УДК 004.8, 81'322

КЛАССИФИКАЦИЯ СУЩЕСТВИТЕЛЬНЫХ ТАДЖИКСКОГО ЯЗЫКА ДЛЯ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

Статья поступила в редакцию 25.03.2020, в окончательном варианте – 18.07.2020.

Мадибрагимов Навруз Шавкатович, Рязанский государственный радиотехнический университет имени В.Ф. Уткина (РГРТУ), 390005, Российская Федерация, г. Рязань, ул. Гагарина, 59/1,

аспирант, e-mail: navruzmadibragimov@gmail.com

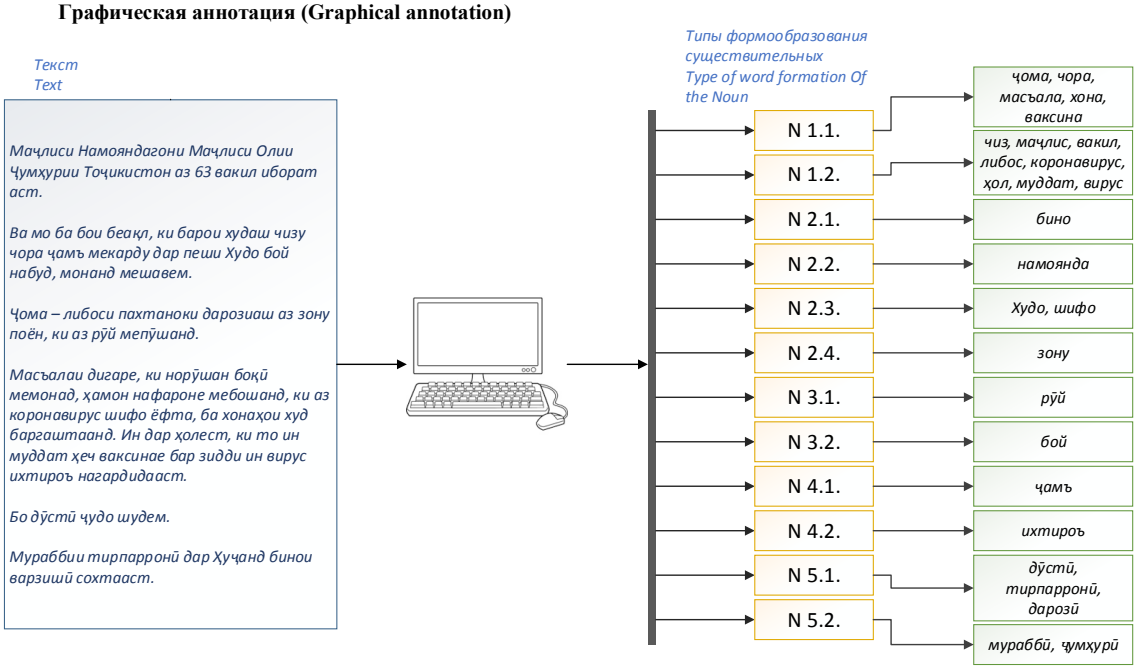
Пруцков Александр Викторович, Рязанский государственный радиотехнический университет имени В.Ф. Уткина (РязГМУ), 390005, Российская Федерация, г. Рязань, ул. Гагарина, 59/1; Рязанский государственный медицинский университет имени академика И.П. Павлова, 390026, Российская Федерация, г. Рязань, ул. Высоковольтная, 9,

доктор технических наук, профессор кафедры вычислительной и прикладной математики РГРТУ, доцент кафедры математики, физики и медицинской информатики РязГМУ; e-mail: mail@prutzkow.com

Таджикская компьютерная лингвистика остро нуждается в развитии, так как много трудов в этой сфере выполнены только на теоретическом уровне. Авторами данной статьи реализован универсальный метод генерации и определения словоформ таджикского языка. В работе описывается автоматическая обработка текстов и ее уровни, рассматривается морфологический уровень. Анализируются особенности таджикского языка и его система морфологии. Выполнен обзор исследований в области автоматической обработки текстов на таджикском языке на морфологическом уровне. Приводится предложенная классификация существительных таджикского языка по типам формообразования. Выделены 5 типов формообразования существительных таджикского языка и 12 подтипов. Для выделенных типов и подтипов охарактеризованы отличительные особенности. Результаты данного исследования послужили основой для программной реализации генерации форм слов таджикского языка в виде интернет-приложения.

Ключевые слова: компьютерная лингвистика, автоматическая обработка текста, таджикский язык, морфология таджикского языка, модель формообразования, генерация и определение форм слов, интернет-приложение

Графическая аннотация (Graphical annotation)



CLASSIFICATION OF NOUNS OF THE TAJIK LANGUAGE FOR NATURAL LANGUAGE PROCESSING

The article was received by the editorial board on 25.03.2020, in the final version – 18.07.2020.

Madibragimov Navruz Sh., Ryazan State Radio Engineering University named after V.F. Utkin (RSREU), 59/1 Gagarin St., Ryazan, 390005, Russian Federation,
postgraduate, e-mail: navruzmadibragimov@gmail.com
Prutzkow Alexander V., Ryazan State Radio Engineering University named after V.F. Utkina (RSREU), 59/1 Gagarin St., Ryazan, 390005, Russian Federation; Ryazan State Medical University named after I. P. Pavlov (RSMU), 390026, Russian Federation, Ryazan, Vysokovoltnaya str., 9
Doct. Sci. (Engineering), Professor of RSREU Department of Computational and Applied Mathematics, Assistant Professor of RSMU Department of Mathematics, Physics, and Medical Computer Sciences, e-mail: mail@prutzkow.com

Tajik computer linguistics remains in dire need of development because many works in this direction have been performed only at a theoretical level. We have implemented the use of a universal method for generating and determining word forms for the Tajik language. Also, we have described natural language processing its levels, consider the morphological level. The features of the Tajik language and the morphology system of the Tajik language are analyzed. Studies are also presented in the field of the Tajik language processing at a morphological level. Classification of the Tajik nouns by type of morphology is explained in detail. We have found 5 types of the Tajik nouns word-forming and 12 subtypes. These types and subtypes are characterized by peculiarity. The results of this study are the basis of software implementation of word form generation of Tajik language. The development of an Internet application for the generation of Tajik word forms is briefly outlined.

Keywords: computer linguistics, natural language linguistics, Tajik language, morphology of the Tajik language, form-building model, generation and recognition of word-forms, Internet application

Введение. Задачи автоматической обработки текстов (АОТ) продолжают оставаться актуальными уже на протяжении нескольких десятилетий. Причины этого: большая часть знаний, накопленных человечеством, хранится в виде текстов (например, в книгах и интернете); развитие информационно-телекоммуникационных технологий, обеспечившее расширение возможностей АОТ.

Наиболее активно развивающимися областями АОТ являются следующие: машинный перевод [9]; выявление в тексте знаний [5]; обработка корпусов текстов [28]; выявление нежелательных почтовых отправлений [10]; анализ персонального стиля автора текста (например, исследование [11]).

Для решения перечисленных, а также ряда других задач обработки текста, используется самый разнообразный математический аппарат: формальные языки и грамматики, алгебраические системы (например, алгебра предикатов [19]), методы математической статистики [18] и другие [22]. На сегодняшний день развитие АОТ на таджикском языке находится на раннем этапе и остро нуждается в дальнейших исследованиях. В перечисленных выше областях АОТ для таджикского языка нет работ с практической реализацией, за исключением [4]. Однако в этой области выполнены много трудов на теоретическом уровне.

Автоматическая обработка текстов и ее современное состояние. Организация взаимодействия человека с компьютером требует решения задачи компьютерного анализа и синтеза естественных языков (ЕЯ). Анализ подразумевает понимание речи или текста, а синтез – их генерацию. АОТ может рассматриваться как «преобразование текста на искусственном или естественном языке с помощью компьютера» [29]. Все запросы пользователя, поступающие в компьютер, и выдаваемые пользователю в любом виде ответы преобразуются и обрабатываются в виде текста. Поэтому задача АОТ является важной для разработки любого естественно-языкового интерфейса. Для преобразования текстового запроса пользователя в смысловое представление, а также для того, чтобы смысловое представление обратно преобразовать в текстовое представление, «необходимо выполнить АОТ на морфологическом, синтаксическом и семантическом уровнях» [23]. На морфологическом уровне происходит определение частей речи, выделение грамматической основы слова, приведение слова к стандартной словарной форме, то есть нахождение основы в словоформе. На синтаксическом уровне определяется синтаксическая структура предложения. На семантическом уровне устанавливаются смысловые связи и отношения между словами [30].

Морфологический уровень автоматической обработки текстов. Морфологический уровень АОТ предназначен для решения задач генерации (получения формы слова с заданными грамматическими значениями) и определения форм слов (распознавания по заданной словоформе ее грамматического значения и основы (лемматизация)) [21].

Существует большое число методов морфологического анализа и синтеза [23, 32]. В работах [19, 31] приводятся описания модели формообразования и метода генерации и определения форм слов

различных ЕЯ. Кратко опишем модель формообразования, необходимого для дальнейшего изложения.

Модель формообразования является алгебраической:

$$E = \{\Psi, \Omega\},$$

где Ψ – множество слов ЕЯ; Ω – множество операций над словами.

Множество Ψ включает множество нормальных форм слов (основ) S_M, множество остальных форм слов F_M ($S_M \cap F_M = \varnothing$), пустое слово $\wp$:

$$\Psi = S_M \cup F_M \cup \{\wp\}.$$

Множество Ω включает множество прямых операций Θ_M; множество обратных операций Θ'_M:

$$\Omega = \Theta_M \cup \Theta'_M.$$

При этом $\Theta_M \cap \Theta'_M = \varnothing$; $|\Theta_M| = |\Theta'_M| = N_\Theta$.

Множество Θ_M включает три операции:

1) префикс $P + A$, где $A \in \Psi$;

2) постфикс $A + P$;

3) замена первого слева вхождения подстроки H на подстроку P $H \rightarrow P$.

Для каждой прямой операции должна существовать обратная по действию операция.

Множество Θ_M может включать другие операции, необходимые для описания формообразования ЕЯ.

Последовательность операций называется цепочкой преобразований. Последовательность прямых операций называется прямой цепочкой R. Эта цепочка преобразует основу в словоформу. Обратная последовательность обратных операций называется обратной цепочкой R'. Обратная цепочка преобразует форму слова в основу. Для каждой прямой цепочки существует обратная цепочка.

Необходимость поддержания такой взаимосвязи между основой и словоформой налагает ограничения на цепочки и составляющие их операции. Цепочка преобразований должна обладать следующими свойствами:

1) однозначность результата: цепочка всегда приводит к одному и тому же результату;

2) обратимость действия: применение к форме А прямой цепочки, а затем обратной цепочки не изменяет ее: $A = (A(R))(R')$.

В [21] показано, что предложенные модель и метод применимы к пяти ЕЯ различных групп и семейств. Также в [21] доказано утверждение, что получение любой грамматической формы любого языка (даже неестественного) с морфологией можно представить в виде цепочки преобразований. Следовательно, словоизменение любого ЕЯ, в том числе и таджикского, можно описать в терминах модели формообразования. Таджикский язык был выбран по причине того, что один из авторов статьи является носителем этого языка и ему хорошо известна его морфология. В то же время этот язык имеет ряд особенностей, важных с точки зрения проведения морфологического анализа фраз.

Конкретная цель исследования. Целью данного исследования является формализация образования форм слов таджикского языка на основе предложенной модели формообразования. В этой статье будет представлена формализация образования существительных таджикского языка. Формализация образования форм слов состоит из нижеследующих этапов:

1) классифицировать слова по типам формообразования; слова будут относиться к определенному типу формообразования, если их словоизменение можно описать одинаковым способом преобразований;

2) описать словоизменения выделенных типов цепочками преобразований.

Обзор таджикского языка и его морфологии представлен в приложении 1. Обзор исследований в области АОТ на таджикском языке на морфологическом уровне в приложении 2.

Классификация существительных таджикского языка по типам формообразования. Метод определения и генерации словоформ путем представления образования форм слов последовательностью преобразований требует классифицировать основы слов ЕЯ. На рисунке 1 приведен алгоритм классификации основы ЕЯ по типу формообразования.

В настоящее время производится работа по классификации слов таджикского языка по правилам изменения слов и принципов словообразования. Данная классификация послужит для описания слов с помощью метода генерации и определения словоформ [12–13]. В работе [4] перечислены аффиксы для частей речи таджикского языка, и данные префиксы и постфиксы представлены в разделах каждой части речи в виде таблиц. Для существительного таджикского языка определены 1 801 префикс и постфикс. В данной работе исследуются 1 381 слово – основы существительных таджикского языка из словаря [2] и добавленные авторами слова. Стоит отметить, что небольшой словарь из книги [2] содержит наиболее употребляемые слова в таджикском языке. Так как словоизменение аффиксацией существительных таджикского языка происходит в зависимости от окончания слов, то у нас появилась необходимость добавить несколько слов в словарь для полноты и достоверности исследования.

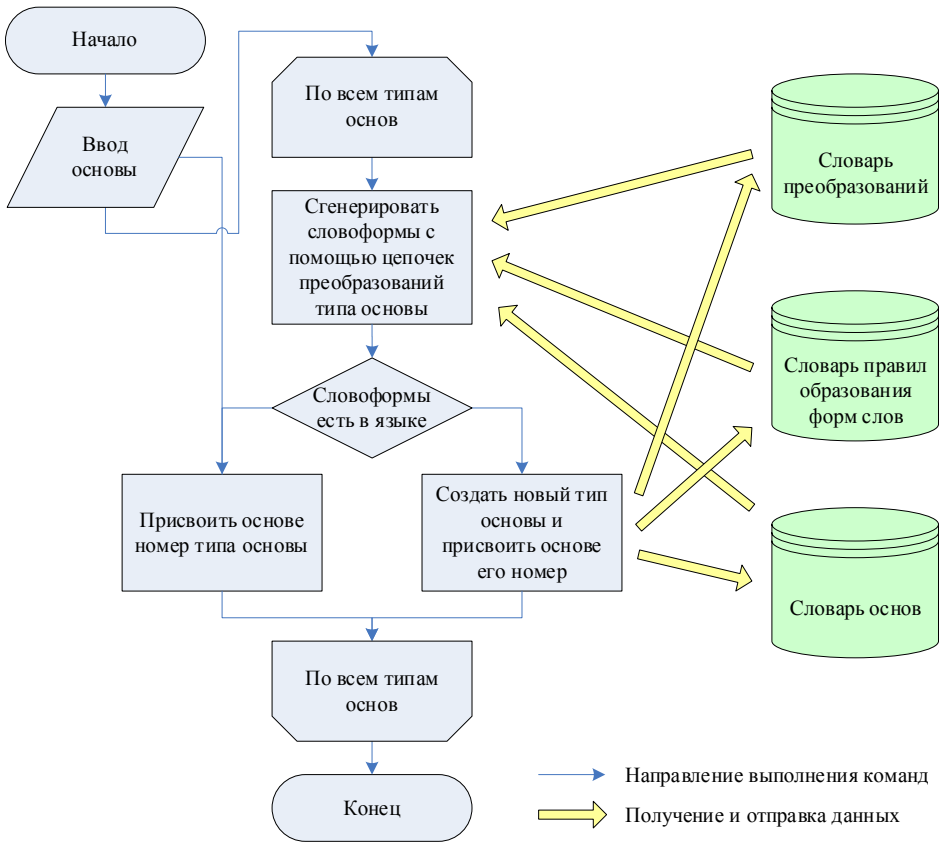


Рисунок 1 – Алгоритм классификации основы естественного языка по типу формообразования

Слова были исследованы в алфавитном порядке и классифицированы по типам формообразования. Классификация производилась следующим образом. Рассматривалось первое слово; описывались всевозможные преобразования его словоизменения. Рассматриваемое слово приписывалось к одному типу словоформ существительных. Затем рассматривалось второе слово, и если порядок его словоизменения совпадал с предыдущим словом (словами), то оно относилось к существующему типу, иначе оно становилось представителем нового типа и т.д.

После анализа формообразования существительных были выделены следующие типы и подтипы формообразования существительных таджикского языка:

1. **N 1.** Основы слов, которые заканчиваются на согласные буквы (*б, в, г, ғ, д, ж, з, к, қ, л, м, н, п, р, с, т, ф, х, ҳ, ч, ҷ, ш*). Этот тип словоформ описывается тем, что при морфологическом преобразовании слов основы не меняются, а постфиксы и вспомогательные слова приписываются по общим правилам, не меняя основу. Местоименные суффиксы применяются из 1 вида постфиксов (*+ат, +аш* и т.п.) [12].

1.1. **N 1.1.** Основы слов, которые заканчиваются на согласные буквы (*б, в, г, ғ, д, ж, з, к, қ, л, м, н, п, р, с, т, ф, х, ҳ, ч, ҷ, ш*) и на гласные буквы «*а*», «*я*», для которых в зависимости от окончания и стиля множественное число образуется только с помощью добавления постфикса *-ҳо* [2]. Пример: *телефон – телефонам* (мой телефон), *телефонат* (твой телефон), *телефонаш* (его/её телефон), ..., *телефони* (изафет), *телефонро* (телефона) ..., *телефонҳо* (телефоны) ...

1.2. **N 1.2.** Основы слов, множественное число которых образуется при помощи постфиксов *-ҳо* и *-он*. Пример: *сухан* (слово) – *суханам* (мое слово), *суханат* (твое слово), *суханаш* (его/её слово), ..., *сухани* (изафет), *суханро* (слова – род. п.), *суханҳо* (слова – мн. ч.), *суханон* (слова – мн. ч.) ...

2. **N 2.** Основы слов, которые заканчиваются на гласные буквы (*а, е, ё, и, о, у, ӯ, э, ю, я*). При морфологическом преобразовании слов основы не меняются, а постфиксы и вспомогательные слова приписываются по общим правилам, не меняя основу.

2.1. **N 2.1.** Основы слов, которые заканчиваются на гласные буквы (*е, ё, и, о, у, ӯ, э, ю, я*), множественное число которых образуется при помощи постфикса *-ҳо*. Местоименные суффиксы применяются из 2-го вида (*+ят, +яш* и т.д.). Пример: *савдо* (торговля) – *савдоям* (моя торговля),

савдоят (твоя торговля), *савдоятон* (ваша торговля), *савдояшон* (их торговля), ..., *савдои* (изафет), *савдоро* (торговли – р.п.), *савдоҳо* (торговли – мн. ч.) ...

2.2. **N 2.2.** Основы слов, которые заканчиваются на гласные буквы «*а», «я»*, множественное число которых образуется при помощи постфиксов *-ҳо* и *-гон*. Местоименные суффиксы применяются из 1-го вида постфиксов (*+ат, +аш* и т.п.). Пример: *нависанда* (писатель) – *нависандаам* (мой писатель, я писатель), ..., *нависандаашон* (их писатель), *нависандаҳо* (писатели), *нависандагон* (писатели); *ҳамсоя* (сосед)– *ҳамсояам* (мой сосед), *ҳамсояат* (твой сосед), ..., *ҳамсояҳо* (соседи – мн. ч.), *ҳамсоягон* (соседи – мн. ч.) ...

2.3. **N 2.3.** Основы слов, которые заканчиваются на гласные буквы *«о»*, для которых свойственно образование множественного числа, при помощи постфиксов *-ҳо* и *-ён*. Местоименные суффиксы, применяются из 2-го вида (*+ят, +яш* и т.д.). Пример: *бобо* (дедушка) – *бобоям* (мой дедушка), ..., *бобояшон* (их дедушка), *бобоҳо* (дедушки – мн. ч.), *бобоён* (дедушки – мн. ч.) ...

2.4. **N 2.4.** Основы слов, которые заканчиваются на гласные буквы *«у»,* «*ӯ*», множественное число которых образуется при помощи постфиксов *-ҳо* и *-вон*. Местоименные суффиксы применяются из 2-го вида (*+ят, +яш* и т.д.). Пример: *оҳу* (антилопа) – *оҳуям* (моя антилопа), ..., *оҳуяшон* (их антилопа), *оҳуҳо* (антилопы – мн. ч.), *оҳувон* (антилопы – мн. ч.); *абрӯ* (бровь) – *абрӯям* (моя бровь), ..., *абрӯшон* (их бровь), *абрӯҳо* (брови – мн. ч.), *абрӯвон* (брови – мн. ч.) ...

3. **N 3.** Основы слов, которые заканчиваются на «*й»* (и краткое). При добавлении местоименных суффиксов буква «*й»* в конце слова удаляется и применяются из 2-го вида. Все другие постфиксы и вспомогательные слова применяются по общим правилам.

3.1. **N 3.1.** Основы слов, множественное число которых образуется только при помощи постфикса *-ҳо*. Пример: *най* (флейта) – *найи* (изафет), *найро* (флейты – р. п.), *наям* (моя флейта), *наят* (твоя флейта), *наяшон* (их флейта), *найҳо* (флейты – мн. ч.) ...

3.2. **N 3.2.** Основы слов, множественное число которых образуется при помощи постфиксов *-ҳо* и *-ён*. При добавлении местоименных суффиксов буква «*й»* в конце удаляется и применяются из 2-вида. Остальные постфиксы и вспомогательные слова добавляются по общим правилам. Пример: *бой* (богатый человек) – *бойи* (изафет), *бойро* (богатого человека), *боям* (мой богатый человек), *боят* (твой богатый человек), *бояшон* (их богатый человек), *бойҳо, боён* (богатые люди) ...

4. **N 4.** Основы слов, которые заканчиваются на безгласную букву «ъ». Постфиксы и вспомогательные слова добавляются по общим правилам, не изменяя основу.

4.1. **N 4.1.** Предпоследняя буква основы, т. е. буква, стоящая перед «ъ», – согласная. Местоименные суффиксы применяются по принципу типа N 1.1. Пример: *навъ* (сорт) – *навъам* (мой сорт), *навъаш* (твой сорт), ... *навъро* (сорта – р. п.) ...

4.2. **N 4.2.** Предпоследняя буква основы, буква, стоящая перед «ъ», – гласная. Местоименные суффиксы применяются по принципу типа N 2.1. Пример: *мавзӯъ* (тема) – *мавзӯъям* (моя тема), *мавзӯъяш* (его/её тема), ... *мавзӯъро* (темы – р. п.) ...

5. **N 5.** Основы слов, которые заканчиваются на букву «*ӣ»* (*и заданок* (название буквы «*ӣ»* (и с макроном) на таджикском языке). При добавлении постфикса буква «*ӣ»* в конце заменяется на «*и»*. Местоименные суффиксы, как и слова с гласным окончанием, применяются из 2-го вида. Вспомогательные слова применются по общим правилам.

5.1. **N 5.1.** Основы слов, множественное число которых образуется только при помощи постфикса *-ҳо*. Пример: *сабзӣ* (морковь) – *сабзии* (изафет), *сабзиро* (моркови – р. п.), *сабзиям* (моя морковь), *сабзият* (твоя морковь), *сабзиҳо* (моркови – мн. ч.) ...

5.2. **N 5.2.** Основы слов, множественное число которых образуется при помощи постфиксов *-ҳо,* и *-ён*. Пример: *мураббӣ* (тренер) – *мураббии* (изафет), *мураббиро* (тренера – р. п.), *мураббиям* (мой тренер), *мураббият* (твой тренер), *мураббиҳо, мураббиён* (тренеры – мн. ч.) ...

Представленный словарь из 1381 существительного охватывает все слова-основы со всевозможными видами окончаний, что является достаточной базой для выявления всех типов и подтипов словоформ.

Местоименные суффиксы бывают двух видов:

1-й вид – постфиксы *+ам, +ат, +аш, +амон, +атон, +ашон,* применяются к существительным с окончанием на согласную букву;

2-й вид – постфиксы *+ям, +ят, +яш, +ямон, +яшон, +ятон,* используют существительные с окончанием на гласную букву.

Обобщая классифицирование типов существительных, можно сделать выводы, что по типам N 1, N 2 и N 4 используется только операция добавления постфиксов справа, а в типах N 3 и N 5 используется операция замены одной подстроки символов другой подстрокой.

В результате классификации существительных по типам формообразования была получена следующая статистика для указанного выше списка из 1381 слова (табл. 1).

Таблица 1 – Типы формообразования существительных таджикского языка и количество соответствующих им слов

№	Тип формообразования	Количество слов
1	N 1.1	792
2	N 1.2	398
3	N 2.1	56
4	N 2.2	37
5	N 2.3	10
6	N 2.4	4
7	N 3.1	10
8	N 3.2	3
9	N 4.1	3
10	N 4.2	8
11	N 5.1	36
12	N 5.2	24
Всего		1381

Большинство слов-существительных относятся к первому типу N 1, из них в типе N 1.1 слов почти два раза больше, чем в типе N 1.2. В целом тип N 1 включает в себя на порядок больше слов относительно других типов. Это означает, что для большинства существительных таджикского языка существуют единые правила словоизменения.

Описание словоизменения типов формообразования цепочками преобразований. В таджикском языке существуют 19 односложных (простых) префиксов: ***ба, бар, бе, би, бо, боз, бу, во, дар, ма, ме, на, но, то, фар, фур, ҳам, ҳаме, ҳар*** [16]. Перечисленные префиксы составляют исчерпывающий список простых префиксов таджикского языка, и этот перечень нельзя дополнить [7]. Префиксы в таджикском языке могут формироваться из трёх простых префиксов, иначе говоря, они могут быть трехуровневой сложности. Сочетая два или три простых префикса, можно образовать составные префиксы, при этом у двойных или тройных префиксов не будут повторяющиеся префиксы [4]. Но не все сочетания простых префиксов дают правильные составные префиксы. В работе [4] выполнялась процедура распознавания префиксов экспертом среди всевозможных сочетаний простых префиксов. В результате был выявлен 81 префикс, и этот набор является исчерпывающим.

В [4] на основе работы [7] и собственных исследований авторов составлен список из 113 простых постфиксов. В таджикском языке постфиксы могут быть составлены максимум из восьми простых постфиксов. Так как с помощью комбинаторно-статистического метода практически невозможно определить все составные постфиксы таджикского литературного языка, то итерационными процедурами (обработкой небольшой коллекции текстов) на данный момент составлен список постфиксов в размере 2 533 [4].

Словоформы в таджикском языке бывают словоизменительные, словообразовательные и словосочетательные [4]. Существительные в таджикском языке, принимая префиксы, образуют другие существительные, создавая словообразовательную словоформу. Мы рассматриваем словоизменительные и словосочетательные словоформы, поэтому изучаем словоизменение с помощью постфиксов. В ходе исследования было выявлено, что постфиксы существительных в таджикском языке могут быть составлены максимум из шести простых постфиксов. Это выяснено при формировании сложных постфиксов в виде сочетания неповторяющихся простых постфиксов. В таджикском языке выражения из нескольких слов можно составить из одной словосочетательной словоформы. Например, выражение «*ҳамаи аспҳои майдаи моро ва*» («*всех наших маленьких лошадок и…*») можно выразить одной словоформой «*аспчаяконамонрою*» или «*аспчаяконамонрову*», и такой прием равносильно используется как в разговорной речи, так и в литературном таджикском языке. На рисунке 2 приведен пример части графа формообразования слова «*асп*» («*лошадь*») (тип N 1.2).

Пространство словоформ можно расширить очень значительно, учитывая многоуровневость словоформ таджикского языка, а также наличием собранной большой базой аффиксов. Представления цепочками преобразований помогут при разработке программного обеспечения для решения задач генерации и определения форм слов таджикского языка.

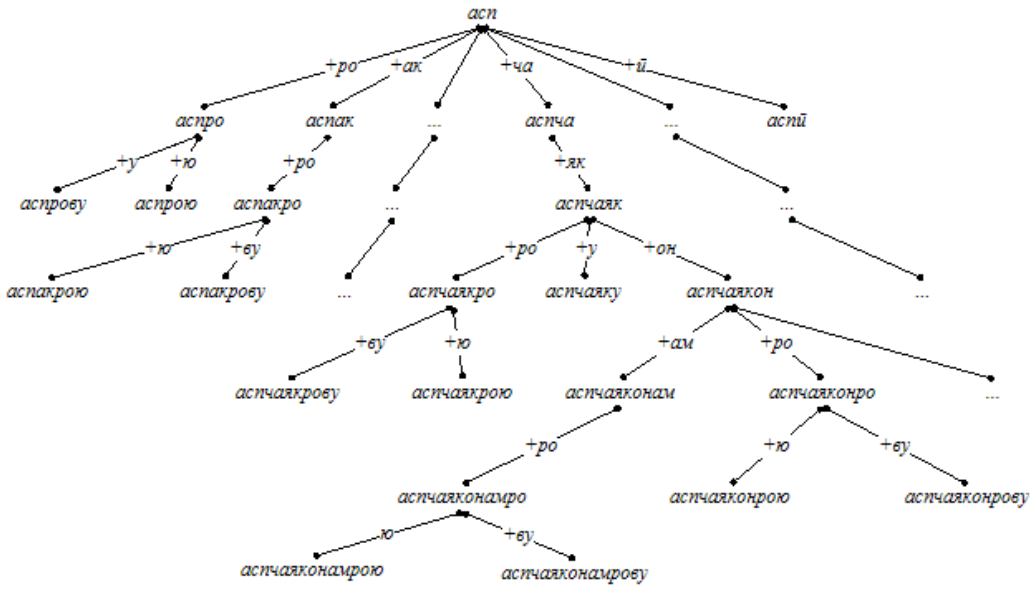


Рисунок 2 – Пример некоторых цепочек преобразований слова «*асп*» («*лошадь*»)

Разработка интернет-приложения для генерации таджикских словоформ. Классификации типов существительных таджикского языка реализована авторами в виде интернет-приложения на языке PHP7, с помощью фреймворка CodeIgniter версии 3.1.10. Интернет-приложение разработано на основе собственных знаний авторов о морфологии таджикского языка и консультаций специалистов в этой области. Внешний интерфейс интернет-приложения использует HTML, CSS, Javascript, объединенный в единый инструментарий – фреймворк Bootstrap версии 4.0 компании Twitter. Словарь основ существительных использован из словаря книги [2], а также есть добавления некоторых слов от самих авторов для полноты исследования. Из списка аффиксов существительных в диссертации Г.М. Довудова [4] была сформирована база постфиксов.

Практическая реализация интернет-приложения включает базу данных, состоящую из таблиц частей речи, аффиксов и словаря основ с классифицированными типами слов. Авторы назвали интернет-приложение «Ибора» (*ибора – тадж. «словосочетание»*).

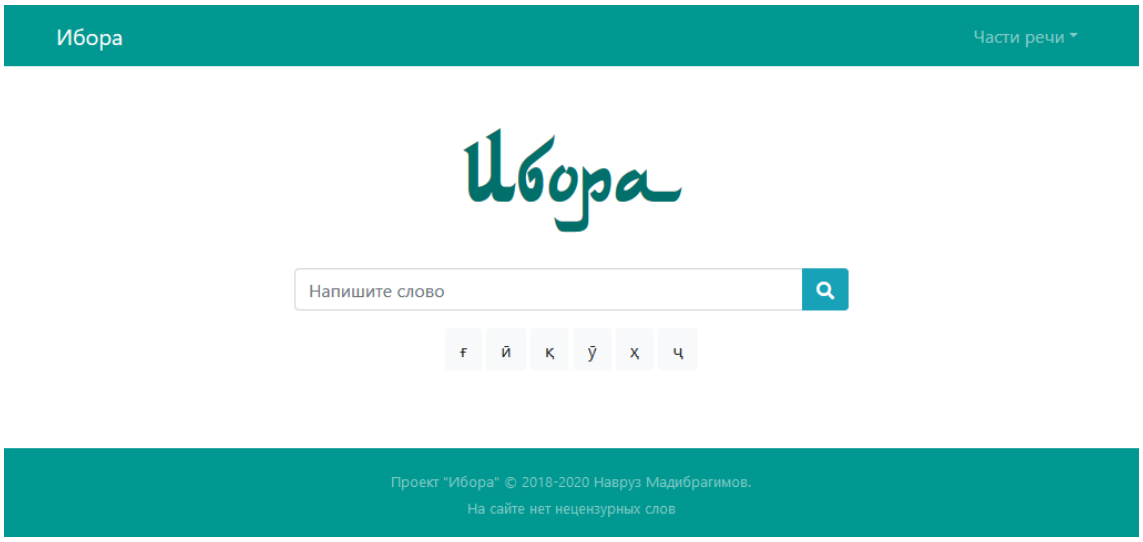
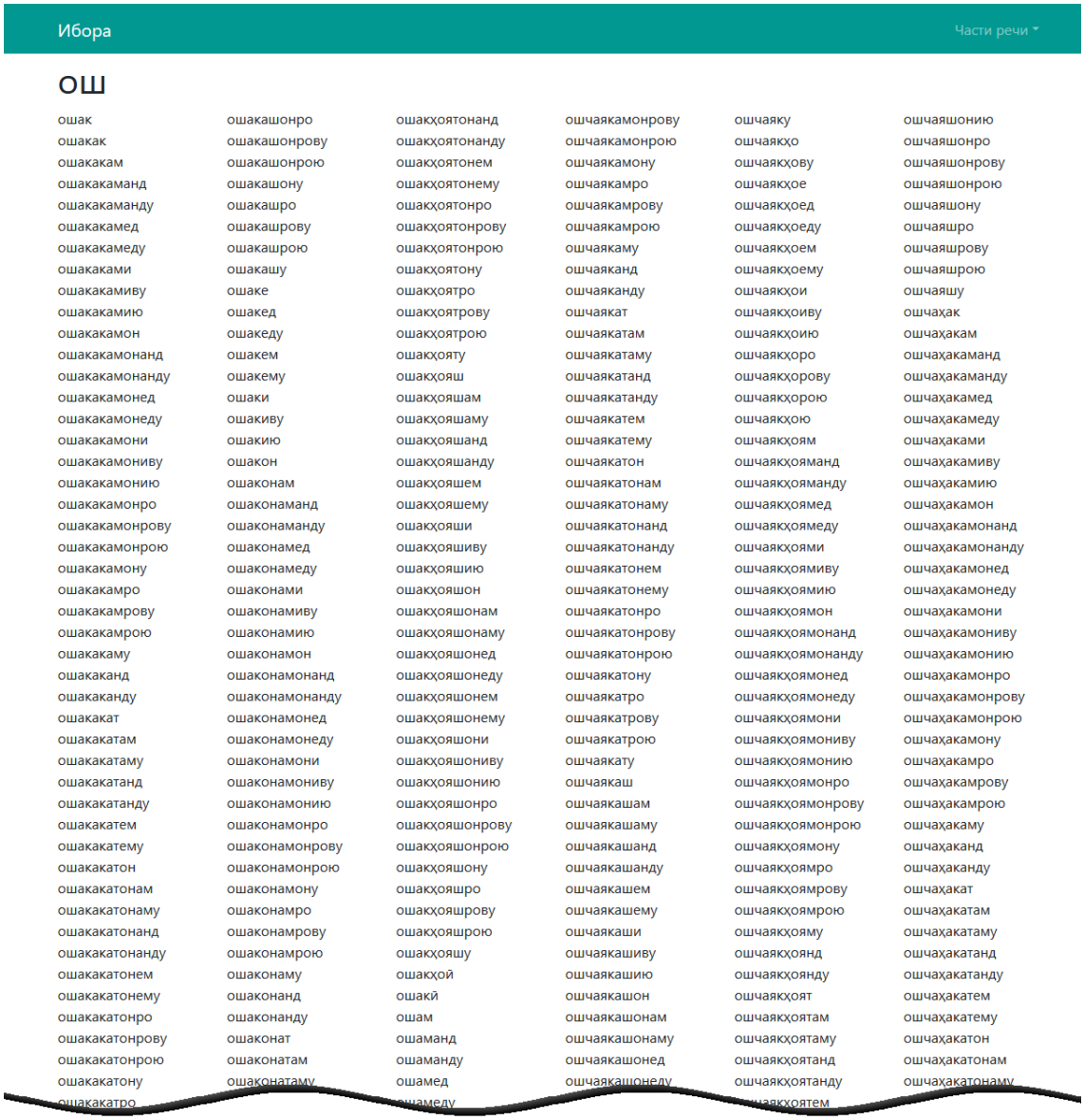


Рисунок 3 – Главная страница интернет-приложения «Ибора»

В интернет-приложении создана форма генерации словоформ. В форме есть кнопки ввода таджикских букв, которых нет на кириллической клавиатуре. При отправке слова через форму генерируются всевозможные варианты словоформ выбранного слова и выводятся на странице в виде списка в четыре столбца (рис. 4). Генерация осуществляется по правилам классификации на типы, описанные выше в данной статье.



Ибора

Части речи

ОШ

ошак
ошакак
ошакакам
ошакакаманд
ошакакаманду
ошакакамед
ошакакамеду
ошакаками
ошакакамиву
ошакакамию
ошакакамон
ошакакамонанд
ошакакамонанду
ошакакамонед
ошакакамонеду
ошакакамони
ошакакамониву
ошакакамонию
ошакакамонро
ошакакамонрову
ошакакамонрою
ошакакамону
ошакакамро
ошакакамрову
ошакакамрою
ошакакаму
ошакакаманд
ошакакаманду
ошакакатам
ошакакатаму
ошакакатанд
ошакакатанду
ошакакатем
ошакакатему
ошакакатон
ошакакатонам
ошакакатонаму
ошакакатонанд
ошакакатонанду
ошакакатонем
ошакакатонему
ошакакатонро
ошакакатонрову
ошакакатонрою
ошакакатону

ошакашонро
ошакашонрову
ошакашонрою
ошакашону
ошакашро
ошакашрову
ошакашрою
ошакашу
ошаке
ошакед
ошакеду
ошакем
ошакему
ошаки
ошакиву
ошакию
ошакон
ошаконам
ошаконаманд
ошаконаманду
ошаконамед
ошаконамеду
ошаконами
ошаконамиву
ошаконамию
ошаконамон
ошаконамонанд
ошаконамонанду
ошаконамонед
ошаконамонеду
ошаконамони
ошаконамониву
ошаконамонию
ошаконамонро
ошаконамонрову
ошаконамонрою
ошаконамону
ошаконамро
ошаконамрову
ошаконамрою
ошаконаму
ошаконанд
ошаконанду
ошаконат
ошаконатам
ошаконатаму

ошакҳоятонанд
ошакҳоятонанду
ошакҳоятонем
ошакҳоятонему
ошакҳоятонро
ошакҳоятонрову
ошакҳоятонрою
ошакҳоятону
ошакҳоятро
ошакҳоятрову
ошакҳоятрою
ошакҳояту
ошакҳояш
ошакҳояшам
ошакҳояшаму
ошакҳояшанд
ошакҳояшанду
ошакҳояшем
ошакҳояшему
ошакҳояши
ошакҳояшиву
ошакҳояшию
ошакҳояшон
ошакҳояшонам
ошакҳояшонаму
ошакҳояшонед
ошакҳояшонеду
ошакҳояшонем
ошакҳояшонему
ошакҳояшони
ошакҳояшониву
ошакҳояшонию
ошакҳояшонро
ошакҳояшонрову
ошакҳояшонрою
ошакҳояшону
ошакҳояшро
ошакҳояшрову
ошакҳояшрою
ошакҳояшу
ошакҳой
ошакй
ошам
ошаманд
ошаманду
ошамед
ошамеду

ошчаякамонрову
ошчаякамонрою
ошчаякамону
ошчаякамро
ошчаякамрову
ошчаякамрою
ошчаякаму
ошчаяканд
ошчаяканду
ошчаякат
ошчаякатам
ошчаякатаму
ошчаякатанд
ошчаякатанду
ошчаякатем
ошчаякатему
ошчаякатон
ошчаякатонам
ошчаякатонаму
ошчаякатонанд
ошчаякатонанду
ошчаякатонем
ошчаякатонему
ошчаякатонро
ошчаякатонрову
ошчаякатонрою
ошчаякатону
ошчаякатро
ошчаякатрову
ошчаякатрою
ошчаякату
ошчаякаш
ошчаякашам
ошчаякашаму
ошчаякашанд
ошчаякашанду
ошчаякашем
ошчаякашему
ошчаякаши
ошчаякашиву
ошчаякашию
ошчаякашон
ошчаякашонам
ошчаякашонаму
ошчаякашонед
ошчаякашонеду

ошчаяку
ошчаякҳо
ошчаякҳову
ошчаякҳое
ошчаякҳоед
ошчаякҳоеду
ошчаякҳоем
ошчаякҳоему
ошчаякҳои
ошчаякҳоиву
ошчаякҳоию
ошчаякҳоро
ошчаякҳорову
ошчаякҳорою
ошчаякҳою
ошчаякҳоям
ошчаякҳояманд
ошчаякҳояманду
ошчаякҳоямед
ошчаякҳоямеду
ошчаякҳоями
ошчаякҳоямиву
ошчаякҳоямию
ошчаякҳоямон
ошчаякҳоямонанд
ошчаякҳоямонанду
ошчаякҳоямонед
ошчаякҳоямонеду
ошчаякҳоямони
ошчаякҳоямониву
ошчаякҳоямонию
ошчаякҳоямонро
ошчаякҳоямонрову
ошчаякҳоямонрою
ошчаякҳоямону
ошчаякҳоямро
ошчаякҳоямрову
ошчаякҳоямрою
ошчаякҳояму
ошчаякҳоянд
ошчаякҳоянду
ошчаякҳоят
ошчаякҳоятам
ошчаякҳоятаму
ошчаякҳоятанд
ошчаякҳоятанду
ошчаякҳоятем

ошчаяшонию
ошчаяшонро
ошчаяшонрову
ошчаяшонрою
ошчаяшону
ошчаяшро
ошчаяшрову
ошчаяшрою
ошчаяшу
ошчаҳак
ошчаҳакам
ошчаҳакаманд
ошчаҳакаманду
ошчаҳакамед
ошчаҳакамеду
ошчаҳаками
ошчаҳакамиву
ошчаҳакамию
ошчаҳакамон
ошчаҳакамонанд
ошчаҳакамонанду
ошчаҳакамонед
ошчаҳакамонеду
ошчаҳакамони
ошчаҳакамониву
ошчаҳакамонию
ошчаҳакамонро
ошчаҳакамонрову
ошчаҳакамонрою
ошчаҳакамону
ошчаҳакамро
ошчаҳакамрову
ошчаҳакамрою
ошчаҳакаму
ошчаҳакаманд
ошчаҳакаманду
ошчаҳакат
ошчаҳакатам
ошчаҳакатаму
ошчаҳакатанд
ошчаҳакатанду
ошчаҳакатем
ошчаҳакатему
ошчаҳакатон
ошчаҳакатонам

ошчаҳоямонеду
ошчаҳоямони
ошчаҳоямониву
ошчаҳоямонию
ошчаҳоямонро
ошчаҳоямонрову
ошчаҳоямонрою
ошчаҳоямону
ошчаҳоямро
ошчаҳоямрову
ошчаҳоямрою
ошчаҳояму
ошчаҳоянд
ошчаҳоянду

Количество словоформ – 1263
К списку

Рисунок 4 – Страница сгенерированных словоформ слова «*ош*» (русск. «*плов*»)

Поскольку база слов основ существительных у нас относительно небольшая, всего 1382 слов, то в интернет-приложении создан список слов – основ существительных, каждое из которых в свою очередь является ссылкой на страницу сгенерированных словоформ данного слова.

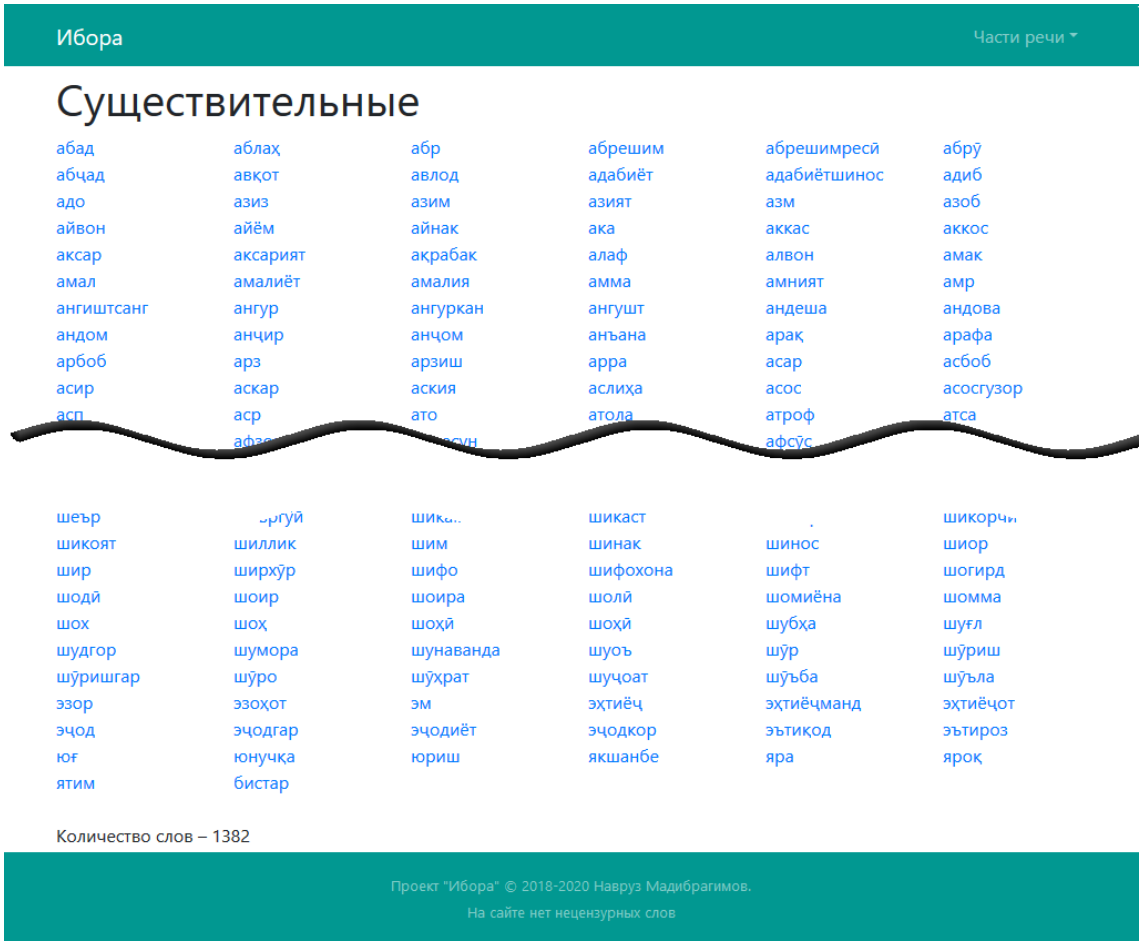


Рисунок 5 – Страница слов – основ существительных таджикского языка

Словоформы существительных в таджикском языке, особенно словосочетательные словоформы, могут образовываться в большом количестве. В таблице 2 приведена статистика, полученная в результате генерации словоформ с помощью описанного интернет-приложения.

Таблица 2 – Количество словоформ существительного

№	Тип формообразования	Количество словоформ одного слова
1	N 1.1	1265
2	N 1.2	1355
3	N 2.1	1266
4	N 2.2	1356
5	N 2.3	1356
6	N 2.4	1356
7	N 3.1	1265
8	N 3.2	1355
9	N 4.1	1265
10	N 4.2	1266
11	N 5.1	1266
12	N 5.2	1356

Таким образом, с использованием универсального метода генерации мы получили исчерпывающий перечень словосочетательных словоформ существительных таджикского языка.

Большое количество словоформ соответствует морфологии таджикского языка, однако не все из них встречаются в текстах часто. Согласно [4], словоформы с более чем 3-мя постфиксами

составляют менее 1 % из свыше 28 млн словоупотреблений, которые были исследованы в разнообразных текстах. В настоящее время в интернет-приложении решена задача генерации словоформ и продолжается реализация задачи определения форм слов.

Заключение. В ходе исследования морфологии таджикского языка были получены следующие результаты. **1.** Проведен анализ морфологии таджикского языка. **2.** Предложено для формализации формообразования таджикского языка использовать модель формообразования, которая применима к ЕЯ различных групп и семейств. **3.** Выполнен обзор современного состояния компьютерной морфологии таджикского языка. **4.** Классифицировано 1381 существительное таджикского языка. **5.** Определены 5 типов формообразования слов и их 12 подтипов. Для каждого типа и подтипа выделены характерные особенности.

Выполненная классификация основ на типы ускоряет перебор основ и аффиксов на сочетаемость при генерации словоформ существительных таджикского языка в разработанном авторами интернет-приложении. Все полученные результаты генерации словоформ были проверены и одобрены специалистами по морфологии таджикского языка.

Таджикский язык и его система морфологии очень близка к языку фарси. Однако применение представленных в статье методов исследований к языку фарси осложняется использованием другой письменности (в таджикском языке применяется кириллица, а в фарси – иранская письменность).

В настоящее время продолжается исследование прилагательных и глаголов таджикского языка, результаты которого также будут опубликованы. Разработанное интернет-приложение после его дополнения средствами работы с этими частями речи в итоге должно стать частью программного комплекса для АОТ на таджикском языке. Этот комплекс должен будет послужить инструментом для обработки информации на таджикском языке и оказать помощь носителям этого языка, исследователям и другим лицам, интересующимся рассматриваемой темой.

Приложение 1 – Основные сведения о таджикском языке и его морфологии

Таджикский язык относится к индоевропейскому семейству, иранской группе и является языком флективно-аналитического типа [15]. Аналитизм замечается в имени существительном, а также в глаголах, где наряду со старыми флективными формами есть и много новых аналитических форм. Современный алфавит таджикского языка построен на основе кириллицы и содержит 35 букв [32].

Таблица 3 – Алфавит таджикского языка

Аа	Бб	Вв	Гг	Ғғ	Дд	Ее
Ёё	Жж	Зз	Ии	Ӣӣ	Йй	Кк
Ққ	Лл	Мм	Нн	Оо	Пп	Рр
Сс	Тт	Уу	Ӯӯ	Фф	Хх	Ҳҳ
Чч	Ҷҷ	Шш	Ъъ	Ээ	Юю	Яя

Таблица 4 – Буквы, не входящие в русский алфавит

Описание	Г с палочкой	И с макроном	К выносное	У с макроном	Х выносное	Ч выносное
Буква	Ғ	Ӣ	Қ	Ӯ	Ҳ	Ҷ
Фонема	[ʁ]	[i]	[q]	[ɵ]	[h]	[dʒ]

Морфология таджикского языка произошла от морфологической системы древнеиранского языка, но сейчас она отличается значительно. В существительных таджикского языка нет понятий падежных склонений. Падежные отношения синтаксически передаются порядком слов в предложении и согласованием, изафетной конструкцией, а также сочетанием с предлогами и послелогами (частями слова, добавляемыми после основы). Одним из основных средств связи слов в предложении служит изафет – особый безударный показатель [15]. Например:

«Агар кас гузаштаи аҷдоди худро надонад, инсони комил нест!» [31]

(*Если человек не знает прошлого своих предков, он не идеальный человек!*)

Агар кас гузашта[и](изафет) аҷдод [и](изафет) худ[ро](послелог) (предлог)[на]дон[ад](послелог), инсон[и](изафет) комил нест!

В таджикском языке нет категории грамматического рода и нет определенных артиклей, но существуют изменения слов по лицам и числам. Все слова по умолчанию в основном обозначают лиц мужского рода. Когда требуется показать принадлежность к женскому роду, то основные слова сопровождаются вспомогательными словами «*зан*» или «*духтар*» (девушка или женщина) [17].

Имеет место отметить, что есть так называемый неопределенный артикль *як* (один) и *-е*. Артикль *як* используется перед существительным, а второй добавляется к существительному как послелог или суффикс. Прямое дополнение излагается с помощью суффикса *-ро*, пример: *Тимурро дидам* (Я увидел Тимура). Следует отметить, что для существительных таджикского языка в единственном числе характерен один признак, а для множественного числа – два признака. Существительные во множественном числе создаются с помощью суффиксов *-ҳо* или *-он (-ён, -гон, -вон)*.

В литературном таджикском языке есть морфологические свойственности, присущие арабским словам. Поэтому в таджикском языке одно слово может иметь два разных вида множественного числа. Например: *дарахтҳо=дарахтон (деревья); нависандаҳо=нависандагон (писатели).*

Глаголы тоже не разделяются на роды и не имеют явно выраженного вида. Для того чтобы выразить подобные категории, в предложениях необходимо менять порядок слов с помощью предлогов, послелогов, изафета и глаголов-связок.

Существуют два основных времени в простых глаголах: прошедшее и настоящее. Например: *рондан* (водить) – *ронд* (прошедшее время), *рон* (настоящее время). Некоторые простые глаголы супплетивны, это значит, подобно неправильным глаголам английского языка, один и тот же простой глагол может быть написан и озвучен по-разному. Например: *омадан* (прийти) – *омад* (прошедшее время), *биё* (настоящее время).

В таджикском языке есть совершенный и несовершенный вид глаголов, которые получаются префиксацией. Оба вида глаголов могут встречаться в трех временах: настоящем, прошедшем и предположительном прошедшем. Существует так называемый аорист – это тип прошедшего времени. Также есть четыре наклонения: изъявительное, сослагательное, повелительное и предположительное. Но при этом есть особое аудитивное наклонение, которое еще называется неочевидным [15].

Существует активный и пассивный залог. Для создания пассивного залога необходимо использовать вспомогательный глагол *шудан* (становиться) и деепричастия прошедшего времени основного глагола. В предложениях глаголы согласуются с подлежащим в лице и числе. Когда встречается сочетание существительного и глагола, получаются сложные глаголы.

В таджикском языке есть определенный порядок:

подлежащее – дополнение – сказуемое.

В изафетном сочетании определение встречается после определяемого слова [15]. Пример:

Мо себ[и] сурх дорем. (У нас есть красное яблоко.)

Себ – яблоко, *сурх* – красное, *себи сурх* – красное яблоко.

Прилагательные в предложениях могут играть роль обстоятельства, также дополнения, объясняя значение существительных и глаголов. Отсюда следует что, одно и то же прилагательное в предложениях может обозначать разные смыслы в зависимости от контекста. Например, «*нав*» может переводиться на русский язык как «новый» или «только что», а слово «*нағз*» – «хороший» или «хорошо».

Отличительные особенности таджикского языка на морфологическом уровне:

- неизменность (как правило) корня или основы при словообразовании;
- многоуровневость аффиксации (до 3 уровней префиксов и до 8 уровней постфиксов: *бар-на-ме-гардам, хоҳ-иш-манд-тар-ин-ҳо-яшон-ро-ву*);
- возможность внутриязыкового представления некоторых флективных форм в более аналитической форме;
- невозможность присутствия более одной согласной фонемы.

Использование (учет) этих особенностей позволяет выявить ряд неточностей в графике, фонетике и морфологии таджикского языка [7]. Также эти особенности помогут упростить сложности в процессе АОТ на таджикском языке. Сложность таджикского языка с позиции АОТ проявляется, к примеру, при генерации словоформ, в связи с необходимостью учета многоуровневости аффиксации и большого числа аффиксов.

Приложение 2 – Обзор исследований в области автоматической обработки текстов на таджикском языке на морфологическом уровне

Автоматизация морфологического анализа таджикских словоформ находит свое начало исследований в 1990 году. З.Д. Усманов и М.А. Исмоилов обозревали принципы распознавания словоформ на таджикском языке и элементов баз префиксов и постфиксов, с помощью которых создаются данные формы слов [25, 26].

М.А. Исмоилов в своих трудах [6, 7] описал концептуальную модель и изложил основы, как автоматизировать морфологический анализ словоформ таджикского языка. В этих работах проблема рассмотрена комплексно с позиции системного анализа.

Ф.А. Абдуллаев поднимал проблему об экспертном морфоанализе для форм слов таджикского языка, которые могут содержать в себе много корней [1]. Л.А. Гращенко разработал программу автоматической конверсии таджикско-персидского письма, в рамках которого предлагает морфологические представления словоформ [3]. Его новый подход заключается в том, что он подвергает таджикские заимствованные слова морфологическому анализу, учитывая морфологические основы других языков, из которых были заимствованы эти слова.

Р.М. Исмоилова изучила словоформы, образующиеся из существительных, прилагательных и числительных. Она предлагает представить словоформы в виде фрагментов предложений, близких к ним по значению [8]. Р.С. Назаров сделал вклад по расширению базы префиксов и постфиксов. Он также исследовал присоединение аффиксов к основам слов частей речи [16].

Д.Д. Собиров предложил метод и алгоритм распознавания глаголов в предложениях таджикского языка [25]. Его предложенный метод основывается на модельном алгоритме, объясняющем способ распознавания глагольных конструкций в предложениях таджикского языка. «В качестве глагольной конструкции в таджикском языке выступают простой глагол, сложно-именной глагол, сложно-деепричастный глагол и составной глагол или сочетание глагола с другими частями речи» [14]. В работе Д.Д. Собирова изучается распознавание простых и сложных глаголов. Он предложил теоретическую процедуру, которая любую словоформу таджикского языка разделяет на корни и аффиксы, а также определяет части речи. В итоге его изучений глаголов и глагольных конструкций была собрана исчерпывающая база глагольных постфиксов.

В своей диссертации [4] Г.М. Довудов исследует решение проблемы автоматического анализа таджикских словоформ. Этот труд основан на тщательном описании базы префиксов, постфиксов и слов – основ таджикских словоформ, по возможности исчерпывающим образом. Он выдвинул алгоритм, который объясняет автоматическое распознавание морфов в таджикских словоформах. В результате своих исследований Г.М. Довудов разработал «полуавтоматический морфоанализатор» [4], который основывается на базе морфов, состоящей из 81 префикса, 76 539 корней и 128 760 постфиксов.

В итоге анализа исследований по компьютерной морфологии таджикского языка сделан вывод о том, что многие работы по этой тематике охватывают автоматизацию отдельных особенностей или разделов морфологии языка. Наиболее полными работами являются труды М.А. Исмоилова и работы Г.М. Довудова. Эти работы послужили основой для исследования, результаты которого представлены в данной статье.

Библиографический список

1. Абдуллоев Ф. А. Моделирование процесса морфологического анализа многокоренных слов таджикского языка / Ф. А. Абдуллоев // Доклады АН РТ. – 2000. – Т. 43, № 4. – С. 4–7.

2. Арзуманов С. Д. Таджикский язык / С. Д. Арзуманов, А. Сангинов. – Душанбе : Маориф, 1988. – 416 с.

3. Гращенко Л. А. Математические основы автоматизированной таджикско-персидской конверсии графических систем письма : дис. ... канд. физ.-мат. наук: 05.13.18 / Л. А. Гращенко. – Душанбе, 2010. – 115 с.

4. Довудов Г. М. Компьютерный морфологический анализ таджикских словоформ : дис. … канд. техн. наук: 05.13.11: защищена 06.04.18 / Гулшан Мирбахоевич Довудов. – Душанбе, 2018. – 161 с.

5. Загорулько Ю. А. Искусственный интеллект. Инженерия знаний : учеб. пособие для вузов / Ю. А. Загорулько, Г. Б. Загорулько. – Москва : Юрайт, 2018. – 93 с.

6. Исмоилов М. А. Математическая модель морфологического анализа и синтеза слов таджикского языка / М. А. Исмоилов // Доклады АН РТ. – 1998. – Т. 41, № 9. – С. 63–68.

7. Исмоилов М. А. Основы автоматизированного морфологического анализа слов таджикского языка / М. А. Исмоилов. – Душанбе : ПИО НПИЦентр, 1994. – 156 с.

8. Исмоилова Р. М. К вопросу автоматизации морфологического анализа словоформ таджикского языка, образованных из имен числительных / Р. М. Исмоилова // Доклады АН РТ. – 1990. – Т. 33, № 10. – С. 652–655.

9. Карасев И. В. Системы машинного перевода / И. В. Карасев, Е. А. Артюшина // Успехи современного естествознания. – 2011. – № 7. – С. 117–118.

10. Ковалев С. С. Современные методы защиты от нежелательных почтовых рассылок / С. С. Ковалев, М. Г. Шишаев // Труды Кольского научного центра РАН. – 2011.

11. Косимов А. А. Разработка основ автоматической системы распознавания автора незнакомого текста : дис. ... канд. тех. наук: 05.13.11 / Абдунаби Абдурауфович Косимов. – Душанбе, 2018. – 104 с.

12. Мадибрагимов Н. Ш. Автоматизация морфологического анализа в промышленных системах обработки текстов / Н. Ш. Мадибрагимов // Современные технологии в науке и образовании – СТНО – 2020 : сб. тр. Междунар. науч.-техн. и науч.-метод. конф.: в 10 т. – Рязань : Рязан. гос. радиотехн. ун-т им. В.Ф. Уткина, 2020. – Т. 4. – С. 34–38.

13. Мадибрагимов Н. Компьютерные модели формообразования слов и их применение для описания морфологии таджикского языка / Н. Мадибрагимов // Современные технологии в науке и образовании – СТНО – 2018 : сб. тр. Междунар. науч.-техн. и науч.-метод. конф. : в 10 т. / под общ. ред. О. В. Миловзорова. – Рязань : Рязан. гос. радиотехн. ун-т, 2018. – Т. 1. – С. 65–68.

14. Мадибрагимов Н. Ш. Современные тенденции развития автоматического морфологического анализа таджикских словоформ / Н. Ш. Мадибрагимов // Современные технологии в науке и образовании – СТНО – 2019 : сб. тр. Междунар. науч.-техн. и науч.-метод. конф. : в 10 т. / под общ. ред. О. В. Миловзорова. – Рязань : Рязан. гос. радиотехн. ун-т, 2019. – Т. 4. – С. 12–15.

15. Махадов М. Самоучитель таджикского языка / М. Махадов. – Душанбе : Маориф, 1993. – 271 с.

16. Назаров Р. С. О множестве префиксов таджикского литературного языка / Р. С. Назаров // Доклады АН РТ. – 2006. – Т. 49, № 3. – С. 221–225.

17. Ниязмухаммадов Б. Морфология таджикского языка. На таджикском языке / Б. Ниязмухаммадов, Л. Бузург-зода. – Сталинабад : Таджикгосиздат, 1941.

18. Пиотровский Р. Г. Текст, машина, человек / Р. Г. Пиотровский. – Ленинград : Наука, 1975. – 327 с.

19. Правиков А. А. Разработка рекомендательной системы с естественно-языковым интерфейсом на основе математических моделей семантических объектов / А. А. Правиков, В. А. Фомичев // Бизнес-информатика. – 2010. – № 4 (14). – С. 3–11.

20. Пруцков А. В. Алгебраическое представление модели формообразования естественных языков / А. В. Пруцков // Cloud of Science. – 2014. – Т. 1, № 1. – С. 88–97.

21. Пруцков А. В. Генерация и определения форм слов естественных языков на основе их последовательных преобразований / А. В. Пруцков // Вестник Рязанского государственного радиотехнического университета. – 2009. – № 27. – С. 51–58.

22. Пруцков А. В. Задачи автоматической обработки текста на естественных языках и возможные математические подходы к их решениям / А. В. Пруцков // Вестник Рязанского государственного радиотехнического университета. – 2016. – № 1 (55). – С. 81–86.

23. Пруцков А. В. Модели, методы и программы автоматической обработки форм слов в естественно-языковых интерфейсах : дис. … д-ра техн. наук: 05.13.11 / Александр Викторович Пруцков. – Рязань, 2015. – 279 с.

24. 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.

25. Собиров Д. Д. Метод и алгоритм распознавания глаголов в предложениях таджикского языка / Д. Д. Собиров // Доклады АН РТ. – 2012. – Т. 55, № 2. – С. 120–125.

26. Усманов З. Д. Автоматическое распознавание элементов таджикского словаря, порождающих задание словоформы / З. Д. Усманов, М. А. Исмоилов // Доклады АН РТ. – 1990. – Т. 33, № 11. – С. 725–728.

27. Усманов З. Д. Концепция автоматического распознавания словоформ таджикского языка / З. Д. Усманов, М. А. Исмоилов // Доклады АН РТ. – 1990. – Т. 33, № 1. – С. 16–18.

28. Худойбердиев Х. А. Разработка параллельного корпуса таджикского и русского языков / Х. А. Худойбердиев, О. М. Солиев, П. А. Солиев // Новые информационные технологии в автоматизированных системах. – 2019.

29. Языкознание. Бол. энцикл. словарь / гл. ред. В. Н. Ярцева. – 2-е изд. – Москва : Бол. рос. энцикл., 1998. – 685 с.

30. Якубовский К. И. Обзор современных лингвистических технологий и систем / К. И. Якубовский, К. А. Якубовская // Вестник МГУП имени Ивана Федорова. – 2015. – № 2. – С. 315–319.

31. Ғафуров Б. Ғ. Тоҷикон: Таърихи қадимтарин, қадим, асрҳои миёна ва даврaи нав / Б. Ғ. Ғафуров. – Душанбе : Дониш, 2008. – 870 сах.

32. Perry J. R. A Tajik Persian Reference Grammar / J. R. Perry. – Boston : Brill, 2005.

33. Prutskov A. V. Algorithmic Provision of a Universal Method for Word-Form Generation and Recognition / A. V. Prutskov // Automatic Documentation and Mathematical Linguistics. –2011. –Vol. 45, № 5. – P. 232–238.

34. Tang X. English Morphological Analysis with Machine-learned Rules / X. Tang // Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation. – 2006. – Y06-1005. – P. 35–41.

References

1. Abdulloyev F. A. Modelirovaniye protsessa morfologicheskogo analiza mnogokorenykh slov tadzhikskogo yazyka [Modeling the process of morphological analysis of multicore words of the Tajik language]. *Doklady AN RT* [Reports of the RT Academy of Sciences]. Dushanbe, 2000, vol. 43, no 4, pp. 4–7.

2. Arzumanov S. D., Sanginov A. *Tadzhikskiy yazyk* [Tajik language]. Dushanbe, Maorif Publ., 1988. 416 p.

3. Grashchenko L. A. *Matematicheskiye osnovy avtomatizirovannoy tadzhiksko-persidskoy konversii graficheskikh sistem pisma* [Mathematical foundations of the automated Tajik-Persian version of graphic writing systems]. Dushanbe, 2010. 115 p.

4. Dovudov G. M. *Kompyuternyy morfologicheskiy analiz tadzhikskikh slovoform* [Computer morphological analysis of Tajik word forms]. Dushanbe, 2018. 161 p.

5. Zagorulko Yu. A., Zagorulko G. B. *Iskusstvennyy intellekt. Inzheneriya znaniy : uchebnoe posobiye dlya vuzov* [Artificial Intelligence. Knowledge Engineering : textbook manual]. Moscow, Yurayt Publ., 2018. 93 p.

6. Ismoilov M. A. Matematicheskaya model morfologicheskogo analiza i sinteza slov tadzhikskogo yazyka [Mathematical model of morphological analysis and synthesis of Tajik language words]. *Doklady AN RT* [Reports of the RT Academy of Sciences]. Dushanbe, 1998, vol. 41, no 9, pp. 63–68.

7. Ismoilov M. A. *Osnovy avtomatizirovannogo morfologicheskogo analiza slov tadzhikskogo yazyka* [Fundamentals of automated morphological analysis of the words of the Tajik language]. Dushanbe, 1994. 156 p.

8. Ismoilova R. M. K voprosu avtomatizatsii morfologicheskogo analiza slovoform tadzhikskogo yazyka, obrazovannykh iz imennykh chislitelnykh [On the issue of automation of morphological analysis of word forms of the Tajik language, formed from nominal numerals]. *Doklady AN RT* [Reports of the RT Academy of Sciences]. Dushanbe, 1990, vol. 33, no 10, pp. 652–655.

9. Karasev I. V., Artyushina Ye. A. Sistemy mashinnogo perevoda [Machine translation systems]. *Uspekhi sovremennogo yestestvoznaniya* [Successes of Modern Natural Science], 2011, no 7, pp. 117–118.

10. Kovalev S. S., Shishayev M. G. Sovremennyye metody zashchity ot nezhelatelnykh pochtovykh rassylok [Modern methods of protection against unwanted mailings]. *Trudy Kolskogo nauchnogo tsentra RAN* [Transactions of the Kola Science Center of the Russian Academy of Sciences], 2011.

11. Kosimov A.bA. *Razrabotka osnov avtomaticheskoy sistemy raspoznavaniya avtora neznakomogo teksta* [Development of the basis of an automatic recognition system for the author of an unfamiliar text]. Dushanbe, 2018. 104 p.

12. Madibragimov N. Sh. Avtomatizatsiya morfologicheskogo analiza v promyshlennykh sistemakh obrabotki tekstov [Automation of morphological analysis in industrial text processing systems]. *Sovremennyye tekhnologii v nauke i obrazovanii – STNO – 2020* [Modern technologies in science and education – STNO – 2020]. Ryazan, 2020. pp. 34–38.

13. Madibragimov N. Kompyuternyye modeli formoobrazovaniya slov i ikh primeneniye dlya opisaniya morfologii tadzhikskogo yazyka [Computer models of morphogenesis of words, and their application for description of tajik language morphology]. *Sovremennyye tekhnologii v nauke i obrazovanii – STNO – 2018* [Modern technologies in science and education – STNO – 2018]. Ryazan, 2018, pp. 65–68.

14. Madibragimov N.Sh. *Sovremennyye tendentsii razvitiya avtomaticheskogo morfologicheskogo analiza tadzhikskikh slovoform* [Modern trends in the development of automatic morphological analysis of Tajik word forms]. *Sovremennyye tekhnologii v nauke i obrazovanii – STNO – 2019* [Modern technologies in science and education – STNO – 2019]. Ryazan, 2019. pp. 12–15.

15. Makhadov M. *Samouchitel tadzhikskogo yazyka* [The self-learning manual of the Tajik language]. Dushanbe, Maorif Publ., 1993.

16. Nazarov R. S. O mnozhestve prefiksov tadzhikskogo literaturnogo yazyka [On the opinion of the prefixes of the Tajik literary language]. *Doklady AN RT* [Reports of the RT Academy of Sciences]. Dushanbe, 2006, vol. 49, no 3, pp. 221–225.

17. Niyazmukhammadov B., Buzurg-zoda L. *Morfologiya tadzhikskogo yazyka. Na tadzhikskom yazyke* [Morphology of the Tajik language. In Tajik language]. Stalinabad, Tadzhikgosizdat Publ., 1941.

18. Piotrovskiy R. G. *Tekst, mashina, chelovek* [Text, machine, human]. Leningrad, Nauka Publ., 1975. 327 p.

19. Pravikov A. A., Fomichev V. A. Razrabotka rekomendatelnoy sistemy s yestestvenno-yazykovym interfeysom na osnove matematicheskikh modeley semanticheskikh obektov [Development of a recommendation system with a local-language interface based on mathematical models of semantic objects]. Biznes-informatika [Business Informatics], 2010, no. 4 (14), pp. 3–11.

20. Prutskov A. V. Algebraicheskoye predstavleniye modeli formoobrazovaniya yestestvennykh yazykov [Algebraic representation of the morphogenesis model of natural languages]. *Cloud of Science*, 2014, vol. 1, no 1, pp. 88–97.

21. Prutskov A. V. Generatsiya i opredeleniya form slov yestestvennykh yazykov na osnove ikh posledovatelnykh preobrazovaniy [Generation and definition of word forms of natural languages based on their successive transformations]. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta* [Bulletin of the Ryazan State Radio Engineering University], 2009, no. 27, pp. 51–58.

22. Prutskov A. V. Zadachi avtomaticheskoy obrabotki teksta na yestestvennykh yazykakh i vozmozhnyye matematicheskiye podkhody k ikh resheniyam [Tasks of automatic text processing in natural languages and possible mathematical approaches to their solutions]. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta* [Bulletin of the Ryazan State Radio Engineering University]. 2016, no. 1 (55), pp. 81–86.

23. Prutskov A. V. *Modeli, metody i programmy avtomaticheskoy obrabotki form slov v yestestvenno-yazykovykh interfeysakh* [Models, methods and programs for automatic processing of word forms in natural language interfaces]. Ryazan, 2015. 279 p.

24. Prutskov A. V., Rozanov A. K. Metody morfologicheskoy obrabotki tekstov [Methods of morphological word processing]. *Prikaspiyskiy zhurnal: upravleniye i vysokiye tekhnologii* [Caspian Journal: Control and High Technologies], 2014, no. 3 (27), pp. 119–133.

25. Sobirov D. D. Metod i algoritm raspoznavaniya glagolov v predlozheniyakh tadzhikskogo yazyka [Verb recognition method and algorithm in sentences of the Tajik language]. *Doklady AN RT* [Reports of the RT Academy of Sciences], 2012, vol. 55, no. 2, pp. 120–125.

26. Usmanov Z. D. Avtomaticheskoye raspoznavaniye elementov tadzhikskogo slovarya, porozhdayushchikh zadaniye slovoformy [Automatic recognition of Tajik vocabulary elements generating the word form task]. *Doklady AN RT* [Reports of the RT Academy of Sciences], 1990, vol. 33, no. 11, pp. 725–728.

27. Usmanov Z. D. Kontseptsiya avtomaticheskogo raspoznavaniya slovoform tadzhikskogo yazyka [The concept of automatic recognition of Tajik word forms]. *Doklady AN RT* [Reports of the RT Academy of Sciences], 1990, vol. 33, no. 1, pp. 16–18.

28. Khudoyberdiyev Kh. A., Soliev O. M., Soliev P. A. Razrabotka parallelnogo korpusa tadzhikskogo i russkogo yazykov [Development of a parallel corpus of Tajik and Russian languages]. *Novyye informatsionnyye tekhnologii v avtomatizirovannykh sistemakh* [New information technologies in automated systems], 2019.

29. Yartseva V. N. (ed.) *Yazykoznaniye. Bolshoy entsiklopedicheskiy slovar* [Linguistics. Big Encyclopedic Dictionary]. Moscow, 1998. 685 p.

30. Yakubovskiy K. I., Yakubovskaya K. A. Obzor sovremennykh lingvisticheskikh tekhnologiy i sistem [Overview of modern linguistic technologies and systems]. *Vestnik MGUP imeni Ivana Fedorova* [Vestnik MGUP by Ivan Fedorov]. 2015, no 2, pp. 315–319.

31. Gafurov B. G. *The Tajiks: Prehistory, Ancient, and Medieval History*. Dushanbe, Donish Publ., 2008. 870 p.

32. Perry J. R. A Tajik Persian Reference Grammar. Boston, Brill, 2005.

33. Prutskov A. V. Algorithmic Provision of a Universal Method for Word-Form Generation and Recognition. *Automatic Documentation and Mathematical Linguistics*, 2011, vol. 45, no. 5, pp. 232–238.

34. Tang X. English Morphological Analysis with Machine-learned Rules. *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, 2006, Y06-1005, pp. 35–41.