распределяются равномерно по всему предложению. Начинают со случайного слова в предложении и проверяют, является ли данное слово ключевым кандидатом. Если да, то это слово удаляется, и продвигаются на следующие n шагов. Также можно вернуться к началу предложения, если необходимо. Длина шага n это длина предложения, разделенная на желаемое количество пропусков. Если слово не является ключевым словом, или уже удалено, рассматривается следующее рядом стоящее слово. Процесс повторяется до тех пор, пока определенное количество слов не будет удалено, или пока в списке не останется ключевых слов.

4. Описание эксперимента

Эксперимент, описанный в предыдущем разделе, был проведен для англо-казахской и казахско-английской языковых пар, используя комбинацию «предложение на целевом языке и два вида вспомогательных данных». В качестве системы машинного перевода использовалась платформа Apertium и ее соответствующие языковые пары.

В эксперименте приняли участие студенты 2, 3, 4 курсов Казахского Национального университета им. аль-Фараби. При проведении эксперимента для англо-казахского и русско-казахского направления перевода участвовали студенты, которые обучаются на казахском отделении; для казахско-английского направления – студенты английского отделения; а для казахско-русского направления – студенты русского отделения. Примеры заданий для англо-казахской и русско-казахской пар языков представлены ниже в таблицах 1-4.

Количество участников эксперимента составило 60 человек для казахско-английской пары и 80 человек для русско-казахской пары языков. Задание состояло из следующих пунктов: заполнить свои данные на листе, прочитать инструкцию и поставленную задачу, заполнить пропуски недостающими словами. Инструкция заключалась в следующем: «Таблица ниже содержит предложения на английском (казахском) языке (столбец «Оригинальный текст»), машинный перевод этих предложений (столбец «Машинный перевод»), и перевод предложений на казахский (английский) язык с пропущенными словами (столбец «Заполнить пробелы в тексте»)». Задача: «Используя предоставленный машинный перевод как подсказку, заполните пропуски в предложениях словами. Если вы не знаете, какое слово должно стоять в пропуске, заполните его словом, которое, по вашему мнению, наиболее подходят к смыслу предложения. В одном пропуске должно быть одно слово». В каждом листе число предложений, в которых необходимо было заполнить пробелы, равнялось 10. На заполнение пропусков в предложении у участников ушло в среднем 20–25 минут.

Таблица 1. **Пример для англо-казахского направления перевода**

Оригинальный текст	Машинный перевод	Заполнить пробелы в тексте
citizenship of the republic of kazakhstan shall be acquired and terminated as prescribed by law shall be indivisible and equal regardless of the grounds of its acquisition	қазақстан республикасының азаматтығы заңға сәйкес алынады және тоқтатылады ол қандай негізде алынғанына қарамастан бірыңғай және тең болып табылады	қазақстан республикасының азаматтығы заңға _____ алынады және тоқтатылады ол қандай _____ алынғанына қарамастан бірыңғай және _____ болып табылады
we need renewed instruments of interaction between the state the non-government sector and business	бізге мемлекеттің үкіметтік емес сектормен және бизнеспен өзара іс-қимылының жаңартылған тәсілдері керек	бізге мемлекеттің үкіметтік емес _____ және бизнеспен өзара іс-қимылының жаңартылған _____ керек

Таблица 2. *Пример для казахско-английского направления перевода*

Оригинальный текст	Машинный перевод	Заполнить пробелы в тексте
Азаматтарды шақыру учаскелеріне тіркеу тұрғылықты мекенжайы немесе уақытша болу орны бойынша жүргізіледі.	Citizens registration to draft offices is carried out according to permanent residency place or temporary stay.	Citizens registration to draft offices is _____out according to permanent residency place or _____stay.
сыбайлас жемқорлыққа қарсы күрес пен құқық бұзушылықтардың алдын алуға ерекше назар аударылатын болады	special attention will be paid to fight against corruption and prevention of crimes	special attention will be paid to fight _____ corruption and prevention of _____

Таблица 3. **Пример для казахско-русского направления перевода**

Оригинальный текст	Машинный перевод	Заполнить пробелы в тексте
Мемлекеттік тілі – қазақ тілі. Орыс тілі ұлтаралық қарым-қатынас тілі мәртебесіне ие.	Государственный язык – казахский язык. Русский язык межнациональное общение статус языка имеет.	Государственный язык – казахский. Русский _____ имеет статус языка межнационального _____.
Тукан құс Оңтүстік Американдан. Ол өзінің ерекше тұмсығымен таңданарлық.	Тукан птица из Южной Америки. Он своим особенным клювом удивлять.	Птица тукан _____ Южной Америки. Он _____своим необыкновенным _____.

Таблица 4. **Пример для русско-казахского направления перевода**

Оригинальный текст	Машинный перевод	Заполнить пробелы в тексте
Законодательные функции выполняет Парламент Республики Казахстан. Он состоит из двух Палат – Сената и Мажилиса.	Заң шығару функция Қазақстан Республиканың Парламенті атқарады. Ол құрылған екі Палатаның – Сенаттың және Мәжілістің.	Заң _____ қызметін Қазақстан Республикасының Парламенті жүзеге асырады. Ол екі Палатадан – Сенаттан және _____ тұрады.
С неба посыпались снежинки. Они кружатся в воздухе и падают на землю.	Аспаннан қар себеледі. Олар ауада айналып және жерге құлайды.	Аспаннан қар _____. Олар _____ айналып жерге құлайды.

Тематика данных в заданиях разная:

- новости;
- художественная литература;
- учебная литература;
- обычная разговорная речь.

Все предложения были отобраны из имеющихся параллельных корпусов.

Процесс удаления ключевых слов является одним из самых важных в эксперименте, так как необходимо определить те слова, которые помогают понять суть предложения. Этот фактор зависел от длины предложения: чем длиннее предложение, тем больше пропусков. В нашем случае, пропуски зависели от длины предложения на целевом языке. Все ключевые слова относились к разным частям речи, которые входят в список разрешенных частей речи (существительные, прилагательные, глаголы, наречия). Также учитывался тот факт, что участники могли заполнить пробелы синонимами, в этом случае, синоним засчитывался как правильный вариант.

При подсчете результатов, сначала считалось количество корректных ответов, количество некорректных ответов, количество корректных синонимов для каждой работы, заполненной участниками. Далее суммировалось количество корректных ответов и корректных синонимов. Полученное число делилось на общее количество пропусков для каждой работы, и рассчитывался общий процент правильных ответов (формула 1).

$$\text{Прав.отв. (\%)} = \frac{\text{(Кол-во коррект.отв.+кол-во кор.синонимов)}}{\text{(общее кол-во пропусков)}} \qquad (1)$$

Для получения окончательных результатов по эксперименту высчитывалось среднее по полученным результатам по всем работам (формула 2).

$$\text{Результат (\%)} = \frac{\text{(}\Sigma\text{прав.отв.)}}{\text{(общее кол-во всех работ)}} \qquad (2)$$

5. Полученные результаты

Результаты оценки понимания сути текста (ассимиляции) по 4 направлениям перевода: англо-казахский, казахско-английский, русско-казахский, казахско-русский – представлены в таблицах 5-6.

Таблица 5. *Результаты англо-казахской языковой пары*

Англо-казахская пара языков	Казахско-английская пара языков
62,8%	68,8%

Таблица 6. **Результаты русско-казахской языковой пары**

Русско-казахская пара языков	Казахско-русская пара языков
76,3%	81,7%

Как видно по результатам, пользователь, которому необходимо перевести какой-либо текст с английского языка на казахский язык, и не владеющий английским языком, сможет понять почти две трети сути переводимого текста. В случае перевода с казахского языка на английский язык, не владея казахским, ¬– чуть больше двух третей переводимой информации. По результатам русско-казахской языковой пары с помощью машинного перевода пользователи усвоили три четверти и более переводимой информации в тексте. Полученные результаты дают возможность предположить, что такая оценка метода отражает значимость вклада машинного перевода в понимание сути текста пользователем. Такой эксперимент можно проводить для любой языковой пары для разных систем машинного перевода. Наличие параллельных корпусов значительно упрощает подготовку заданий для подобной оценки.

ЛИТЕРАТУРА

1. Ginestı-Rosell, M., Ramırez-Sánchez, G., Ortiz-Rojas, S., Tyers, F. M., & Forcada, M. L. (2009). Development of a free Basque to Spanish machine translation system. Procesamiento del Lenguaje Natural, 43, 187–195.

2. Jones, D., Herzog, M., Ibrahim, H., Jairam, A., Shen, W., Gibson, E., & Emonts, M. (2007, April). ILR-based MT comprehension test with multi-level questions. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers (pp. 77–80). Association for Computational Linguistics.

3. Trosterud, T., & Unhammer, K. B. (2012, June). Evaluating North Sámi to Norwegian assimilation RBMT. In Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation (FreeRBMT 2012).

4. Taylor, W. L. (1953). “Cloze procedure”: a new tool for measuring readability. Journalism Bulletin, 30(4), 415–433.

5. O'Regan, J., & Forcada, M. L. (2013). Peeking through the language barrier: the development of a free/open-source gisting system for Basque to English based on apertium. org.

6. Van Slype, G. (1979). Critical study of methods for evaluating the quality of machine translation. Prepared for the Commission of European Communities Directorate General Scientific and Technical Information and Information Management. Report BR, 19142.

7. Somers, H., & Wild, E. (2000, November). Evaluating machine translation: the Cloze procedure revisited. In Translating and the Computer 22: Proceedings of the Twenty-second International Conference on Translating and the Computer.

8. Ageeva, E., Tyers, F. M., Forcada, M. L., & Pérez-Ortiz, J. A. (2015). Evaluating machine translation for assimilation via a gap-filling task. In EAMT-2015: 18th Annual Conference of the European Association for Machine Translation (pp. 137–144).

УДК 81'33

SYNTACTIC AND SEMANTIC REPRESENTATIONS FOR THE DEVELOPMENT OF KAZAKH-RUSSIAN PHRASEOLOGICAL MACHINE TRANSLATION

Zh. Meirambekkyzy[1], A. Choroshilov[2]
[1] Moscow Pedagogical State University, Moscow, Russia
[2] The Institute of informatics problems of the Federal research center «Computer science and control» of the Russian academy of sciences (IPI RAN), Moscow, Russia

In this work we will consider the problem of automated translation from Kazakh language into Russian. The formal model can be applied to languages of the Turkic group. Formalized representation on the basis of formal grammatical systems can be used for different purposes. These methods for the introduction of formal grammatical models of parsing and syntactic transfer of the sentence structure. It developed a program of Kazakh-Russian translations, based on the method of phraseological machine translation.

Keywords: Kazakh syntax, Kazakh-Russian translation, formal models, transformation, linguistic rules, phraseological method, machine translation.

СИНТАКТИКО-СЕМАНТИЧЕСКИЕ ПРЕДСТАВЛЕНИЯ ДЛЯ РАЗВИТИЯ КАЗАХСКО-РУССКОГО ФРАЗЕОЛОГИЧЕСКОГО МАШИННОГО ПЕРЕВОДА

Ж. Мейрамбеккызы[1], А.А. Хорошилов[2]
[1] Московский педагогический государственный университет
[2] Институт проблем информатики Федерального исследовательского центра РАН

Концепция фразеологического метода автоматизированного перевода была сформулирована в 1975 г. известным российским ученым Г.Г. Белоноговым. Согласно этому методу механизм обработки текста происходит преимущественно фразами, синтаксическими конструкциями, выражающие *понятия* и *мыслительные образы*. В статье предлагаются способы формализации казахских синтаксических конструкций для семантико-синтаксического анализа текста. В рамках фразеологического метода были рассмотрены распространенные фразеологические единицы казахского языка в переводе на русский язык, а также сформулированы особенности

трансформаций. Для моделирования лингвистических единиц были использованы правила формально-граммматических систем с дальнейшим внедрением в блоки синтаксического анализа. Основные три ключевых компонента трансформации, на наш взгляд, важно применить в переводе с казахского языка на русский: субъектно-предикатные трансформации; локальные перестановки глагольных словосочетаний, в том числе конструкции со вспомогательными глаголами; замена составных глаголов, вспомогательных глаголов в один переводной эквивалент.

Ключевые слова: казахско-русский машинный перевод, формальные модели, лингвистические правила, локальные трансформации, фразеологический метод перевода.

В данной работе будут рассмотрены задачи автоматизированного перевода с казахского языка на русский, формальные модели, в том числе могут быть применены для языков тюркской группы. Формализованное представление на базе формально-грамматических систем можно использовать для разных целей и в разных метаязыках. Данные методы по внедрению формально-грамматической модели разбора и трансфера синтаксической структуры предложения были применены для системы автоматизированного перевода «Метафраз». Это разработанная программа казахско-русского перевода, основанная на фразеологическом методе машинного перевода. В данной работе будут представлены способы формализации синтаксиса казахского языка, в связи с чем, ставились следующие задачи:

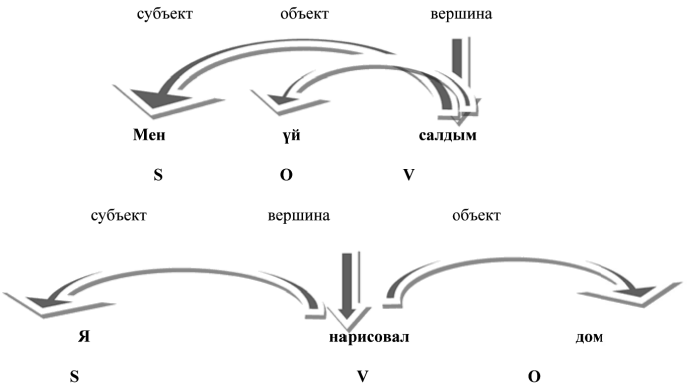
- сформулировать гипотезы трансформаций и правила перевода, с тем, чтобы имеющейся системой все это показать и подтвердить
- выявить распространенные стандартные синтаксические конструкции
- последовательное включение формально-грамматических моделей в блоки синтаксического анализа фразеологического машинного перевода
- лингвистический эксперимент, связь с отображением структуры предложения SOV в структуре предложений SVO (казахский и русский)

В работе рассмотрены вопросы синтаксических трансформаций при переводе с казахского языка на русский. Модели представлены в традиционном для лингвистики виде деревьев, но для

системы Метафраз существует отдельный способ с применением собственного набора тегов и правил. Исследование проводилось на материале казахского корпуса и параллельных текстов, которые были использованы при создании и тестировании данного переводчика. Базовые процедуры системы фразеологического машинного перевода могут использоваться как в системе фразеологического машинного перевода, так и в технологических процессах автоматизированного формирования грамматических таблиц для семантико-синтаксического анализа и синтеза и составления двуязычных словарей.

1. Лингвистический эксперимент, связь с отображением структуры предложения SOV в структуре предложений SVO (казахский и русский)

В нашей работе мы применяем семантико-синтаксический анализ текстов, желая представить его формализованную структуру. Для этого необходимо выделить в тексте лексически значимые единицы и установить наличие связей между ними. Для того, чтобы сравнить структуру предложений языков типов SVO и SOV, корректно будет применить структуру зависимостей. Дерево зависимостей не только показывает отношение доминации подчиняющего слова над зависимым, но также наглядно отражает линейность и последовательность членов предложения:



Данные примеры приведены для простых предложений казахского языка, однако со сложными предложениями субъектно-

объектный порядок также будет неизменен, а вся придаточно-определительная часть будет до предиката. Включение в алгоритм перевода правила относительно порядка слов важно для правильного порядка слов на выходе. Путем семантико-синтаксического анализа исходного языка мы получаем формализованную структуру SOV, далее необходима перестановка. Трансформация осуществляет замену словоформ исходного текста на словоформы на выходном языке. Преобразует информацию о синтаксической структуре казахского текста в информацию, необходимую для синтеза русского текста. На выходе, в итоге семантико-синтаксического синтеза получим правильный порядок слов русского предложения. Процедура замены словоформ исходного текста на словоформы на выходном языке осуществляется при помощи машинного словаря. В результате замены выходной текст представляется сначала последовательностью слов и словосочетаний, расположенных параллельно эквивалентам исходного текста. После этого преобразуется в связное предложение. В этот момент произойдет перестановка объекта и предиката в SVO. В русских предложениях сказуемое согласуется с подлежащим, и, если не содержатся фразеологические единицы, то зависимые слова – «слуги» согласовываются со словами – «хозяевами». Также благодаря типологической классификации по порядку слов в предложении, мы можем извлечь конструкции из предложений и определить фиксированный порядок слов внутри них, или фразеологическую единицу из предложения, что важно для фразеологического метода и расширения словаря фраз. В лингвистической науке считается, что универсальным явлением для всех языков является наличие субъектно-предикативной конструкции, конструкции с родительным падежом (или генитивные конструкции). Существует разные критерии классификации языков по порядку слов[1]. Например, по генитивным конструкциям обозначений обладателя и обладаемого. В английском возможны две: *the house of Tom / Tom's house*, в русском языке также возможны двойные конструкции: *дом отца, отцовский дом*. Для казахского языка один вариант конструкции и четкий порядок слов, сначала следует генитив (Р.п.) – «слуга», затем существительное «хозяйн»: *әкенің үйі*. Ниже приведем таблицу последовательности

[1] Дж. Гринберг. «Антропологическая лингвистика». Вводный курс. М., 2004, С. 141

слов внутри конструкций/фраз в 2 типах языков. Она оформлена по статье известного американского лингвиста Дж.Гринберг.[1] Автор проводил исследование на выбранных 30 языках, среди которых сербский и турецкий. Т.к. эти языки по структуре близки к рассматриваемым нам языкам, мы по аналогии применили эти данные из работы. В фиолетовых строках те конструкции, которым необходимы трансформации в обратный порядок слов при синтезе. В белых строках порядок слов совпадает в исходном и выходном языках.

Таблица 1. **По данным из статьи Дж. Гринберга:**

тип III казахский, турецкий	**тип II русский, сербский**	**примеры**
объект глагола перед глаголом	объект глагола после глагола	**отанды** сүю любить **родину**
прилагательные ставятся перед именами существительными, которые они определяю	прилагательные ставятся перед именами существительными, которые они определяют	**сұр** аспан **серое** небо
послелоги	Предлоги	X кейн после X
существительное следует за определяющим его числительным	существительное следует за определяющим его числительным	**қырық** үй **сорок** домов
вспомогательный глагол следует за главным	вспомогательный глагол предшествует главному	оқығым **келмейді** **не хочу** учиться
существительное следует за указательномым местоимением	существительное следует за указательномым местоимением	**бұл** уй **этот** дом
генитив предшествует имени существительному	Генитив следует за именем существительным, от которого зависит	**ананың** сүты молоко **матери** (**материнское** молоко)

[1] Дж. Гринберг. «Некоторые грамматические универсалии, преимущественно касающиеся порядка значимых элементов». Новое в лингвистике. Вып. 5., М., 1970. С. 114–162.

Продолжение таблицы 1

наречия перед прилагательными, которые они определяют	наречия перед прилагательными, которые они определяют	красиво оформленный
относительное предложение предшествует существительному	существительное предшествует относительному предложению	**дом**, который построил Джек Джек салған **уй**

В дальнейшем список может продолжаться. Применяя морфосинтаксическую таблицу системы Метафраз конструкции были формализованы, что позволяет провести локальнкую трансформацию. По системе Метафраз слова имеют числовую-буквенную форму, в разметку предложения были добавлены вышеописанные лингвистические правила. В результате получаем перевод с правильной трансформацией и синтаксически правильной расстановкой на выходнмо языке.

Таблица 2

лингвистические правила	**выражение в мнемонических обозначениях в системе Метафраз**
объект глагола перед глаголом	object V
вспомогательный глагол следует за главным	V V4
генитив предшествует имени существительному	Gen2 N
относительное предложение предшествует существительному	Aa N

Существуют разные способы, метаязыки, модели формализации синтаксических конструкций. Однако машинная обработка предложений с использованием контекстно-свободных грамматик широко распространена в разработках естественных и формальных языков. Вышеописанные лингвистические правила также можно формально представить, применяя правила КС-грамматики. Ниже приведем список правил для казахского и русского языков:

S(каз) →	**NP VP**
NP →	N| NP NP | AdjP NP | NPPP |
VP →	V | VPVP | PPVP | AdvVP | VP Aux
Adj →	AdjP NP| AdjP AdjP |
AdvP →	Adv| AdvP AdvP | AdvP VP
PP →	NP PP

Таблица 1. **Для казахского языка**

S →	**NP VP| VP NP**
NP →	N| NP NP | AdjP NP | NPPP | NP S | S NP | PP NP | NP PartP | NP VPINF |
VP →	V | VPVP | VPNP | VPPP | AdvPVP|VPS |VP AdvP | AdvPartP VP| VP AdvPartP |
AdjP →	Adj| AdvPAdjP| AdjP NP| AdjP PP|
AdvP →	Adv| AdvP AdvP | AdvP NP | AdvP VP
PP →	P | PP NP

Таблица 2. **Для русского языка**

S – Нетерминал, предложение; V – Глагол; N – Существительное; Adj – Прилагательное; Adv – Наречие; PP – Предложная группа, послелоги; Aux – Вспомогательные глаголы.

3. Синтаксические трансформации при переводе с казахского на русский

В последние годы меняются разные подходы и методы построения моделей для машинного перевода, где начинают внедрять лингвистические правила. В статистические методы встраиваются грамматические особенности, а в правиловый метод включается статистика. Одним из основных этапов перевода – трансформация, представляет собой замена слов и словосочетаний исходного текста на эквиваленты выходного языка и преобразование информации о синтаксической структуре исходного языка в информацию. А моделирование языковых трансформаций – явление только начинающее практиковаться в системе машинного перевода.

Грамматические трансформации обуславливаются различными причинами – как чисто грамматического, так и лексического характера, хотя основную роль играют грамматические факторы, т.е. различия в строе языков. В нашем исследовании необходимо будет отразить наиболее распространенные стандартные синтаксические конструкции. В казахско-русском процессе перевода трансформации могут происходить на разных уровнях и этапах перевода: синтаксические трансформации, перестановки от леоветвящей структуры к правому ветвлению, переход от пассива к активу, переход от двойного маркирования казахского к зависимостному на примере изафетов, замена типа синтаксической связи.

В нашем исследовании основные тезисы трансформации касаются локальных перестановок, т.к. – для фразеологического метода важнейшей языковой единицей является словосочетание. Однако в дальнейшем, получив полный набор всевозможных локальных трансформаций для структуры казахского языка в переводе на русский, мы можем получить правила и закономерность для глобальных трансформаций.

В тюркских языках присутствует класс глаголов, имеющих одновременно два варианта употребления. В одном из них такие глаголы идентичны обычным лексическим глаголам, в другом они выполняют грамматическую функцию подобно вспомогательным глаголам *быть* и *иметь* в европейских языках. Такие глаголы в тюркологической литературе чаще всего называются вспомогательными, их несколько десятков. Например, существуют четыре вспомогательные глаголы, выражающие настроящее время или продолжительность действия, встречаются во всех языках тюркской группы.

«Отыр» указывает на действие, происходящее в данный момент речи.

Арман кітап **оқып отыр**. Арман **читает** книгу.

«Жатыр» придает действию постоянный характер и указывает на длительность:

Атам ұйықтап жатыр. Дедушка спит.

«Жүр» указывает на продолжительность действий и их повторяемость:

Балалар далада **ойнап жүр**. Дети играют во дворе.

«Тұр» придает действию характер повторения, указывая на то, что действие происходит в данный момент речи:

Бөлмеде радио **қосылып тұр**. В комнате **включено** радио[1]

Для сравнения, *стоять*, др.-тюрк. – *tur;* казахский – *тұр;* киргизский – *тур;* татарский – *тор;* кумыкский – *тур*; карачаево-балкарский – *тур.*[2]

Кроме 4 основных часто встречаемых существует множество других вспомогательных глаголов, каждый имеющий свою функциональность: *алу* брать и *беру* давать. Эти глаголы передают грамматическое значение аппликатива, вводя в предложение нового участника, являющегося бенефициантом описываемой ситуации. При этом *ал* делает бенефициантом субъект основного глагола (сделать нечто для себя), *бер* некоторое третье лицо (сделать нечто для кого-то другого)

Ол кітап ***сатып алды*** *Он* ***купил*** *книгу*
Он книга.ACC ***покупать*** *CONV* ***брать*** *PST*
Ол кітабімді **сатып берді** Он [помог] продать мою книгу
Үй тапсырмасын **жасап берші**, өтінемін!
Сделай мне домашнее задание, пожалуйста!

В разных языках в разной форме. Например, в татарском: *укып калу – сумел прочитать.* В казахском они не грамматикализовались и выделены как отдельные глаголы: *ауырып қалу – заболеть*, *ұмытып қалу – забывать*, *ойланып қалу – думать.* Далее эти вспомогательные глаголы по аналогии присоединяются к главному и находится в постпозиции:

- кел – приходить
- жібер
- бар – идти
- кет – уходить
- кір – входить
- өт – проходить
- түс – спускаться
- біл – знать
- көр – пробовать, пытаться; дәмін татып қарау – попробовать на вкус

---

[1] Оралбаева Н. 1971. Аналитические формы глагола в современном казахском языке. Алма-Ата.

[2] Гращенков П.В.«Тюркские конструкции со вспомогательным глаголом и деепричастием на -п» [URL: http://pandia.ru/text/78/153/53477.php]

• түс – спускаться, падать
• қара – смотреть\ присматривать

Итак, при обработке казахского текста формальное выделение глаголов, состоящих из двух-трех слов, позволит на выходе выбрать правильный эквивалент глагола на русском. Для более общего моделирования можно будет структурировать эти составные глаголы как одну фразовую структуру на выходном языке. Здесь нет необходимости для локальных перестановок, в фразеологический словарь нужно внести весь список глаголов со вспомогательными элементами и тем самым расширить словарь. В рамках системы Метафраз это делается своим особым метаязыком с помощью грамматической таблицы с мнемоническими обозначениями. Синтаксическая структура выходного текста в значительной мере определяется синтаксической структурой фразеологических словосочетаний, выбранных из словаря. Именно поэтому подобные фразы – составные глаголы, будут удобны в рамках фразеологического метода. В качестве фразеологических элементов данные сочетания вместе с вспомогательными глаголами будут покрываться словарными фразеологическими словосочетаниями, представляя переводными эквивалентами из одного глагола русского языка: *ұмытып қалу – забывать.*

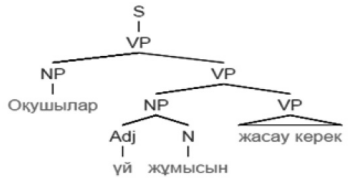
Данное исследование не приводит полный список всех возможных конструкций казахского языка, однако в дальнейшем планируется расширять его. Приведем некоторые примеры: конструкции со словом *керек*/*тиіс* интерпретируются в зависимости от расположения по отношению к главному слову. Если следует *керек* перед существительным, то оно выполняет роль прилагательного. *Керек заттар – нужные вещи.*

Телефон кітапшасынан ***керек нөмірді*** *жаздым*
Я записала ***нужные номера*** *из телефонной книжки*

Если слово *керек* встречается после сказуемого-глагола, то оно является его частью, составляя глагольное словосочетание. *Білу керек – необходимо знать.* При трансформации глагольное словосочетание *VP + керек/тиіс* меняется местами и на выходе имеет противоположный порядок слов. *Нужно/необходимо + VP*

Оқушылар үй жұмысын ***жасау керек.*** *Школьники должны делать домашнее задание*

[S [VP [NP Оқушылар] [VP[NP [Adj үй] [N жұмысын]] [^VP жасау керек]]]



Вот так выглядит структура выходного предложения **без** локальной трансформации:

[S [VP [NP Школьники] [VP[NP [Adj домашнее] [N задание]] [^VP должны делать]]]

Как известно, в рамках автоматизированного перевода должен производиться анализ смысловой структуры исходного текста, заключающийся в трансформации исходного представления текста в его формализованное представление. Затем происходит соотнесение выявленных наименований понятий и их взаимосвязей на исходном языке с системой понятий на выходном языке, и далее построение формализованного представления смысловой структуры текста на выходном языке и локальная перестановка глагольного словосочетания. Итак, с учетом лингвистического правила в процессе синтеза мы получаем правильный вариант трансформации:

[S [VP [NP Школьники] [VP [^VP должны делать][NP [Adj домашнее] [N задание]]]

По аналогии можно эту же модель применять и для сложных предложений и для сложносочиненных предложений. Для слож-

неподчиненных предложений казахского языка необходимо будут дополнительные правила по распознаванию деепричастий и причастий в конце придаточного предложения в составе сложного. Ниже следует краткий набор синтаксических конструкций и грамматических особенностей казахского языка, которые могут составить сочетания слов – фразы. Для их формализации можно применить одно из вышеописанных правил:

• порядок слов в предложении казахского языка: предикат в конце, определение перед определяемым словом

• порядок слов внутри словосочетаний, изафеты

• различие прилагательных от наречий по расположению(как в рус.)

Айнала ***үнсіз(Adj)*** *дала. Кругом* ***безмолвная*** *степь* (Г.Мустафин).

Жауынгер ***үнсіз(Adv)*** *қайтып кетті. Солдат* ***молча*** *ушел*

• конструкции со вспомогательными глаголами (*«сатып алу» – купить*)

• наклонения (условное / повелительное / желательное) *Мен Астанаға* ***барғым келеді*** *– Я хочу поехать в Астану*

• послелоги

• сочетания глаголов со словами *«керек, қажет» – нужно*

• союзы

• Если перед существительным стоит числительное, то окончание множественного числа в существительном не употребляется. Числительные и прилагательные в функции определения перед существительными не изменяются ни в числе, ни в падеже.

• Причастие, стоящее перед определяемым существительным, не изменяется ни по падежам, ни по числам в отличие от русского языка, где причастие согласуется с определяемым словом в роде, числе, падеже. Имена числительные (количественные и порядковые), употребляясь в предложении в функции определения, не изменяются в числе и в падеже.

• Генитив предшествует имени существительному

• относительное предложение предшествует существительному, в русс. *Дом, который...*

В данной работе мы выявили типы трансформаций, а также было установлено, что для наилучшей передачи смысла одного из компонентов предложения, довольно часто применяется не одна трансформация, а последовательность из нескольких трансформа-

ций. Таким образом, можем выявить, что основные виды трансформации в переводе с казахского языка на русский это:

– субъектно-предикатные трансформации

– локальные перестановки глагольных словосочетаний, в том числе конструкции со вспомогательными глаголами;

– замена составных глаголов, вспомогательных глаголов в один переводной эквивалент.

Заключение

В рамках выполнения данной работы были разработаны общие принципы построения и формализации машинной грамматики казахского языка на основе лингвистических знаний и формальной грамматики. При разработке этих принципов была проанализирована грамматика казахского языка, написаны правила по системе Контекстно-свободной грамматики с учетом особенностей синтаксиса казахского языка. Также выявлены распространенные стандартные синтаксические конструкции

Приведены и визуализированы способы синтаксических трансформаций при переводе. Однако, в дальнейшем можно расширить исследование, анализируя на более объемном корпусе текстов и собрать конечный список синтаксических единиц.

ЛИТЕРАТУРА

1. Хорошилов А.А. Теоретические основы и методы построения систем фразеологического машинного перевода// докт.диссер./ М., 2006., С. 251.

2. Elena Kozerenko, Alexander Khoroshilov, Alexei A. Khoroshilov Syntactic Parameters in the Phrasal Machine Translation // Proceedings of ICAI'13, WORLDCOMP'13, July 22–25, 2013, Las Vegas, Nevada, USA – CRSEA Press, USA, 2013, Vol.II. P. 890–895.

3. Хорошилов Ал-др А. и др. Системы фразеологического машинного перевода текстов с русского языка на английский и с английского на русский// Белоногов Г.Г., Зеленков Ю.Г., Кузнецов Б.А., Новоселов А.П., Хорошилов Ал-др А., Хорошилов Ал-др А. и др.

4. Системы фразеологического машинного перевода – технология XXI века/ Белоногов Г.Г., Зеленков Ю.Г., Новоселов А.П., Хорошилов Ал-др А., Хорошилов Ал-сей А. // Материалы междунар. конф. НТИ-97

«Информационные ресурсы, интеграция, технологии». – М.: ВИНИТИ, 1997. – С.42–43 (0,2 п.л).

5. Дж. Гринберг. «Некоторые грамматические универсалии, преимущественно касающиеся порядка значимых элементов». Новое в лингвистике. Вып. 5., М., 1970. С. 114–162.

6. Оралбаева Н. 1971. Аналитические формы глагола в современном казахском языке. Алма-Ата.

7. Гращенков П.В.«Тюркские конструкции со вспомогательным глаголом и деепричастием на -п» [URL: http://pandia.ru/text/78/153/53477.php]

УДК 81’32

ON VARIOUS APPROACHES TO MACHINE TRANSLATION FROM RUSSIAN TO KAZAKH

A. Makazhanov[1], B. Myrzakhmetov[1,2], Zh. Kozhirbayev[1,3],
[1] National Laboratory Astana, 53 Kabanbay Batyr ave., Astana, 010000, Kazakhstan
[2] Nazarbayev University, SST, 53 Kabanbay Batyr ave., Astana, 010000, Kazakhstan
[3] L.N. Gumilyov Eurasian National University, Faculty of Information Technologies, 2 Satpayev str., Astana, 010008, Kazakhstan
bagdat.myrzakhmetov@nu.edu.kz

In this work we compare a number of approaches to machine translation (MT) form Russian to Kazakh. We focus specifically on this pair of languages for a number of reasons. First, these languages are relatively understudied in terms of MT research, as well as, natural language processing (NLP) research in general. Kazakh, in particular, has been actively studied with modern methods for less than a decade. Second, this pair of languages poses several processing challenges rooted in their nature: both languages are morphologically complex and tend to have free order constituents, which makes long term dependencies rather frequent. From the perspective of data-driven approaches to NLP that means increased data sparseness and high OOV rates. Lastly, apart from scientific curiosity there is a strong practical demand for high quality MT between the languages in question. Kazakh is the state language of Kazakhstan, while Russian, due to a strong Soviet heritage, largely remains a language of professional communication and conduct. This frequently results in paperwork being initially prepared in Russian and then translated into Kazakh. Thus, high quality MT systems are in demand as they would greatly reduce manual labor of the professional translators.

We categorize the approaches that we compare into data-driven, linguistically motivated and hybrid ones. In the first category we compare a phrase-based statistical MT (SMT) and a neural MT (NMT) approaches. For the latter we experiment with three different neural architectures. As the result of this comparison we conclude that while NMT is a promising research direction one needs a lot more computational resources and, perhaps, even more data to achieve the level of accuracy offered by SMT. As for linguistically motivated and hybrid approaches we compare a rule-based approach with a so called factored model, which is essentially an SMT model that takes into account various linguistic factors, such as parts of speech, lemmata, morphology, etc.

Although this comparison has shown that factored models should be strongly favored, we must note that the Russian-Kazakh pair for the rule-based system that was used in the experiment is still a work in progress. Lastly, one final comparison between the best performing models from each category, i.e. a pure data-driven SMT-model and a hybrid factored model, has favored the former.

While we acknowledge that the present work makes no significant contribution to the NLP research in general, we want to point out that, to the best of our knowledge, for the particular language pair considered herein experiments on NMT and factored SMT have never been performed before. We speculate that one possible reason for this is the absence of an accessible Russian-Kazakh parallel corpus that is suitable for those experiments in terms of both size and quality. With this in mind we also provide a detailed description of the parallel data set that we used for our experiments and which we plan to make available in the future.

Keywords: Machine translation; RBMT; NMT; SMT; factored SMT.

О РАЗЛИЧНЫХ ПОДХОДАХ К МАШИННОМУ ПЕРЕВОДУ С РУССКОГО НА КАЗАХСКИЙ ЯЗЫК

А. Макажанов[1], Б. Мырзахметов[1,2], Ж. Кожирбаев[1,3]

[1] Национальная Лаборатория Астана, пр. Кабанбай батыра 53, Астана, 010000, Казахстан

[2] Назарбаев Университет, ШНТ, пр. Кабанбай батыра 53, Астана, 010000, Казахстан

[3] Евразийский Национальный Университет им Л.Н. Гумилева. Факультет информационных технологий, ул. Сатпаева 2, Астана, 010008, Казахстан

bagdat.myrzakhmetov@nu.edu.kz

В данной работе мы сравниваем между собой различные подходы к машинному переводу (МП) с русского на казахский язык. Мы рассматриваем именно эту пару языков по ряду причин. Во-первых, эти языки относительно мало изучены как с точки зрения исследований в области обработки естественного языка (ОЕЯ) в целом, так и исследований в области МП в частности. Во-вторых, автоматическая обработка этих языков сопровождается рядом трудностей, обусловленных природой оных: оба языка имеют сложную морфологию и относительно свободный порядок слов, что делает довольно частым наличие удаленных связей. Такое поведение языков влечет за собой повышенную степень разрежённости данных и высокую частоту OOV слов – «кошмар» основанных на данных подходов к ОЕЯ. И наконец, кроме научного любопытства качественный МП между данными

языками имеет высокую практическую значимость и соответствующий спрос. Казахский язык является государственным языком Республики Казахстан, в то время как русский, в силу сильного Советского наследия, в большинстве случаев остается языком профессионального общения и деятельности. Такое положение довольно часто приводит к тому что, официальные документы сначала изготавливаются на русском, а затем переводятся на казахский язык. Таким образом, существует спрос на высококачественные системы МП между этими языками, так как такие системы смогут сильно облегчить труд профессиональных переводчиков.

Мы различаем следующие подходы к МП: основанные на анализе данных, лингвистически мотивированные, и гибридные. Среди подходов, основанных на анализе данных, мы сравниваем фразовый статистический МП (СМП) и нейронный МП (НМП). В случае с НМП мы экспериментируем с тремя разными архитектурами нейронных сетей. По итогам сравнения подходов, основанных на анализе данных, мы пришли к выводу, что хотя НМП является многообещающим направлением исследований, для достижения точности лучших СМП систем может понадобиться гораздо больше вычислительных ресурсов и, возможно, больше обучающих данных. Для сравнения лингвистически мотивированных и гибридных подходов, мы экспериментируем с системой, основанной на грамматических правилах, и так называемой факторной моделью, являющейся модификацией СМП, которая берет в расчет различные лингвистические факторы, такие как части речи, словарные формы, морфологию и т.д. Хотя результат данного сравнения показал, что гибридный подход (факторная модель) гораздо предпочтительней лингвистически мотивированного подхода, следует отметить, что работа над русско-казахской парой в системе, основанной на грамматических правилах, использованной в наших экспериментах, еще ведется и далека от совершенства. Наконец, окончательную точку в наших экспериментах поставило сравнение лучших подходов в своих категориях, т.е. СМП (подход, основанный на анализе данных) и факторной модели (гибридный подход). Данное сравнение показало, что при, тех ресурсах, которыми мы располагаем, статистический машинный перевод является наиболее точной моделью МП для перевода с русского на казахский язык.

Мы признаем, что настоящая работа не делает значимого вклада в исследования в области ОЕЯ в целом, однако следует отметить, что для конкретной пары языков, рассмотренной в этой работе, эксперименты по нейронному машинному переводу и по использованию факторной модели МП проводятся впервые. Мы полагаем, что одной из возможных причин этого является отсутствие русско-казахского параллельного корпуса, подходящего для подобных экспериментов, как по размеру, так и по качеству. С учетом этого, мы также подробно описываем параллельный корпус, который мы использовали в своих экспериментах и планируем сделать доступным в будущем.

Ключевые слова: Машинный перевод; машинный перевод на основе грамматических правил; нейронный машинный перевод; статистический машинный перевод; факторная модель перевода.

1. Introduction

We roughly categorize approaches to MT into data-driven, linguistically motivated, and hybrid ones. Data-driven approaches require a training procedure, which, in turn, requires special language resources, so called parallel corpora, i.e. sets of sentence pairs, where sentences in each pair are translations of each other. A classic example of a data-driven approach is statistical MT, where one has to train so called translation and language models (Koehn, 2010). In SMT training these models and maximizing the product of their outputs for a given source sentence and a set of candidate translations, i.e. decoding procedure, are performed separately and require several tools to be used. Such a modular approach became a source of criticism with the development of another popular data-driven approach neural MT. Most of the existing NMT systems encode source sentences into fixed-length vectors, which are then subjected to a series of non-linear transformations, and decoded into target sentences (Bahdanau et al., 2014; Cho et al., 2014). In contrast to SMT, all of this is done in a single framework. Such convenience however comes at a cost, as to be effective NMT systems usually require longer training time and more computational resources than SMT systems. At any rate, SMT and NMT both share a common characteristics in that exercising them requires little to none linguistic background.

Currently the bulk of research in MT focuses on data-driven approaches. The early works on MT, however, attempted another approach, where after a direct word to word translation words in target languages were put in a desired order. This approach was further developed to account for grammar of both source and target language, and translation was approached as a transfer of structure and meaning with the help of hand-crafted rules (Arnold, 1994). This linguistically motivated approach is commonly referred to as rule-based machine translation (RBMT). Compared to systems based on data-driven approaches, systems based on RBMT have the advantage of being less computationally prohibitive and easier for debugging. However, when it comes to development of such systems one needs to have a solid

linguistic background and know the grammars of languages involved in translation.

Lastly there are hybrid approaches that attempt to advantage from both the power of data and statistical methods, as well as from linguistic insight. A good example of such an approach is a so called factored SMT model proposed by Koehn and Hoang (2007). This approach incorporates grammatical information into a statistical model in a form of any number of factors, e.g. lemmata (dictionary forms), part of speech tags, morphology, etc. The model was shown by the authors to be superior to SMT ("pure" data driven), especially when languages involved in translation exhibit various degrees of morphological complexity. This approach, however, is harder to implement than SMT, because not only it needs a parallel corpus, which is for many languages a rare commodity, it also requires this corpus to be enriched with linguistics annotation, which is equally (if not more) difficult to obtain.

In this work we compare all of the aforementioned MT approaches on the task of translating from Russian to Kazakh language. For the comparison we use free open-source tools: Moses open SMT toolkit (Koehn et al., 2007) for building SMT and factored SMT models; Tensorflow (Abadi, et al. 2016) for building NMT models; and Apertium platform (Forcada et al., 2011) for testing an existing RBMT system. In each case we utilize those tools to build, or use as is, the most basic models or models with modifications proposed by others (Koehn and Hoang, 2007; Luong et al., 2015). In this sense the present work does not contribute theoretical novelty. If anything, it is an exploratory study, whose goal is to point out future research directions in the framework of a larger project that, among others, has a goal of building an MT system for translating between Russian and Kazakh languages.

In order to make the comparison fair (more or less) we evaluate all of the approaches in the same fashion, namely in terms BLEU (Papineni et al., 2002), a de facto metric for automatic evaluation of MT systems. To this end we prepare a test set of 1500 pairs of Russian and Kazakh sentences, where sentences in each pair are mutual translations of each other. Similarly, when comparing data-driven approaches we use the same training set. A detailed description of the data set used in this work is given in the next section. Here we just would like to point out that in the future we plan to release the data set after proper formatting, documentation, and, possibly, additional extension.

2. Data set

Our data set contains 890K+ parallel sentences extracted from online news articles across 15 websites, a complete list of which is given in Table 1. The chosen websites all belong to state bodies, national companies, and other quasi-governmental establishments, who adhere to the same language policy of making multilingual releases in Kazakh and Russian languages. Moreover, such websites almost always provide page-level alignment, e.g. a Russian version of a page contains a direct link to a Kazakh version of the same page and vice versa. Thus, we solve the document alignment at the stage of crawling.

We extract parallel pages form the aforementioned websites with our own Python-based implementations of site-specific crawlers, which extract only specific content ignoring boilerplates and other irrelevant information. Once we obtain parallel documents, we create a parallel corpus in three stages: alignment, cleaning, and filtering. We begin by lemmatizing both source and target documents (storing the originals) as this was shown to improve the accuracy of sentence alignment (Assylbekov et al., 2016a). We then proceed to align sentences in the lemmatized versions of the documents, and restore the originals via the ladder-style output of Hunalign alignment tool (Varga et al., 2007). The "Aligned" column of Table 1 gives the number of sentence pairs aligned by this initial step. After the initial alignment we perform a cleaning procedure, which boils down to removal of: (i) duplicate sentences; (ii) sentence pairs, where each of a pair is identical to the other; (iii) sentence pairs, where at least one of the pair does not contain a single alphabetic character; (iv) sentences longer than 50 and shorter than 3 alphanumeric tokens. As shown in Table 1 (the "Cleaned" column) this way we remove over 70K noisy sentence pairs. Such a cleaning procedure removes only blatant misalignments, however there are many aligned pairs that, although pass the cleaning test, are not in fact valid translations of each other. In order to filter out such cases we again follow Assylbekov et al. (2016a) and train a Random Forrest classifier on a manually labeled data set of 1600 sentence pairs. The classifier uses a combination of 36 features and for a given pair of sentences estimates the probability of the pair being parallel. We fed the aligned and cleaned sentence pairs to the classifier and created the list of pairs ranked by the classifier's estimations. We then removed the pairs for which the estimation was given as 0, thereby creating the

final data set with a total of 893234 parallel sentences (cf. “Filtered” column of Table 1).

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

Website	Aligned	Cleaned	Filtered	Test set	Tuning set
www.akorda.kz	80 333	75 240	75 199	150	100
www.primeminister.kz	111 193	81 060	79 483	140	95
www.ortcom.kz	77 468	73 770	73 610	130	100
www.nurotan.kz	63 268	57 043	56 563	90	60
www.astana.gov.kz	105 929	91 010	90 762	150	100
www.strategy2050.kz	372 249	347 560	345 372	495	310
www.adilet.gov.kz	59 489	30 083	28 744	75	50
www.economy.gov.kz	15 179	11 417	11 398	20	15
www.dkz.mzsr.gov.kz	24 813	7 624	7 232	15	15
www.kaztag.kz	52 643	45 653	45 505	75	50
www.almaty.gov.kz	25 176	18 211	18 036	30	20
www.mfa.gov.kz	12 463	10 466	10 405	20	10
www.palata.kz	36 394	33 355	33 206	70	50
www.expo2017astana.com	7 244	6 027	5 963	10	5
www.emer.gov.kz	14 199	11 855	11 756	30	20
Total	1 069 078	909 189	893 234	1500	1000

After the parallel corpus has been assembled, we need to sample out a test set for evaluation purposes and a tuning set for setting parameters when working with data-driven approaches. Both of those sets should be of high quality and representative of the data they are drawn from. To ensure quality we sample only those sentence pairs that were ranked the highest (i.e. most probable to be parallel) at the filtering stage of building the parallel corpus. To ensure representativeness and balance, from each source website we draw sentence pairs in proportion to the ratio of that website’s data volume to the size of the entire corpus. As it can be seen from Table 1 (columns “Test set” and “tuning set”), this way we have selected 1500 and 1000 sentences for the test and tuning sets respectively.

3. Comparison of Data-driven Approaches

We now proceed with the comparison of statistical and neural MT models. For building the SMT model we use Moses open SMT toolkit (Koehn et al., 2007) with standard parameters. For building 3-gram language models we use KenLM (Heafield, 2011) with modified Kneser-Ney smoothing (Kneser and Ney, 1995).

Table 2. **Results of the automatic evaluation of the data-driven approaches**

Model	BLEU score
SMT	**34.15**
Basic NMT	3.9
Attention NMT, 2 layers	9.14
Attention NMT, 4 layers	11.0

After initial training of the SMT model we perform the tuning procedure on the held-out tuning set (cf. Table 1). To this end we use the minimum error rate training optimization (Och, 2003) algorithm.

When it comes to building the NMT models we experiment with three different architectures: (i) basic encoder-decoder recurrent neural net based on the long-short term memory (LSTM) architecture (Sutskever, 2014); and LSTM with the attention mechanism (Luong et al., 2015) with (ii) two and (iii) four layers. We build three NMT models based on the aforementioned architectures using Tensorflow (Abadi, et al. 2016) API. In all three cases we use vocabularies of 50000 most frequent (in the training set) Russian and Kazakh words, and train the models on Tesla k3025 GPU using a 20000 iterations limit with 128 hidden units, and 0.2 dropout probability.

We evaluate the SMT and NMT models automatically in terms of the BLEU metric (Papineni et al., 2002) on the test set of 1500 sentence pairs (cf. Table 1). The results of the evaluation are given in Table 2. As it can be seen the performance of the NMT models increases with the increase in the complexity of the neural architecture which they are based upon. Nevertheless the SMT model performs more than three times better than the best NMT model. We speculate that such a drastic underperformance may have been caused by the limitations we had to impose on the NMT models due to insufficient amount of computational resources. Thus, in order to train the models

in adequate time we had to limit the training procedure to 20000 iterations. We also had to limit the vocabularies to a rather modest size of 50000 words, which may not be enough for morphologically complex languages such as Kazakh and Russian. In the future we plan to further investigate the impact of those parameters as well as to experiment with neural architectures.

4. Comparison of Linguistically Motivated and Hybrid Approaches

For a rule-based MT system we use Apertium platform (Forcada et al., 2011) with the readily available Russian-Kazakh system (revision 82385). We use this software as is, adding only a small post processing step where we clean out the symbols used internally by Apertium, e.g. * (asterisk) preceding unanalyzed words.

Table 3. **Mapping POS-tags for factored SMT**

Grammatical category	Russian tag	Kazakh tag	Common tag
Noun (common, proper)	NOUN	NOUN	NOUN
Adjective	ADJF	ADJ	ADJ
Adjective, short	ADJS	–	ADJ
Comparative degree	COMP	ADJ/ADV	ADJ/ADV
Adverb	ADVB	ADVB	ADVB
Pronoun	NPRO	PRON	PRON
Imitative word	–	IMT	NOUN
Numeral	NUMR	NUM	NUMR
Verb, finite	VERB	VERB	VERB
Verb, infinite	INFN	–	VERB
Verb, auxiliary	–	AUX	VERB
Participle	PRTF	VERB	VERB
Participle, short	PRTS	VERB	VERB
Verbal adverb	GRND	VERB	VERB
Predicative	PRED	–	MOD
Modal word	–	MOD	MOD
Preposition	PREP	–	ADP

Postposition	–	ADP	ADP
Conjunct	CONJ	CONJ	CONJ
Particle	PRCL	PART	PRCL
Interjection	INTJ	INTJ	INTJ
Punctuation	PUNCT	PUNCT	PUNCT
Symbol	–	SYM	SYM
Unknown word	UNK	X	X

For implementing the hybrid approach we build a factored SMT model using Moses open SMT toolkit (Koehn et al., 2007) yet again. This time, however, we also need to carry out additional data processing to supply the model with linguistic factors. For the initial experiments we chose to supply lemmata and POS tags. To jointly lemmatize and tag Russian sentences we use pymorphy2 a Python-based morphological analyzer and generator for Russian and Ukrainian languages (Korobov, 2015). For the same purpose we apply to Kazakh sentences a data-driven morphological analyzer and disambiguator for Kazakh (Makhambetov et al., 2015).

Table 4. **Results of the automatic evaluation of linguistically motivated and hybrid approaches**

Model	BLEU score
RBMT	6.41
Factored SMT	21.77
SMT	**24.73**

In order to adhere to the same tokenization scheme and minimize the OOV rate we pre-tokenize both Russian and Kazakh sentences using our own tokenizer that identifies initials, numerals, dates, URLs, common abbreviations and named entities, and tags them appropriately. Last but not least we map the tagset used for tagging Russian and Kazakh into a single tagset. This mapping is given in Table 3.

To train the factored SMT model we sample 175440 sentence pairs from the training set, which is roughly 1/5 of the set. We do this for the same reason we limited our NMT models – insufficient computation resources. The result of the comparison of the RBMT and factored

SMT systems is given in Table 4. In order to complete the comparison across all of the approaches we report the result of the SMT system (pure data-driven approach) trained in equal conditions as the factored SMT (training set of 175440 sentences, no tuning). As in the previous case evaluation is performed automatically by comparing the outputs of the model to the same test set of 1500 sentence pairs (cf. Table 1) and computing BLEU scores.

As it can be seen the factored model outperforms the RBMT model, achieving the BLEU score of 21.77, more than three times that of the Apertium-based RBMT system. We must state here, however, that in the framework of the Apertium project the Russian-Kazakh pair is still at the development stage, and results of our experiments should not be generalized to rule-based approaches to MT from Russian to Kazakh. Nevertheless, a pure data-driven approach without any linguistic insight, i.e. the SMT model, outperforms the hybrid approach. The net improvement of the SMT model over the factored SMT model is 2.96 BLEU, which corresponds to a noticeable 13.6% relative improvement. Here we should note that the SMT model obviously benefits from the increase in the size of training data (compare "SMT", Table 2 and "SMT" Table 4). It is reasonable to expect that the performance of the factored model will also increase with the increase in the amount of training data, and whether that increase will be larger than that of the SMT model remains to be seen. Be that as it may, in the future we would like to continue our experiments with factored SMT models, and apart from enlarging the training set, we plan to consider additional factors, e.g. morphology, syntax, etc. We will also try to use more advanced tools for morphological processing, which we had certain technical difficulties with using in the present work, e.g. HFST-based analyzer (Washington et al., 2014) and LSTM-based tagger for Kazakh (Toleu et al., 2017).

5. Related Work

Machine translation from Russian to Kazakh languages has been studied mainly from the perspective of rule-based and statistical approaches. To our knowledge, the present work is the first one to experiment with the hybrid and neural models for Russian to Kazakh MT.

Most of the work on rule based-approaches has been focused on developing the Russian-Kazakh language pair for the Apertium plat-

form (Forcada et al., 2011). Implementing a language pair involves building a number of processing tools such as morphological transducers, taggers, etc. Washington et al. (2014) developed a free open source morphological transducer capable of analyzing and generating Kazakh word forms. The authors report 98.6% precision and 57.9% recall for the analysis. Later Assylbekov et al. (2016b) developed an HMM-based tagger which in combination with constraint grammar-based tagger achieved 90.7% disambiguation accuracy. Although, to our knowledge, currently this tagger is not used in the Apertium project, it is fully compatible with it and could be used in principle. Apart from the tools, one has to also define various transfer rules. Abduali et al. (2015) describe development of structural transfer rules and lexical selection. Rakhimova (2015) investigates the problem of semantic analysis and synthesis of pretext.

The work on statistical MT from Russian to Kazakh has started fairly recently. Myrzakhmetov et al., (2016) study the problem of document alignment, an initial step in building parallel corpora, which in the present work we skip due to the idiosyncratic properties of our data source (cf. Section 2). The authors compare rule- and learning-based methods, and conclude that, although having lower recall, the former should be favored due to a slightly higher precision and simplicity. Assylbekov et al. (2016a) experiment with various approaches to sentence alignment and present a machine learning classifier for estimating alignment accuracy. The classifier is tested on a manually aligned sample and achieves 94% F-score. Myrzakhmetov and Makazhanov (2017) report on the experiments on phrase-based Russian to Kazakh SMT. They compare a basic SMT system with another SMT system that was trained on data, where source (Russian) sentences were lemmatized. They also attempt cleaning the data by re-aligning the training set using Hunalign (Varga et al., 2007) in combination with a hand-crafted dictionary. The results show that lemmatization alone boosts the performance of the basic model by 6.3% and when it is combined with cleaning relative performance gain is 9.5%.

6. Conclusion and Future Work

We have experimented with three approaches to machine translation from Russian to Kazakh language. To evaluate a data-driven approach we have built and tested a phrase based SMT model and three

NMT models with various neural architectures. For a hybrid approach we have built a factored SMT model that enriches the basic statistical model with linguistic factors, such as lemmata, POS tags, etc. Lastly for a pure linguistically motivated approach we use an existent system based on Apertium platform (Forcada et al., 2011).

We have compared the aforementioned approaches in a unified manner of automatic evaluation on a common carefully prepared test set of 1500 sentence pairs. The comparison has shown that a data driven approach, specifically statistical MT, performs significantly better than others. We note, however, that other models may have had disadvantages of being undertrained or underdeveloped, and we plan to work on eliminating those disadvantages in the future.

Acknowledgements

We would like to thank Dr. Francis M. Tyers for his help with the experiment on translation with Apertium platform.

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

REFERENCES

1. Abadi, Martín, et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467.

2. Abduali Balzhan, Akhmadieva Zhadyra, Zholdybekova Saule, Tukeyev Ualsher, Rakhimova Diana. (2015). Study of the Problem of Creating Structural Transfer Rules and Lexical Selection for the Kazakh-Russian Machine Translation System on Apertium Platform. In TurkLang 2015, pp. 5–9.

3. Douglas Arnold, Lorna Balkan, Siety Meijer, R. Lee Humphreys, Louisa Sadler. (1994). Machine Translation: an Introductory Guide, London: NCC Blackwell.

4. Assylbekov Z., Myrzakhmetov B., and Makazhanov A. (2016a). Experiments with Russian to Kazakh Sentence Alignment. Izvestija KGTU im.I.Razzakova, vol. 38, no. 2, pp. 18–23.

5. Assylbekov, Z., Washington, J., Tyers, F., Nurkas, A., Sundetova, A., Karibayeva, A., Abduali, B. and Amirova, D. (2016b). A free/open-source hybrid morphological disambiguation tool for Kazakh. In TrucLing 2016, pp. 18–26.

6. Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. CoRR. abs/1409.0473, pp. 1–15.

7. Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, Yoshua Bengio. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. CoRR. abs/1409.1259, pp. 1–9.

8. Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G. and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. Machine Translation 24(1), pp. 1–18.

9. K. Heafield. (2011). KenLM: faster and smaller language model queries. In Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation, pp. 187–197.

10. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL, pp. 177–180.

11. Koehn, P. and Hoang, H. (2007). Factored Translation Models. In EMNLP-CoNLL, pp. 868–876.

12. Koehn, P. (2010). Statistical Machine Translation. Cambridge, UK: Cambridge University Press, 433 P.

13. Korobov, Mikhail. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. International Conference on Analysis of Images, Social Networks and Texts, pp. 320–332.

14. R. Kneser and H. Ney. (1995). Improved backing-off for m-gram language modeling. In Acoustics, Speech, and Signal Processing, vol. 1. IEEE, pp. 181–184.

15. Luong, Thang and Pham, Hieu and Manning, Christopher D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In EMNLP, pp. 1412–1421.

16. Makhambetov O., Makazhanov A., Sabyrgaliyev I., and Yessenbayev Z. (2015). Data-driven morphological analysis and disambiguation for Kazakh. In CICLing, pp. 151–163.

17. Myrzakhmetov B. and Makazhanov A. (2017). Initial Experiments on Russian to Kazakh SMT. Research in Computing Science, vol. 117, pp. 153–160.

18. Myrzakhmetov B., Sultangazina A., Makazhanov A. (2016). Identification of the parallel documents from multilingual news websites. In Proceedings of the 2016 10th IEEE International Conference on Application of Information and Communication Technologies, pp. 197–201.

19. F. J. Och. (2003). Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics – Volume 1, pp. 160–167.

20. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. (2002). Bleu: A method for automatic evaluation of machine translation. In ACL, pp. 311–318.

21. Rakhimova, D. (2015). Research of problem of the semantic analysis and synthesis of pretext in the Russian-Kazakh machine translation. In Proceedings of the International Conference on Turkic Languages Processing (TurkLang 2015), pp. 59–67.

22. Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. (2014). Sequence to sequence learning with neural networks." Advances in neural information processing systems, pp. 3104–3112.

23. Toleu, Alymzhan and Tolegen, Gulmira and Makazhanov, Aibek. (2017). Character-Aware Neural Morphological Disambiguation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, pp. 666–671.

24. Varga, D., Halacsy, P., Kornai, A., Nagy, V., Nemeth, L., Tron, V. (2007). Parallel corpora for medium density languages. Amsterdam Studies. The Theory And History Of Linguistic Science Series. 4.292, pp. 1–7.

25. Jonathan Washington, Ilnar Salimzyanov, and Francis Tyers. (2014). Finite-state morphological transducers for three Kypchak languages. In LREC, pp. 3378–3385.

УДК 004.89

BUILDING A BILINGUAL DICTIONARY OF POLYSEMOUS WORDS FOR THE MACHINE TRANSLATION OF THE KAZAKH LANGUAGE

U. Tukeyev ualsher.tukeyev@gmail.com,
Z. Zhumanov z.zhake@gmail.com,
A. Karibayeva a.s.karibayeva@gmail.co,
D. Amirova amirovatdina@gmail.com,
A. Sundetova sun27aida@gmail.com,
B. Abduali balzhanabdualy@gmail.com
Al-Farabi Kazakh National University, Almaty, Kazakhstan

In this article we consider the creation of linguistic resources for Kazakh-English and Kazakh-Russian language pairs to find the correct translation in a specific context. These language pairs represent an interesting task of translating ambiguous words in the text. To perform this task, it is important to build linguistic data for the Kazakh language. One of such construction methods is based on large amounts of annotated data. These data are the linguistic basis for machine translation systems, such data can be in form of dictionaries, corpora. Quality of machine translation depends on the volume of language database and on the depth of description of natural languages. Dictionaries have a great influence on work of machine translation systems. An important problem of a dictionary is its transparency and limitations; it can not be absolutely complete because the lexical composition of a language is always subject to change. Thus, the process of creating a dictionary becomes infinite, as each dictionary can improve daily.

Nowadays there are many machine translation systems for the language pair. But the lack of reliable mechanisms for determining the meaning of a word reduces the accuracy (quality) of the translation since a word can have not one but a whole series of meanings in another language. And automatically determining the correct translation of words that depend on the context is a very difficult task. The solution of lexical disambiguation problem for a long time was conceived as the main task, the completion of which will make it possible to achieve almost perfect machine translation.

The solution of the disambiguation problem is an important task of machine translation and requires a large number of linguistic data. In this problem, the meanings (possible translations) of ambiguous words are taken from a bilingual dictionary. A bilingual dictionary is a special dictionary used to translate words or phrases from the source language into the target language. In addition to translation, a bilingual dictionary usually defines part of speech and other char-

acteristics of words. Dictionaries are a fundamental element in the development of the linguistic base of natural languages. Linguistic resources, such as bilingual languages, are the basis for obtaining the correct translation in rule-based machine translation such as Apertium. The formation of linguistic resources for machine translation requires a wide and complete presentation of linguistic information about the natural language that needs to be processed. To this date, there are different types of machine translation, such as rule-based, statistical, hybrid, and others. In all systems of machine translation, the linguistic database of languages plays a very important role. Rule-based machine translation uses linguistic information contained in dictionaries and grammars of the source language and the target language. Statistical machine translation is based on a collection of aligned parallel corpora and includes calculation of this translation's probability and choice of the most probable one. The number of parallel corpora for the Kazakh language is small in comparison with the number of a monolingual text corpus. In any machine translation system, linguistic data is expanded through translations with different senses of the source language into the target language. Currently, one of the main problems of natural language processing is ambiguity. This paper will consider the creation of linguistic resources for solution of the ambiguous translation task in machine translation based on the rules in the free/open source Aperium platform and the results of extracting the necessary data, that is, alternative translations of ambiguous words from parallel natural language corpora based on a bilingual dictionary, will be demonstrated.

Keywords: machine translation; linguistic data; ambiguity; bilingual dictionary; corpus; Apertium.

ФОРМИРОВАНИЕ ДВУЯЗЫЧНОГО СЛОВАРЯ МНОГОЗНАЧНЫХ СЛОВ ДЛЯ МАШИННОГО ПЕРЕВОДА КАЗАХСКОГО ЯЗЫКА

У. А. Тукеев, ualsher.tukeyev@gmail.com,
Ж. М. Жуманов, z.zhake@gmail.com,
А. С. Карибаева, a.s.karibayeva@gmail.com,
Д. Т. Амирова, amirovatdina@gmail.com,
А. М. Сундетова, sun27aida@gmail.com,
Б. А. Абдуали, balzhanabdualy@gmail.com
Казахский национальный университет им. аль-Фараби, Алматы, Казахстан

В этой статье мы рассмотрим создание двуязычных словарей для казахско-английской и казахско-русской языковых пар на платформе Апертиум. Эти языковые пары представляют интересную задачу перевода многозначных слов в тексте. Существует множество разных методов

формирования лингвистических данных, используемые в машинном переводе. Один из таких методов построения основывается на больших аннотированных данных. Эти данные являются лингвистической базой для систем машинного перевода, такими данными могут быть словари, корпуса. Качество машинного перевода зависит от объема языковой базы данных и от глубины описания естественных языков. Словари оказывают большое влияние на работу систем машинного перевода. Важной проблемой словаря является его прозрачность и ограниченность; он не может быть абсолютно полным, потому что лексический состав языка всегда подлежит изменению. Таким образом, процесс создания словаря становится бесконечным, поскольку каждый словарь может улучшаться ежедневно.

В наше время существует множество систем машинного перевода для пары языков. Но отсутствие надежных механизмов определения значения слова снижает точность (качество) перевода, так как слово может иметь не одно, а целый ряд значений на другом языке. И автоматическое определение правильного перевода слов, зависящих от контекста – очень трудная задача. Разрешение лексической многозначности с давних пор задумывалось как главная задача, решение которой позволит добиться почти идеального машинного перевода, в котором двуязычный словарь играет важную роль.

Решение задачи многозначности – важная задача машинного перевода и требует большого количества лингвистических данных. В этой задаче смыслы (возможные переводы) многозначных слов берутся из двуязычного словаря. Двуязычный словарь – это специальный словарь, используемый для перевода слов или словосочетаний с исходного языка на целевой язык. В дополнение к переводу двуязычный словарь обычно определяет часть речи и другие характеристики слов.

На сегодняшний день существуют разные виды машинного перевода, такие как основанные на правилах, статистический, гибридный и другие. Во всех системах машинного перевода немало важную роль играет лингвистическая база данных языков.

Машинный перевод, основанный на правилах, использует лингвистическую информацию, содержащуюся в словарях и грамматиках исходного языка и целевого языка. Статистический машинный перевод основан на коллекции выровненных параллельных корпусов и включает в себя вычисление вероятности данного перевода и выбор наиболее вероятного. Количество параллельных корпусов для казахского языка невелико по сравнению с количеством одноязычного текстового корпуса. В любой системе машинного перевода лингвистические данные расширяются за счет многозначных переводов исходного языка на целевой.

Словари являются основополагающим элементом в разработке лингвистической базы естественных языков. Лингвистические ресурсы, такие как двуязычные языки, являются основой для получения правильного перевода в машинном переводе основанным на правилах, таких как Апертиум.

Платформа Апертиум насчитывает около 40 автоматических переводчиков и еще больше находится в процессе разработки. Команда Апертиум проявляет большой интерес к региональным языкам и активно поддерживает работу по созданию новых систем машинного перевода с лингвистической базой данных словарями и правилами, указанными в формате XML.

Формирование лингвистических ресурсов для машинного перевода требует широкого и наиболее полного представления лингвистической информации о естественном языке, который требуется обработать.

В настоящее время, одной из основных проблем обработки естественного языка является многозначность. В данной статье будет рассматриваться создание многозначных двуязычных словарей для решения задачи лексического выбора в машинном переводе, основанный на правилах в открытой/свободной платформе Апертиум и будут продемонстрированы результаты извлечения слов или лемм, то есть альтернативных переводов многозначных слов из параллельных корпусов естественного языка и двуязычного словаря.

Ключевые слова: машинный перевод; лингвистические данные; многозначность; двуязычный словарь; корпус; Апертиум;

1. Введение

Одной из основных проблем обработки естественного языка является многозначность, в которой немаловажную роль играет словари. Проблема разрешения многозначности как отдельная задача была сформулирована еще в конце 40-х годов XX века, практически одновременно с появлением машинного перевода. Начиная с того времени было разработано немало методов решения данной задачи, тем не менее она остается актуальной и по нынешний день.

Во всех развитых языках присутствуют как однозначные, так и многозначные слова. Способность слов выступать лишь в одном значении называется однозначностью или моносемией. Примеры таких слов: «бинокль», «троллейбус», «suitcase», «noun». Однако большинство слов имеют не одно, а несколько значений. Они называются многозначными или полисемантическими. Способность лексических единиц иметь несколько значений называется многозначностью или полисемией.

В данной статье будут показаны методы для формирования двуязычного словаря Апертиума, используемые при решении задачи лексического выбора при комбинированной технологии в

англо-казахском, казахско-английском, русско-казахском и казахско-русской языковой пары на платформе Апертиум.

Двуязычные словари многозначных слов являются частью общих двуязычных словарей языковых пар. Так как проблема многозначности слов существенно сказывается на качестве машинного перевода, вопрос полноты двуязычных словарей многозначных слов является очень актуальным.

В данной работе рассматривается комплексная технология формирования двуязычных словарей многозначных слов машинного перевода с использованием как существующих электронных двуязычных словарей, так и формирование двуязычных словарей многозначных слов через обработку параллельных двуязычных корпусов.

2. Обзор работ

Формирование лингвистических ресурсов имеет большое значение при приобретении знаний. Существуют некоторые подходы к извлечению лингвистических данных из корпусов и двуязычных словарей.

Многие подходы касаются автоматического приобретения лингвистических ресурсов для машинного перевода(МП). Самый ранний метод для МП просто включал двуязычный словарь, в котором каждое слово было просмотрено и перенесено на его эквивалент на целевом языке. Этот подход полезен для перевода фраз, но не для предложений (Victor M.Sanchez-Cartagena, Miquel Espla-Gomis и др.)

Во (Michael Rosner, Kurt Sultana, 2014) путем объединения двух методов для расширения двуязычного словаря авторы получают верные словарные записи. Они используют два метода для расширения словаря из сопоставимых корпусов. Первый метод – это вектор контекста, а второй – метод эскиза. Последний метод показывает более высокую точность, но он очень чувствителен к размеру корпуса. В то время как векторный метод контекста хорошо работает даже с небольшими корпусами по сравнению с методом эскиза и дает лучшие результаты.

Метод обогащения двуязычных окончательных переводов без каких-либо предварительных знаний представлен в (Ajay Dubey и др.). Авторы представили автоматический метод создания межъязычного словаря, который был сгенерирован небольшим количе-

ством текстов из средств массовой информации, при использовании сопоставимых пар.

Существует много методов автоматической генерации двуязычных словарей. Эти методы создают двуязычные словари, извлекая двуязычные словосочетания из корпусов.

Создание двуязычных словарей для определенных областей, определенных параллельным корпусом, описаны в (Angelina Ivanova и др). Авторы предлагают метод, который состоит из 4 шагов. Первый шаг – это выравнивание предложений, который устанавливает эквивалентности между предложениями параллельных текстов. Следующий шаг – морфологический анализ, собирающий морфологическую информацию для слов. Идентификация эквивалентности между одиночными и многословными выражениями в параллельных предложениях выполняется на третьем этапе – лексическое выравнивание. Четвертый шаг отвечает за выбор более релевантных записей для словаря ().

Авторы (L. Hilgert, L. Lopes и др., 2012) показали, что двуязычные словари также могут быть изучены или обогащены из неприсоединившихся сопоставимых корпусов. Они предлагают комбинированные лингвистические и статистические методы. В статье описана двухэтапная модель перевода

Двуязычной терминологии из сопоставимых корпусов, неоднозначности и выбора лучших вариантов перевода на основе их морфологических знаний.

Турдаков Д. Ю. (2010) использовал названия статей, которые используются для поиска терминов в тексте для создания словаря. После всех терминов, найденных в тексте, они представлены в виде серии наблюдений и значений, соответствующих состояниям в скрытом марковской модели. Для создания словаря терминов они использовали имена всех статей, описывающих соответствующие понятия и имена всех страниц, перенаправляемых к статье. В статье будут рассмотрены методы по формированию списка многозначных слов по частотности по корпусу.

3. Описание методов формирование двуязычного словаря

Технология формирования лингвистических данных подразумевает несколько этапов создания данных со словарей и корпусов.

Для формирования многозначного словаря используются следующие данные:

• Двуязычный словарь данной языковой пары(apertium-eng-kaz.eng-kaz.dix);

Для других языковых пар используется такой же двуязычный словарь в формате .dix. Эти файлы написаны на языке XML, форматы заполнения записей приведены ниже.

• Параллельные корпуса исходного и целевого языка;

Параллельные корпуса для казахского языка собирались из разных источников, таких электронные сайты правительства Республики Казахстан, из художественной литературы, библии и т.д. На сегодняшний день имеется около 26000 предложении в параллельных корпусах для англо-казахского(и обратно) языка, для русско-казахского(и обратно) первоначально было собрано около 2500 предложении.

Второй подход формирование многозначного словаря реализуется из существующего двуязычного словаря системы Апертиум.

Шаги для формирования многозначного словаря из двуязычного словаря(файл с расширением.dix) следующие:

• Пройтись по двуязычному словарю от начала до конца;

• Нахождение леммы исходного языка, которые имеют несколько переводов в целевом;

• Извлечение слова с лексической многозначностью перевода из словаря.

Для формирования из корпусов аналогично извлечению из словаря и корпусов, и включает следующие этапы:

• Сделать список частотных слов из одноязычного корпуса исходного языка;

• Выбор слова с множеством перевода из списка частот, используя двуязычный словарь;

• Создать многозначный словарь с использованием словаря;

• Найти перевод заданных многозначных слов в параллельных корпусах;

• Добавить перевод из целевого корпуса в словарь, при его отсутствии.

Задача перевода многозначного слова требует много ручной работы в корпусах и словарях. Все переводы с параллельного корпуса будут добавлены к двуязычной словарь с описанием частей речи.

При формировании частотного списка переводов корпуса были очищены от слов, не имеющих важность при переводе, такие как цифры, знаки препинания и т.д, то есть “стоп” слов. Список наиболее часто встречающихся форм слов был отсортирован и сокращен до основ леммы при помощи скриптов, в результате чего был составлен список наиболее часто встречающихся лемм. Лемма – это каноническая форма слова, слово без грамматической информации. Например, леммой слова “cats” является “cat”. В английском языке лемма рассматриваемого существительного как правило совпадает с его формой единственного числа. В русском и казахском языках лемма существительного имеет вид формы именительного падежа единственного числа. Для глаголов в английском языке лемма будет иметь вид инфинитива без “to” (или просто инфинитива в русском и казахском языках). Например, леммой слова “was” будет “be”, также как леммой слова “был” будет “быть”, или леммой слова “болды” будет “бол”.

Слова были добавлены в двуязычный словарь вручную, как описано ниже. Двуязычный словарь: содержит переводные соответствия слов и символов двух языков. Он называется так: **apertium-eng-kaz.eng-kaz.dix.**

Если слово имеет несколько переводов, то пишутся все альтернативные переводы:

```
<e><p><l>sister<s n=»n»/></l><r>сіңлі<s n=”n”/></r></p></e>
<e><p><l>sister<s n=”n”/></l><r>әпке<s n=”n”/></r></p></e>
<e><p><l>sister<s n=”n”/></l><r>қарындас<s n=”n”/></r></p></e>
```

В данном случае, слово «sister» на казахском языке имеет 3 перевода. Один из них выбирается позже, посредством лексической выборки. Лексический выбор – это нахождение правильного перевода по контексту, в котором слово исходного языка имеет несколько переводов, которые относятся к одной части речи. Важность задачи разрешения лексической многозначности все же очень велика. В том числе оно может быть полезным целому ряду приложений. Сформированный многозначный двуязычный словарь используется в разных сферах: для улучшения качества машинного перевода, повышения точности методов классификации и кластеризации текстов, в информационном поиске и других приложениях, а также мы используем этот словарь в решении задачи лексического выбора на основе комбинированной технологии, реализующая через правила и статистику.

3.1 Извлечение списка частотных слов с проверкой с помощью «традиционного» словаря

Сделав список частот из одноязычного корпуса, мы стали искать среди них неоднозначные слова. В частности, мы искали слова с множеством переводов, относящиеся к одной части речи. Все возможные переводы были рассмотрены из «традиционного» словаря. Под «традиционным» словарем мы подразумеваем онлайн-словарь с большим объемом данных, таких как lugat.kz, sozdik.kz и т. д. Существует не так много словарей, которые переводят с и на казахский язык с разными вариантами переводов для многозначных слов.

Lugat – это комплекс словарей для нескольких языков. В словарях для англо-казахских и казахско-английских языковых пар содержится 113 318 слов для каждого направления. Словари для казахско-русского и русско-казахского языков содержат 95 118 слов для каждого направления. Создик – казахско-русский и русско-казахский онлайн-словарь.

Следующий шаг после выбора многозначных слов и нахождения переводов – заполнение двуязычного словаря не имеющихся в двуязычном словаре.

Инструменты и ресурсы, которые использовались в работе:

• apertium-eng-kaz.eng-kaz.dix – двуязычный казахско-английский и англо-казахский словарь

• apertium-kaz-rus.kaz-rus.dix – двуязычный казахско-русский и русско-казахский словарь

• Intertext – программное обеспечение для управления параллельными текстами

• Инструменты Апертиума для морфологического анализа и лемматизации.

• Sublime Text – редактор с богатыми возможностями преобразования текста.

Для англо-казахских и казахско-русских языковых пар есть одноязычные и двуязычные словари на платформе Апертиум.

InterText использовался для работы с параллельными текстами на английском, казахском и русском языках. InterText – инструмент управления для выровненных параллельных текстов. Он помогает редактировать и управлять выравниваниями нескольких версий текста на нескольких языках на уровне предложений, а

также поддерживает пользовательские XML-документы и набор символов Юникод(http://wanthalf.saga.cz/intertext).

Инструмент в основном используется лингвистами и экспертами, которые занимаются работой с текстами, особенно с выравниванием текстов, с которыми трудно справиться без глубоких знаний и технических навыков.

4. Результаты

В статье описаны методы формирование многозначного двуязычного словаря необходимых для решения задачи лексической многозначности на свободной/открытой платформе Апертиум. Работа по созданию словаря необходима для реализации комбинированной технологии лексического выбора в Апертиуме, на стадии формирования семантического куба для каждого многозначного слова. Семантический куб – это многомерная таблица данных для каждого многозначного слова с частотой появления с определенным контекстом из параллельного корпуса. Размеры словаря для англо-казахских (и наоборот) и русско-казахских составляют ~ 21 165 и ~ 12 923 записей соответственно. Результаты формирования многозначного словаря исходных языков казахского, русского и английского языка показаны ниже в таблице 1.

Таблица 1. **Результаты по созданию списка многозначных слов из словарей и корпусов**

	Объем параллельного корпуса	Объем двуязычного словаря Апертиум	Количество полученных слов из словаря Апертиума	Количество полученных слов из параллельного корпуса
English-Kazakh	25 324	21 165	432	197
Kazakh-English	25 324	21 165	748	246
Russian-Kazakh	2 256	12 923	91	71
Kazakh-Russian	2 256	12 923	87	63

Данные таблицы 1 показывают формирование многозначного двуязычного словаря из параллельных корпусов и общего двуязычного словаря Апертиума. Количество полученных многозначных слов с двуязычного словаря больше чем количество изъятых из корпуса, так как в параллельном корпусе предложения из официальных сайтов государства, в которых многозначность перевода присутствует в очень малой степени. С увеличением записей в словаре будет повышаться качество и точность перевода.

5. Заключение

Формирование лингвистических данных, а именно многозначного словаря будет представлять интерес для лингвистов, специалистам по прикладной и вычислительной лингвистике, и другим специалистам, интересующихся увеличением качества машинного перевода, информационного поиска и других приложений.

В этой статье был представлен подход для извлечения многозначных слов из разных лингвистических ресурсов: параллельных корпусов и словаря. Полученные лингвистические данные могут использоваться в различных задачах обработки языка, таких как задача лексического выбора в машинном переводе. Формирование более полного двуязычного словаря многозначных слов позволит повысить качество систем машинного перевода казахско-английской и казахско-русской языковых пар. Несомненно, данная тема актуальна и требует продолжения исследований в будущем.

ЛИТЕРАТУРА

1. Victor M.Sanchez-Cartagena, Miquel Espla-Gomis, Juan Antonio Perez-Ortiz. Source-Language Dictionaries Help Non-Expert Users to Enlarge Target-Language Dic-tionaries for Machine Translation.

2. Michael Rosner, Kurt Sultana. Automatic Methods for the Extension of a Bilingual Dictionary using Comparable Corpo. LREC 2014 :3790-3797.

3. Ajay Dubey, Parth Guptay, Vasudeva Varma, Paolo Rossoy. Enrichment of Bilingual Dictionary through News Stream Data. LREC 2014 :3761-3765.

4. Khang Nhut Lam and Feras Al Tarouti and Jugal Kalita. Automatically Creating a Large Number of New Bilingual Dictionaries.

5. Angelina Ivanova. Charles University in Prague Faculty of

Mathematics and Physics M.Sc. Thesis. Creating a Bilingual Dictionary using Wikipedia.

6. L. Hilgert, L. Lopes, A. Freitas, R. Vieira, D. N. Hogetop, A. A. Vanin. Building Domain Specific Bilingual Dictionaries. LREC 2012 : 2772-2777.

7. Sadat, F.; Yoshikawa, M.; Uemura, S.: Learning Bilingual Translations from Comparable Corpora to Cross-Language Information Retrieval: Hybrid Statistics-based and Linguis-tics-based Approach. In In Proc. IRAL 2003, 2003.

8. http://wanthalf.saga.cz/intertext.

9. Vondřička, Pavel. Aligning parallel texts with InterText. LREC 2014:160-164.

10. Турдаков Д. Ю. Методы и программные средства разрешения лексической многозначности терминов на основе сетей документов. Москва, 2010. Thesis.

УДК 513.7

MATHEMATICAL MODELS OF ENGLISH LANGUAGE FOR SYSTEM OF MULTILINGUAL SITUATIONS OF MACHINE TRANSLATION

M. Khakimov
Natational university of Uzbekistan it. Mirzo Ulugbek
Tashkent, Republic Uzbekistan
muftah@mail.ru

In the given are resulted work mathematical models of words of sentences on types of English language for the machine translation system «Tarjimon - LMX». System «Tarjimon - LMX» is developed as multilingual on the basis of technology of the modeled computer translator developed for machine translation. Mathematical models are described by means of the source language, developed for mathematical modeling of natural languages. Formulas for calculation maximum and minimum quantities of words or sentences for each of the stated models are stated.

Keywords: mathematical model of a conclusion of words; mathematical model of a conclusion of sentences; English language; the expanded source language; a natural language; technology; the modeled computer translator; a multilingual situation; machine translation; word type; offer type; the machine translation system.

МАТЕМАТИЧЕСКИЕ МОДЕЛИ АНГЛИЙСКОГО ЯЗЫКА ДЛЯ СИСТЕМЫ МНОГОЯЗЫЧНЫХ СИТУАЦИЙ МАШИННОГО ПЕРЕВОДА

М.Х. Хакимов
Национальный Университет Узбекистана им. Мирзо Улугбека
г. Ташкент, Республика Узбекистан
muftah@mail.ru

В данной работе приведены математические модели слов предложений по типам английского языка для системы машинного перевода «Tarjimon - LMX». Система «Tarjimon - LMX» разрабатывается как многоязычная на основе технологии моделируемого компьютерного переводчика разработанной именно для машинного перевода. Математические модели описываются с помощью входного языка, разработанного для математического моделирования естественных языков. Изложены формулы для расчета мак-

симальных и минимальных количеств слов или предложений для каждого из изложенных моделей.

Ключевые слова: математическая модель вывода слов; математическая модель вывода предложений; английский язык; расширяемый входной язык; естественный язык; технология; моделируемый компьютерный переводчик; многоязычная ситуация; машинный перевод; тип слова; тип предложения; система машинного перевода.

Введение

Каждый естественный язык (ЕЯ) является сложной системой, состоящих математически неструктурированных и не формализованных составных частей. Однако проведенные исследования над ЕЯ показывают, что не структурированность и не формализованность ЕЯ, можно привести к структурированному и формализованному виду, используя линейную методологию – выявление состава слова и построением логико-лингвистических (семантических) моделей по типам слов и предложений, далее построением математических моделей с помощью входного языка [2]. Данную методологию можно определить как степень формализации языка. Степень формализации в свою очередь определяет степень формализации семантики ЕЯ и точность алгоритма. Поверхностное понимание степени формализации ЕЯ, что формализованный язык – абстрактная, полностью оторванная от содержания конструкции с простой логической структурой приводит к низкой технологии машинного перевода [1]. Формализация позволяет выделить различные части ЕЯ, исследовать динамику их связей и главным образом даст возможность описания семантической структуры. Все эти качества очень существенны, когда используется общее ядро системы, т.е. когда над всеми ЕЯ входящие в данную среду перевода применяется единый системный подход, независимо с какого на какой ЕЯ осуществляется перевод.

Так как, английский язык (АЯ) также предназначен для включения в систему машинного перевода «Tarjimon - LMX», то он должен быть исследован с точки зрения формализации на принципах моделируемой компьютерной технологии. Следовательно, необходимо построение математических моделей слов и предложений по типам АЯ.

1. Математические модели вывод слов по типам

1.1. Общая математическая модель вывода слов

Лексический анализ словообразования АЯ показывает [4], что словоформа на АЯ может состоять из: предлога, префикса, корня, аффикса образующего слова, аффикса образующего форму и аффикса изменяющего слова. Согласно данной структуры, а также на выведенных логико-лингвистических моделей [4] и семантических баз данных АЯ [3], общая математическая модель построения слов на АЯ будет выражена как:

$$L1_{h30}(D, T, K(C,P,G,M,N,F), A1, A2, A3) =$$
$$=\downarrow\$_{[i,1\text{-}52]}D_i \oplus \downarrow\$_{[j,1\text{-}96]}T_j \oplus \$_{[i1,1\text{-}h0]}K_{i1}(C,P,G,M,N,F) \oplus$$
$$\downarrow\$_{[j1,1\text{-}86]}A1_{j1} \oplus \$_{[j2,1\text{-}93]}A2_{j2} \oplus \downarrow\$_{[j3,1\text{-}2]}A3_{j3}$$

Здесь, подкоренным словом К может быть любая часть речи – существительное, прилагательное, глагол, местоимения, наречия и числительное. Присоединение к этим подкоренным словам предлога, префикса и различных аффиксов даст образование слов. Переменная h0 обозначает общее количество слов имеющих корень. h30 = {min(93*h0), max(79852032*h0)} – минимальное и максимальное количество слов, которые могут быть выведены на основе данной закономерности.

1.2. Математическая модель вывода имен существительных

При выводе имен существительных на АЯ подкоренным словом может являться существительное, прилагательное, глагол, местоимение и окончание «not». Математическая модель вывода имен существительных АЯ на основе тринадцати вариантов логико-лингвистических моделей [4] и семантических баз данных АЯ [3], после некоторых преобразований будет выражена как:

$$C_{h31}(K(C), C(S), C(A1), D, C(T), M, C, P, K(G), K(M)) = (\$_{[i,1\text{-}h1]}K(C_i) \oplus (\downarrow\$_{[j,1\text{-}77]}C(S_j) \mathbf{V} \$_{[j1,1\text{-}17]}C(A1_{j1})))$$

$$\mathbf{V}\ (\$_{[i,1\text{-}h1]}K(C_i) \oplus \$_{[j,1\text{-}h1]}K(C_j))\ \mathbf{V}\ (\$_{[i,1\text{-}h1]}K(C_i) \oplus \$_{[j,1\text{-}h1]}K(C_j) \oplus \$_{[j1,1\text{-}77]} C(S_{j1}))\ \mathbf{V}\ (\$_{[i,1\text{-}h1]}K(C_i) \oplus$$

$$\$_{[j1,1\text{-}52]}D_{j1} \oplus \$_{[i1,1\text{-}h1]}K(C_{i1}))\ \mathbf{V}\ (\$_{[i1,1\text{-}43]}C(T_{i1}) \oplus \$_{[i,1\text{-}h1]}K(C_i) \oplus \$_{[j,1\text{-}77]} C(S_j))\ \mathbf{V}\ (\$_{[j,1\text{-}h5]}M_j \oplus \$_{[i,1\text{-}h1]}C_i)\ \mathbf{V}$$

$$(\$_{[i,1\text{-}h2]}P_i \oplus \$_{[j,1\text{-}77]}C(S_j)) \mathbf{V} (\$_{[i1,1\text{-}h2]}P_{i1} \oplus \$_{[i,1\text{-}h1]}K(C_i) \mathbf{V} (\$_{[j,1\text{-}h3]} K(G_j) \oplus \$_{[i,1\text{-}h1]}K(C_i)) \mathbf{V} (\$_{[j,1\text{-}h3]}K(G_j) \oplus$$

$$\$_{[i,1\text{-}h2]}P_i) \mathbf{V} (\$_{[j,1\text{-}h3]}K(G_j) \oplus \$_{[j1,1\text{-}77]}C(S_{j1})) \mathbf{V} (\$_{[j,1\text{-}h3]}K(G_j) \oplus \$_{[j1,1\text{-}77]} C(S_{j1}) \oplus \downarrow\$_{[j2,1\text{-}77]}C(S_{j2})) \mathbf{V}$$

$$(\$_{[j,1\text{-}h3]}K(G_j) \oplus \$_{[j1,1\text{-}h5]}K(M_{j1}) \oplus \text{“not”})$$

где h31= $\overline{1,h20}$, i ≠ j, i≠i1, j ≠ j1, j1 ≠ j2 только для выражений находящихся внутри одной скобки, т.е. для данного варианта выводимого слова.

Переменная h1 – количество слов, когда основанием является существительное; h2 – количество слов, где основанием является прилагательное; h3 – количество слов, где основанием является глагол; h4 – количество слов, где основанием является числительное; h5 – количество слов, где основанием является местоимение; h6 – количество слов, где основанием является наречие. Переменные: h31 – порядок слова, h20 = {min(17*h1), max(4004*h1*h2*h3)} минимальное и максимальное количество слов, которые могут быть выведены на основе данной закономерности.

1.3. Математическая модель вывода прилагательных

При выводе имен прилагательных подкоренным словом может служить прилагательное, существительное, глагол, наречие, числительное и местоимение. Вывод прилагательного образуется при присоединении к этим подкоренным словам префикса, суффикса, причастий, предлога и окончаний «d», «ed». Основываясь на двадцати вариантах логико-лингвистических моделей вывода прилагательных [4] и семантических баз данных АЯ [3], после некоторых преобразований формируем следующую математическую модель:

$$P_{h32}(K(P), K(C), P(S), T2, D, P(T), K(G), K(N), T, P, K(F), K(M)) = (\$_{[i,1\text{-}h2]}\, K(P_i) \oplus \$_{[i,1\text{-}h1]}K(C_i) \oplus$$

$$\$_{[i1,1\text{-}48]}P(S_{i1})) \mathbf{V} (\$_{[i,1\text{-}h2]}\, K(P_i) \oplus \$_{[i1,1\text{-}h2]}\, K(P_{i1})) \mathbf{V} (\$_{[i,1\text{-}h2]}\, K(P_i) \oplus \$_{[j,1\text{-}3]} T2_j) \mathbf{V} (\$_{[i,1\text{-}h2]}\, K(P_i) \oplus \$_{[j,1\text{-}52]}D_j)$$

$$\mathbf{V} (\$_{[j,1\text{-}22]}P(T_j) \oplus \$_{[i,1\text{-}h2]}\, K(P_i)) \mathbf{V} (\$_{[j,1\text{-}h3]}K(G_j) \oplus \$_{[i,1\text{-}48]}P(S_i)) \mathbf{V} (\$_{[i,1\text{-}h3]} K(G_i) \oplus \$_{[j,1\text{-}52]}D_j) \mathbf{V} (\$_{[i,1\text{-}h1]}K(C_i)$$

$$\oplus\ \$_{[i1,1\text{-}48]}P(S_{i1}))\ \mathbf{V}\ (\$_{[i,1\text{-}h1]}K(C_i)\ \oplus\ \$_{[i1,1\text{-}h1]}K(C_{i1})\ \oplus\ \text{“d”})\ \mathbf{V}\ (\$_{[i,1\text{-}h1]}K(C_i)\ \oplus\ \$_{[i1,1\text{-}h2]}K(P_{i1}))\ \mathbf{V}\ (\$_{[i,1\text{-}h1]}K(C_i)$$

$$\oplus\ T2_1)\ \mathbf{V}\ (\$_{[i,1\text{-}h1]}K(C_i)\ \oplus\ \$_{[i1,1\text{-}h1]}K(C_{i1})\ \oplus\ \$_{[j,1\text{-}48]}P(S_j)\ \oplus\ \text{“d”})\ \mathbf{V}\ (\$_{[i,1\text{-}h6]}K(N_i) \oplus \$_{[j,1\text{-}3]}\ T2_j)\ \mathbf{V}\ (\$_{[i,1\text{-}h6]}K(N_i)$$

$$\oplus\ \$_{[j,1\text{-}h2]}P_j)\ \mathbf{V}\ (\$_{[i,1\text{-}h6]}K(N_i)\ \oplus\ \$_{[i1,1\text{-}h1]}K(C_{i1})\ \oplus\ \$_{[j,1\text{-}48]}P(S_j)\ \oplus\ \text{“d”})\ \mathbf{V}\ (\$_{[i,1\text{-}h6]}K(N_i)\ \oplus\ \$_{[j,1\text{-}52]}D_j)\ \mathbf{V}$$

$$(\$_{[i,1\text{-}h4]}K(F_i)\ \oplus\ \$_{[j,1\text{-}h1]}K(C_j)\ \oplus\ \text{“ed”})\ \mathbf{V}\ (\$_{[i,1\text{-}h5]}K(Mi)\ \oplus\ \$_{[j,1\text{-}48]}P(S_j))$$

где h32= $\overline{1,h21}$, i ≠ i1 только для выражений находящихся внутри одной скобки, т.е. для данного варианта выводимого слова.

Переменные: h32 – порядок слова, h21= {min(h1+h2), max(48*h1*h2)} минимальное и максимальное количество слов, которые могут быть выведены на основе данной закономерности.

1.4. Математическая модель вывода глагола

При построении глагола подкоренным словом может служить существительное, глагол и прилагательное. При присоединении к этим подкоренным словам префикса, суффикса, предлога и «to be» образуется глагол. Математическая модель вывода глагола на основе девяти вариантов логико-лингвистических моделей [4] и семантических баз данных АЯ [3], после некоторых преобразований будет выражена как:

$$G_{h33}(K(G), K(P), K(C), D, G(T), G(S)) = (\$_{[i1,1\text{-}h3]}K(G_{i1})\ \oplus\ \$_{[j,1\text{-}h2]}K(P_j))\ \mathbf{V}\ (\$_{[i,1\text{-}h3]}K(G_i)\ \oplus\ \$_{[j,1\text{-}h1]}K(C_j))\ \mathbf{V}$$

$$(\$_{[i,1\text{-}h3]}K(G_i)\ \oplus\ \$_{[j,1\text{-}52]}D_j\ \oplus\ \$_{[i1,1\text{-}h3]}K(G_{i1}))\ \mathbf{V}\ (\$_{[i,1\text{-}h3]}K(G_i)\ \oplus\ \$_{[i1,1\text{-}h3]}K(G_{i1}))\ \mathbf{V}\ (\$_{[j,1\text{-}27]}G(T_j) \oplus\ \$_{[i,1\text{-}h3]}K(G_i))$$

$$\mathbf{V}\ (\$_{[i,1\text{-}h1]}K(C_i)\ \oplus\ \$_{[i1,1\text{-}h1]}K(C_{i1}))\ \mathbf{V}\ (\$_{[i,1\text{-}h1]}K(C_i)\ \oplus\ \$_{[j,1\text{-}11]}G(S_j))\ \mathbf{V}\ (\text{“to be”}\ \oplus\ \$_{[j,1\text{-}h2]}K(P_j))\ \mathbf{V}\ (\$_{[i,1\text{-}h2]}K(P_i)\ \oplus\ \$_{[j,1\text{-}11]}G(S_j))$$

где h33=$\overline{1,h22}$, i ≠ i1 только для выражений находящихся внутри одной скобки, т.е. для данного варианта выводимого слова.

Переменные: h33 – порядок слова, h22= {min(27*h3), max(104*h3)} – минимальное и максимальное количество слов, которые могут быть выведены на основе данной закономерности.

1.5. Математическая модель вывода местоимения

При построении местоимения подкоренным словом могут служить местоимения, существительные, числительные и окончания – “ever”, ”of”, “self”, ”selves”. Они вместе образуют местоимение. Математическая модель вывода местоимений на основе семи вариантов логико-лингвистических моделей [4] и семантических баз данных АЯ [3], после некоторых преобразований будет выражена как:

$$M_{h34}(K(M), D, C, K(F)) = (\$_{[j,1\text{-}h5]}K(M_j) \oplus \$_{[i,1\text{-}52]}D_i) \mathbf{V} (\$_{j,\,[1\text{-}h5]}K(M_j) \oplus \$_{[i,1\text{-}h3]}C_i) \mathbf{V} (\$_{[j,1\text{-}h5]}K(M_j) \oplus \$_{[i,1\text{-}h4]}K(F_i) \mathbf{V} (\$_{[j,1\text{-}h5]}K(M_j) \oplus \text{“ever”}) \mathbf{V} (\$_{[j,1\text{-}h5]}K(M_j) \oplus \text{”of”}) \mathbf{V} (\$_{[j,1\text{-}h5]}K(M_j) \oplus \text{“self”}) \mathbf{V} (\$_{[j,1\text{-}h5]}K(M_j) \oplus \text{”selves”})$$

где $h34=\overline{1,h23}$.

Переменные: h34 – порядок слова, h23={min(52*h5), max(h3*h4*h5)} минимальное и максимальное количество слов, которые могут быть выведены на основе данной закономерности.

1.6. Математическая модель вывода наречия

При построении наречия подкоренным словом могут служить существительное и прилагательное. При присоединении к этим подкоренным словам суффикса и предлога образуется наречие. Математическая модель вывода наречий на основе четырех вариантов логико-лингвистических моделей [4] и семантических баз данных АЯ [3], после преобразований будет выражена как:

$$N_{h35}(K(P), N(S), K(C), N(D)) = (\$_{[j,1\text{-}h2]}K(P_j) \oplus \$_{[i,1\text{-}10]}N(S_i)) \mathbf{V} (\$_{[i,1\text{-}h1]}K(C_i) \oplus \$_{[j,1\text{-}10]}N(S_j)) \mathbf{V} (\downarrow\$_{[i1,1\text{-}18]}N(D_{i1}) \oplus \$_{[j,1\text{-}h2]}K(P_j) \oplus \$_{[i,1\text{-}10]}N(S_i)) \mathbf{V} (\downarrow\$_{[i1,1\text{-}18]}N(D_{i1}) \oplus \$_{[i,1\text{-}h1]}K(C_i) \oplus \$_{[j,1\text{-}10]}N(S_j))$$

где $h35 = \overline{1,h24}$.

Переменные: h35 – порядок слова, h24= {min(10*h2), max(180*h1)} минимальное и максимальное количество слов, которые могут быть выведены на основе данной закономерности.

1.7. Математическая модель вывода числительных

При построении числительного подкоренным словом служить само числительное. При присоединении к подкоренному слову суффикса, союза и числительного образуется новое числительное. Математическая модель вывода числительных на основе трех вариантов логико-лингвистических моделей [4] и семантических баз данных АЯ [3], после преобразований будет выражена как:

$$F_{h36}(K(F), F(S), Y) = (\$_{[j,1\text{-}h4]}K(F_j) \oplus \downarrow\$_{[i,1\text{-}4]}F(S_i)) \mathbf{V} (\$_{[j,1\text{-}h4]}K(F_j) \oplus \downarrow\$_{[i,1\text{-}4]}F(S_i) \oplus \$_{[j1,1\text{-}h4]}K(F_{j1})) \mathbf{V}$$
$$\$_{[i,1\text{-}h4]}K(F_i) \oplus \downarrow\$_{[j,1\text{-}3]}Y_j \oplus \$_{[i1,1\text{-}h4]}F(S_{i1}))$$

где h36 = $\overline{1,h25}$, i ≠ i1, j ≠ j1 только для выражений находящихся внутри одной скобки, т.е. для данного варианта выводимого слова.

Переменные: h36 порядок слова, h25 – минимальное и максимальное количество слов, которые могут быть выведены на основе данной закономерности. h25= {min(4*h4), max(8*h4)}.

2. Математические модели вывода предложений

2.1. Математическая модель вывода повествовательных предложений

Математическая модель вывода повествовательного предложения на АЯ на основе двенадцати вариантов логико-лингвистических моделей [4] и семантических баз данных АЯ [3], после преобразований будет выражена как:

$$E1(C, G, N, L1, P, M, F, Y, L, D(E)) = (\$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[h35,1\text{-}h24]}N_{h35}) \mathbf{V} (\$_{[i,1\text{-}3]}L1_i \oplus$$
$$\downarrow\$_{[h32,1\text{-}h21]}P_{h32} \oplus \$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[h321,1\text{-}h21]}P_{h321}) \mathbf{V}$$
$$(\$_{[i,1\text{-}3]}L1_i \oplus \downarrow\$_{[h32,1\text{-}h21]}P_{h32} \oplus$$
$$\$_{[h31,1\text{-}h20]}C_{h31} \oplus \downarrow\$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[h35,1\text{-}h24]}N_{h35}) \mathbf{V} (\$_{[h34,1\text{-}h23]}M_{h34} \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[h31,1\text{-}h20]}C_{h31} \oplus$$
$$\downarrow\$_{[h32,1\text{-}h2]}P_{h32}) \mathbf{V} (\$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[h32,1\text{-}h21]}P_{h32}) \mathbf{V}$$
$$(\$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus$$

$$\downarrow\$_{[h32,1\text{-}h21]}P_{h32} \oplus \downarrow\$_{[h311,1\text{-}h20]}C_{h311} \oplus \downarrow\$_{[h35,1\text{-}h24]}N_{h35} \oplus \downarrow\$_{[h34,1\text{-}h23]}M_{h34} \oplus \downarrow\$_{[h331,1\text{-}h22]}G_{h331}) \mathbf{V} (\$_{[h34,1\text{-}h23]}M_{h34}$$

$$\oplus \$_{[h33,1\text{-}22]}G_{h33} \oplus \downarrow\$_{[h36,1\text{-}h25]}F_{h36} \oplus \downarrow\$_{[h31,1\text{-}h20]}C_{h31} \oplus \downarrow\$_{[i,1\text{-}70]}Y_i \oplus \downarrow\$_{[h32,1\text{-}h21]}P_{h32} \oplus \downarrow\$_{[h311,1\text{-}h20]}C_{h311}) \mathbf{V}$$

$$(\$_{[i,1\text{-}3]}L1_i \oplus \$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[h32,1\text{-}h21]}P_{h32} \oplus \$_{[j,1\text{-}70]}Y_j \oplus \downarrow\$_{[h321,1\text{-}h21]}P_{h321}) \mathbf{V}$$

$$(\downarrow\$_{[h34,1\text{-}h23]}M_{h34} \oplus \$_{[i,1\text{-}91]}L_i \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[h35,1\text{-}h24]}N_{h35}) \mathbf{V} (\$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[i,1\text{-}91]}L_i \oplus$$

$$\$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[h35,1\text{-}h24]}N_{h35}) \mathbf{V} (\$_{[i,1\text{-}91]}L_i \oplus \downarrow\$_{[h34,1\text{-}h23]}M_{h34} \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[j,1\text{-}70]}Y_j \oplus$$

$$\downarrow\$_{[h341,1\text{-}h23]}M_{h341}) \mathbf{V} (\$_{[j,1\text{-}42]}D(E_j) \oplus \$_{[h34,1\text{-}h23]}M_{h34} \oplus \$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[h36,1\text{-}h25]}F_{h36}$$

$$\oplus \$_{[h312,1\text{-}h20]}C_{h312})$$

где h31 ≠ h311, h32 ≠ h321, h33 ≠ h331 только для одного варианта выводимого предложения.

2.2. Математическая модель вывода вопросительных предложений

Математическая модель вывода вопросительного предложения на АЯ на основе десяти вариантов логико-лингвистических моделей [4] и семантических баз данных АЯ [3], после преобразований будет выражена как:

$$E2(G1, G, C, M, L, N, E, Y) = (\$_{[i,1\text{-}20]}G1_i \oplus \$_{[h31,1\text{-}h20]}C_{h31} \oplus \downarrow\$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[h311,1\text{-}h20]}C_{h311}) \mathbf{V}$$

$$(\$_{[i,1\text{-}20]}G1_i \oplus \$_{[h31,1\text{-}h20]}C_{h31} \oplus \downarrow\$_{[h34,1\text{-}h23]}M_{h34} \oplus \downarrow\$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow \$_{[h311,1\text{-}h20]}C_{h311}) \mathbf{V} (\$_{[i,1\text{-}20]}G1_i \oplus$$

$$\$_{[h31,1\text{-}h20]}C_{h31} \oplus \downarrow\$_{[h311,1\text{-}h20]}C_{h311}) \mathbf{V} (\$_{[i,1\text{-}20]}G1_i \oplus \$_{[h34,1\text{-}h23]}M_{h34} \oplus \downarrow \$_{[h31,1\text{-}h20]}C_{h31}) \mathbf{V} (\$_{[i,1\text{-}20]}G1_i \oplus$$

$$\downarrow\$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[h35,1\text{-}h24]}N_{h35}) \mathbf{V} (\$_{[i,1\text{-}91]}L_i \oplus \downarrow\$_{[h34,1\text{-}h23]}M_{h34} \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus$$

$$\downarrow\$_{[h35,1\text{-}h24]}N_{h35}) \mathbf{V} (\$_{[j,1\text{-}15]}E_j \oplus \$_{[i,1\text{-}91]}L_i \oplus \downarrow\$_{[h31,1\text{-}h20]}C_{h31} \oplus \downarrow\$_{[h34,1\text{-}h23]}M_{h34} \oplus \$_{[h33,1\text{-}h22]}G_{h33}) \mathbf{V} (\$_{[i,1\text{-}91]}L_i$$

$$\oplus \$_{[h31,1\text{-}h20]}C_{h31} \oplus \downarrow\$_{[h311,1\text{-}h20]}C_{h311} \oplus \downarrow\$_{[j,1\text{-}70]}Y_j \oplus \downarrow\$_{[h312,1\text{-}h20]}C_{h312}) \ \mathbf{V} \ (\$_{[i,1\text{-}91]}L_i \oplus \$_{[h34,1\text{-}h23]}M_{h34} \oplus$$

$$\downarrow\$_{[h31,1\text{-}h20]}C_{h31} \oplus \downarrow\$_{[j,1\text{-}70]}Y_j \oplus \downarrow\$_{[h311,1\text{-}h20]}C_{h311}) \ \mathbf{V} \ (\downarrow\$_{[h31,1\text{-}h20]}C_{h31} \oplus \downarrow\$_{[h34,1\text{-}h23]}M_{h34} \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus$$

$$\downarrow\$_{[h311,1\text{-}h20]}C_{h311} \oplus \$_{[i,1\text{-}91]}L_i \oplus \downarrow\$_{[h34,1\text{-}h23]}M_{h34})$$

где h31 ≠ h311≠h312 только для одного варианта выводимого предложения.

2.3. Математическая модель вывода восклицательных предложений

Математическая модель вывода восклицательных предложений на АЯ на основе двенадцати вариантов логико-лингвистических моделей [4] и семантических баз данных АЯ [3], после преобразований будет выражена как:

$$E3(C, L, G, M, Y, L1, E, P, D(E), N) = (\$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[i,1\text{-}91]}L_i \oplus \$_{[h33,1\text{-}h22]}G_{h33}) \ \mathbf{V} \ (\$_{[h31,1\text{-}h20]}C_{h31} \oplus$$

$$\downarrow\$_{[h311,1\text{-}h20]}C_{h311} \oplus \$_{[i,1\text{-}91]}L_i \oplus \$_{[h33,1\text{-}h22]}G_{h33}) \ \mathbf{V} \ (\$_{[i,1\text{-}91]}L_i \oplus \downarrow\$_{[h34,1\text{-}h23]}M_{h34} \oplus \$_{[j,1\text{-}70]}Y_j \oplus \$_{[h33,1\text{-}h22]}G_{h33}$$

$$\oplus \downarrow\$_{[j1,1\text{-}70]}Y_{j1} \oplus \downarrow\$_{[i1,1\text{-}3]}L1_{i1} \oplus \downarrow\$_{[h31,1\text{-}h20]}C_{h31}) \ \mathbf{V} \ (\$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[i1,1\text{-}3]}L1_{i1} \oplus \downarrow\$_{[h31,1\text{-}h20]}C_{h31}) \ \mathbf{V}$$

$$(\$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[j,1\text{-}3]}L1_j \oplus \downarrow\$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[i,1\text{-}91]}L_i) \ \mathbf{V} \ (\$_{[j,1\text{-}15]}E_j \oplus \downarrow\$_{[h32,1\text{-}h21]}P_{h32} \oplus \$_{[h31,1\text{-}h20]}C_{h31}$$

$$\oplus \downarrow\$_{[h34,1\text{-}h23]}M_{h34} \oplus \$_{[i,1\text{-}91]}L_i \oplus \downarrow\$_{[h33,1\text{-}h22]}G_{h33}) \ \mathbf{V} \ (\$_{[j,1\text{-}15]}E_j \oplus \downarrow\$_{[h32,1\text{-}h21]}P_{h32} \oplus \downarrow\$_{[i1,1\text{-}3]}L1_{i1} \oplus$$

$$\downarrow\$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[i,1\text{-}91]}L_i \oplus \downarrow\$_{[j1,1\text{-}42]}D(E_{j1}) \oplus \downarrow\$_{[h31,1\text{-}h20]}C_{h31}) \ \mathbf{V} \ (\$_{[j,1\text{-}15]}E_j \oplus \downarrow\$_{[h35,1\text{-}h24]}N_{h35} \oplus$$

$$\downarrow\$_{[h34,1\text{-}h23]}M_{h34} \oplus \$_{[h33,1\text{-}h22]}G_{h33}) \ \mathbf{V} \ (\$_{[j,1\text{-}15]}E_j \oplus \downarrow\$_{[h32,1\text{-}h21]}P_{h32} \oplus \downarrow\$_{[j,1\text{-}70]}Y_j \oplus \downarrow\$_{[h321,1\text{-}h21]}P_{h321} \oplus$$

$$\$_{[i1,1\text{-}3]}L1_{i1} \oplus \downarrow\$_{[h34,1\text{-}h23]}M_{h34} \oplus \downarrow\$_{[h311,1\text{-}h20]}C_{h311}) \ \mathbf{V} \ (\$_{[j,1\text{-}15]}E_j \oplus \downarrow\$_{[i1,1\text{-}3]}L1_{i1} \oplus \downarrow\$_{[h32,1\text{-}h21]}P_{h32} \oplus$$

$$\$_{[h31,1\text{-}h20]}C_{h31}) \ \mathbf{V} \ (\$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[i1,1\text{-}3]}L1_{i1} \oplus \downarrow\$_{[h31,1\text{-}h20]}C_{h31} \oplus \downarrow\$_{[h311,1\text{-}h20]}C_{h311}) \ \mathbf{V} \ (\downarrow\$_{[i,1\text{-}3]}L1_i \oplus$$

$$\$_{[h31,1\text{-}h20]}C_{h31} \oplus \downarrow\$_{[j,1\text{-}42]}D(E_j) \oplus \downarrow\$_{[i,1\text{-}3]}L1_i \oplus \$_{[h311,1\text{-}h20]}C_{h311} \oplus \downarrow \$_{[h312,1\text{-}h20]}C_{h312})$$

где h31 ≠ h311, h33 ≠ h331 только для одного варианта выводимого предложения.

2.4. Математическая модель вывода отрицательных предложений

Математическая модель вывода отрицательных предложений на АЯ на основе шести вариантов логико-лингвистических моделей [4] и семантических баз данных АЯ [3], после преобразований будет выражена как:

$$E4(M, L, X6, Y, G, D(E), L1, C, U) = (\$_{[h34,1\text{-}h23]}M_{h34} \oplus \$_{[i,1\text{-}91]}L_i \oplus \$_{[j,1\text{-}12]}X6_j \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus$$

$$\downarrow\$_{[j1,1\text{-}42]}D(E_{j1}) \oplus \downarrow\$_{[i1,1\text{-}91]}L_{i1} \oplus \downarrow\$_{[h31,1\text{-}h20]}C_{h31} \oplus \downarrow\$_{[j2,1\text{-}42]}D(E_{j2}) \oplus \downarrow\$_{[i2,1\text{-}91]}L_{i2} \oplus \$_{[h34,1\text{-}h23]}M_{h34}) \mathbf{V}$$

$$(\$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[i,1\text{-}91]}L_i \oplus \$_{[j,1\text{-}12]}X6_j \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[j1,1\text{-}42]}D(E_{j1}) \oplus \downarrow\$_{[i1,1\text{-}91]}L_{i1} \oplus \downarrow\$_{[h31,1\text{-}h20]}C_{h31}$$

$$\oplus \downarrow\$_{[j2,1\text{-}42]}D(E_{j2}) \oplus \downarrow\$_{[i2,1\text{-}91]}L_{i2} \oplus \downarrow\$_{[h34,1\text{-}h23]}M_{h34}) \mathbf{V} (\$_{[h34,1\text{-}h23]}M_{h34} \oplus \$_{[i,1\text{-}91]}L_i \oplus \downarrow\$_{[j,1\text{-}12]}U_j \oplus$$

$$\downarrow\$_{[h33,1\text{-}h22]}G_{h33} \oplus \downarrow\$_{[h34,1\text{-}h23]}M_{h34}) \mathbf{V} (\$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[i,1\text{-}91]}L_i \oplus \$_{[j,1\text{-}12]}U_j \oplus \downarrow\$_{[h33,1\text{-}h22]}G_{h33} \oplus$$

$$\downarrow\$_{[h34,1\text{-}h23]}M_{h34}) \mathbf{V} (\$_{[h34,1\text{-}h23]}M_{h34} \oplus \$_{[i,1\text{-}91]}L_i \oplus \$_{[j,1\text{-}12]}X6_j \oplus \$_{[h33,1\text{-}h22]} G_{h33} \oplus \$_{[j1,1\text{-}42]}D(E_{j1}) \oplus$$

$$\$_{[h34,1\text{-}h23]}M_{h34}) \mathbf{V} (\$_{[h31,1\text{-}h20]}C_{h31} \oplus \$_{[i,1\text{-}91]}L_i \oplus \$_{[j,1\text{-}12]}X6_j \oplus \$_{[h33,1\text{-}h22]}G_{h33} \oplus \$_{[j1,1\text{-}42]}D(E_{j1}) \oplus \$_{[h34,1\text{-}h23]}M_{h34})$$

где i ≠ i1 ≠ i2 только для одного варианта выводимого предложения.

ЛИТЕРАТУРА

1. Хакимов М.Х. Формальные системы машинного перевода в многоязычной ситуации. Материалы республиканской научной конференции «Современные проблемы математики, механики и инфор-

мационных технологий», НУУз, Институт Математики и ИТ АН РУз, Т, 2008, с. 297–301.

2. Хакимов М.Х. Расширяемый входной язык математического моделирования естественного языка для многоязычной ситуации машинного перевода. ЎзМУ хабарлари, № 1, 2009, с. 75–80.

3. Абдурахманова Н.З., М.Х. Хакимов. Семантические базы английского языка для многоязычной ситуации компьютерного перевода. Труды республиканской конференции «Проблемы современной математики». 22–23 апреля 2011 г., г. Карши, с. 311–314.

4. Абдурахманова Н.З., М.Х. Хакимов. Логико-лингвистические модели слов и предложений английского языка для многоязычных ситуаций компьютерного перевода. Компьютерная обработка тюрских языков. Первая международная конференция: Труды – Астана: ЕНУ им. Л.Н. Гумилева, 2013, с. 297–302.

СОДЕРЖАНИЕ

СЕКЦИЯ 5. **МАШИННЫЙ ПЕРЕВОД**

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

Труды конференции

Том 2

В авторской редакции

Подписано в печать 28.12.2017. Формат 60×84 1/16.
Усл. печ. л. 19,06. Тираж 100 экз.