Типы прилагательных и местоимений таджикского языка и их использование для генерации словоформ

Н. Ш. Мадибрагимов, А. В. Пруцков

***Аннотация* — Несмотря на информатизацию всех сфер жизни людей, компьютерная лингвистика таджикского языка испытывает нехватку в развитии. Причиной этого является недостаток числа выполненных исследований по этой тематике. В рамках проекта формализации словоизменения естественных языков для автоматической обработки текстов на таджикском языке предлагается классификация прилагательных и местоимений этого языка по типам формообразования. Классификация основана на универсальной модели формообразования, предполагающей, что словоизменение можно представить в виде цепочки преобразований конечной длины. Для 694 прилагательных таджикского языка выделены 5 типов и 2 подтипа формообразования. Для 32 слов, относящихся к местоимениям таджикского языка, выделены 5 типов формообразования. К одному типу формообразования относятся слова, получение форм которых описывается одинаковыми цепочками преобразований. Для выделенных типов и подтипов описаны отличительные особенности, указаны используемые в цепочках типы преобразований. Проведенная классификация продолжает исследование, начатое классификацией существительных таджикского языка. Классификация использована при заполнении лингвистической базы знаний Интернет-приложения, которое доступно другим исследователям и людям, изучающим этот язык, в разных частях мира. С помощью этой базы знаний Интернет-приложение генерирует формы слов таджикского языка. Продолжается классификация остальных частей речи таджикского языка.**

***Ключевые слова* — компьютерная лингвистика, автоматическая обработка текста, таджикский язык, морфология таджикского языка, модель формообразования, генерация форм слов, определение форм слов, Интернет-приложение.**

I. Введение

Процесс информатизации активно внедряется во все сферы нашей жизни, в том числе и обработку текстов. Базой для автоматической обработки текстов являются формальные модели различных уровней естественного языка (морфологического, синтаксического и семантического), алгоритмы анализа и синтеза для этих уровней, а также лингвистические базы знаний. Для заполнения лингвистических баз знаний требуется описать правила естественного языка согласно формальным моделям. Такие описания требуют значительных временных затрат и будут отличаться для различных языков. Однако без этой работы автоматическая обработка текстов невозможна.

Глобальной целью исследования является формализация образования форм слов таджикского языка на основе универсальной модели формообразования [1]. Уже были классифицированы существительные таджикского языка [2]. Целью этой статьи является представление результатов классификации прилагательных и местоимений таджикского языка.

Статья получена 18 октября 2021.

Мадибрагимов Навруз Шавкатович, Рязанский государственный радиотехнический университет имени В.Ф. Уткина (РГРТУ), 390005, Российская Федерация, Рязань, Гагарина, 59/1; Рязанский государственный медицинский университет имени академика И. П. Павлова (РязГМУ), 390026, Российская Федерация, Рязань, ул. Высоковольтная, 9 (e-mail: navruzmadibragimov@gmail.com)

Пруцков Александр Викторович, Рязанский государственный радиотехнический университет имени В. Ф. Уткина (РГРТУ), 390005, Российская Федерация, Рязань, Гагарина, 59/1; Рязанский государственный медицинский университет имени академика И. П. Павлова (РязГМУ), 390026, Российская Федерация, Рязань, ул. Высоковольтная, 9; Рязанский государственный университет имени С. А. Есенина (РГУ), 390000, Российская Федерация, Рязань, ул. Свободы, 46. (e-mail: mail@prutzkow.com)

II. Модель формообразования и алгоритмы генерации и определения словоформ

Модель формообразования [1] предполагает, что получение любой формы слова любого естественного языка с морфологией можно описать цепочкой (последовательностью) преобразований. На основе этой модели разработаны алгоритмы генерации и определения форм слов [3].

III. Классификация слов таджикского языка по типам формообразования

Проведенная классификация слов таджикского языка основана на результатах диссертационной работы Г.М. Довудова [4], выполненной под руководством основоположника компьютерной морфологии таджикского языка З.Д. Усманова. В своей диссертации Г.М. Довудов исследовал аффиксы частей речи таджикского языка. Для прилагательного были выявлены 940 аффиксов. Аффиксы разделены на три типа: словоизменительные, словообразовательные, словосочетательные. Из них в настоящей работе рассматриваются словоизменительные и словосочетательные аффиксы. Общее количество рассматриваемых аффиксов 179.

В качестве словаря исходных слов для классификации использовался словарь из учебника [5].

Классификация прилагательных и местоимений по

типам формообразования (как и проведенная классификация существительных) [2] происходила следующим образом. В алфавитном порядке выбиралось следующее слово и описывались все цепочки преобразований согласно модели формообразования. Исследуемое слово сопоставлялось с выделенными типами формообразования. Если цепочки преобразований слова и типа совпадали, то слово относилось к этому типу. Иначе создавался новый тип формообразования, и исследуемое слово относилось к нему.

IV. Классификация прилагательных таджикского языка по типам формообразования

Были классифицированы 694 прилагательных таджикского языка из словаря учебника [5], а также слова, отсутствующие в [5], но представляющие интерес с точки зрения автора статьи. В словарь были добавлены слова с нетипичными окончаниями основы, так как словоизменение в таджикском языке происходит с помощью аффиксов прилагательных таджикского языка в зависимости от окончания основы слова.

В таджикском языке местоименные суффиксы бывают двух видов:

1-й вид – постфиксы *+ам, +ат, +аш, +амон, +атон, +ашон* чаще всего применяются к прилагательным, основа которых заканчивается на согласную букву;

2-й вид – постфиксы *+ям, +ят, +яш, +ямон, +яшон, +ятон* используются для прилагательных, основа которых заканчивается на гласную букву.

После классификации формообразования прилагательных были выделены следующие типы и подтипы формообразования прилагательных таджикского языка:

S 1. Основы слов, заканчивающиеся 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*, а также на гласные буквы *а, я*. Этот тип характеризуется тем, что при морфологическом преобразовании основы слов не меняются, а постфиксы и вспомогательные слова добавляются по общим правилам. Местоименные суффиксы, применяются из 1-го вида постфиксов (*+ат, +аш* и т.п.).

S 1.1. Основы слов, заканчивающиеся 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.* Применяется энклитический союз *-у,* как эквивалент союза *ва* (соответствует союзу *и* в русском языке) [6]. Пример: *мураккаб* (сложный) – *мураккабу* ... (сложный и …), *мураккабат* (твой сложный …) и т.п.; *баланд* (высокий) – *баландро* (высокого), *баландат* (твой высокий …) и т.п.

S 1.2. Основы слов, заканчивающиеся на гласные буквы *а, я*. Применяется энклитический союз *-ю, -ву*. Пример: *ганда* (плохой) – *гандаю, гандаву* ... (плохой и ...), *гандаам* (мой плохой), *гандаат* (твой плохой) и т.п.

S 2. Основы слов, заканчивающиеся на гласные буквы *е, ё, о, у, ӯ*. При морфологическом преобразовании постфиксы и вспомогательные слова добавляются по общим правилам, не изменяя основу. Местоименные суффиксы, применяются из 2-го вида (*+ят, +яш* и т.д.). Пример: *аъло* (отлично) – *аълоям* (мой отличный …), *аълоят* (твой отличный …), *аълоятон* (ваш отличный …), *аълояшон* (их отличный …), ..., *аълои* (изафет), *аълоро* (отличного – род. падеж), *аълоҳо* (отличные – мн.ч.) и т.д. Стоит отметить, что основы слов, заканчивающиеся на гласные буквы *а, я*, которые ранее были отнесены к типу S 1.2., также могут применяться местоименные постфиксы 2-го вида и можно было отнести к типу S 2., например *гандаам = гандаям* (мой плохой), *гушнаам = гушнаям* (мой голодный). Но в силу того, что первый вариант (*гандаам, гушнаам*) на практике применяется больше, как в письменной, так и в разговорной речи, было решено отнести основы слов, заканчивающиеся на гласные буквы *а, я,* к типу S 1.2.

S 3. Основы слов, заканчивающиеся на «*ӣ*» (и краткое). При добавлении местоименных суффиксов, буква «*ӣ*» в конце слова удаляется. Остальные постфиксы и вспомогательные слова добавляются по общим правилам.

Пример: *хушрӯӣ* (красивый) – *хушрӯи* (изафет), *хушрӯйро* (красивого), *хушрӯям* (мой красивый), *хушрӯят* (твой красивый), *хушрӯяшон* (их красивый), *хушрӯйҳо* (красивые) и т.д.

S 4. Основы слов, заканчивающиеся на безгласный «ъ». Постфиксы и вспомогательные слова добавляются по общим правилам, не изменяя основу. Местоименные суффиксы, применяются по принципу типа S 1.

Пример: *хотирҷамъ* (спокойный; уравновешенный) – *хотирҷамъям* (мой спокойный), *хотирҷамъят* (твой спокойный), ... *хотирҷамъро* (спокойного – род. падеж) ...; *шуҷоъ* (смелый) – *шуҷоъам* (моя смелость), *шуҷоъат* (его/её смелость), ... *шуҷоъро* (смелости – род. падеж) …

S 5. Основы слов, заканчивающиеся 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 добавляются по общим правилам.

Пример: *илмӣ* (научный) – *илмии* (изафет), *илмиро* (научного – род. падеж), *илмиям* (моя научная …), *илмият* (твоя научная) …

В цепочках преобразований типов S 1, S 2 и S 4 используется только операция добавления постфиксов справа, а в цепочках типов S 3 и S 5 – также операция замены одной подстроки символов другой подстрокой.

694 прилагательных распределены по типам формообразования следующим образом (табл. 1).

В случае расширения числа постфиксов увеличится и количество словоформ.

Таблица 1. Типы формообразования прилагательных таджикского языка и количество соответствующих им слов

Тип формообразования	Количество соответствующих типу слов	Количество словоформ одного слова
S 1.1	380	95
S 1.2	71	100
S 2	22	100
S 3	5	95
S 4	2	95
S 5	214	100
Всего	694	

V. Классификация местоимений таджикского языка по типам формообразования

В [4] выделены 161 аффикс местоимений, из которых 64 аффикса словоизменительные и словосочетательные. В ходе этого исследования было добавлено еще 9 аффиксов местоимений (*-ямонанд, -ямонро, -ямро, -ястанд, -ясту, -ятонро, -ятро, -яшонро, -яшам*). В результате были получены 73 аффикса. Были классифицированы 32 местоимения из словаря учебника [5].

Проанализировав формообразования местоимений, авторами статьи были выделены следующие типы формообразования местоимений таджикского языка:

J 1. Основы слов, заканчивающиеся на согласные буквы *д, м, н, р, т, ч, ш.* В этом типе при морфологическом преобразовании основы слов не меняются, постфиксы и вспомогательные слова добавляются по общим правилам. Местоименные суффиксы применяются из 1-го вида постфиксов (*+ат, +аш и т.п.*). Применяется энклитический союз *-у*, как эквивалент союза *ва* (и).

Примеры: *фалон* (кто-то) – *фалонам* (мой кто-то), *фалонат* (твой кто-то), *..., фалонро* (кого-то), *фалону* (кто-то и …) и т.п.

J 2. Основы слов, заканчивающиеся на гласные буквы *а, я, о, е, и, у, ӯ*. Постфиксы и вспомогательные слова добавляются по общим правилам не изменяя основу. Местоименные суффиксы применяются из 2-го вида постфиксов (*+ятон, +яшон и т.д.*). Применяется энклитический союз *-ву, -ю*, как эквивалент союза *ва* (и). К этому типу характерно добавлять постфикс *-ст,* сокращение от глагола-связки *аст* (являться).

Примеры: *мо* (мы) – *мою, мову* (мы и …), *моем* (мы, мы есть), *моему* (мы есть и …), *моро* (нас), *морою, моровy* (нас и …), *морост* (нас …), *мост* (наш …) и т.п.

J 3. Основы слов, заканчивающиеся на «*й»* (и краткое). Местоименные суффиксы применяются из 2-вида постфиксов (*+ятон, +яшон и т.д.*). При добавлении постфиксов, буква «*й»* в конце слова удаляется, исключение составляют постфиксы начинающиеся на *-и, -ро*, а также постфикс *-ӣ*. Вспомогательные слова добавляются по общим правилам. Применяется энклитический союз *-ю*, как эквивалент союза *-ва* (и), и буква «*й»* в конце удаляется.

Примеры: *вай* (он, она) – *вайи* (изафет), *вайию, вайиву* (с изафетом, применяется энклитический союз *-ю, -ву*), *ваю* (он и …), *ваям* (тот мой), *ваямон* (тот наш) и т.п.

J 4. Основы слов, заканчивающиеся 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, применяются по принципу типа J 2. Вспомогательные слова добавляются по общим правилам. Применяется постфикс *-ст,* сокращение от глагола-связки *аст* (являться).

Примеры: *кӣ* (кто) – *кию, киву* (кто и …), *киед* (кто (вы)), *киро* (кого), *кист* (кто является) *кияш* (его кто), *кистӣ* (кем являешься), *кий* (кто (ты)) и т.п.

J 5. Слово *ман* (я). При добавлении постфикса *-ро* последняя буква «*н»* опускается: *маро* (меня). Остальные преобразования производятся по правилам типа J 1.

В цепочках преобразований типов J 1, J 2 используется только операция добавления постфиксов справа, в цепочках типов J 3, J 4 и J 5 – операция замены одной подстроки символов другой подстрокой.

В результате классификации местоимений по типам формообразования была получена следующая статистика для классифицированных 32 местоимений (табл. 2).

Таблица 2. Типы формообразования местоимений таджикского языка и количество соответствующих им слов

Тип формообразования	Количество соответствующих типу слов	Количество словоформ одного слова
J 1	21	38
J 2	7	47
J 3	1	38
J 4	2	47
J 5	1	38
Всего	32	

VI. Генерация таджикских словоформ прилагательных и местоимений в Интернет-приложении

Классификация прилагательных и местоимений, выполненная ранее классификация существительных, а также алгоритм генерации форм слов (рис. 1) реализованы в Интернет-приложении «Ибора» (*тадж. «словосочетание»*), которое расположено по адресу https://www.ibora.tk. При разработке этого Интернет-приложения использовались язык программирования PHP7 и фреймворк CodeIgniter 3.1.10. Словарь Интернет-приложения и базы постфиксов могут расширены при необходимости.

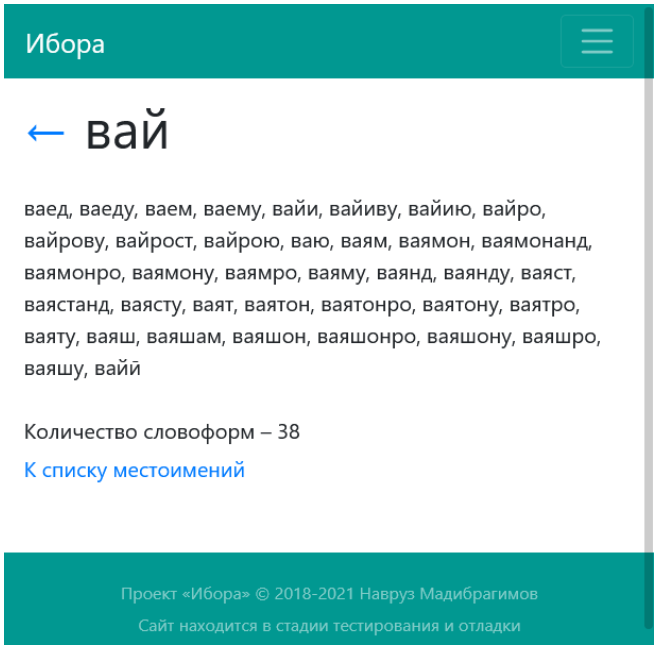

Рис.1. Страница сгенерированных словоформ слова «*вай*» (рус. «*он, (она, оно)*»)

VII. Заключение

В процессе исследования морфологии таджикского языка были получены следующие результаты. 1. Классифицировано 694 прилагательное и 32 местоимения таджикского языка. 2. Определены 5 типов и их 2 подтипа формообразования слов прилагательных, 5 типов формообразования слов местоимений. Для каждого типа и подтипа выделены характерные особенности.

Классификация прилагательных или местоимений таджикского языка по типам формообразования используется для генерации их словоформ в разработанном Интернет-приложении. Полученные результаты генерации словоформ были проверены специалистами по морфологии таджикского языка.

Исследование морфологии таджикского языка продолжается. В настоящее время классифицируются числительные, глаголы, наречия и причастия таджикского языка, результаты которого также будут также опубликованы. Разработанное Интернет-приложение после его дополнения средствами работы с этими частями речи в итоге должно стать частью программного комплекса для автоматической обработки текстов на таджикском языке. Этот комплекс должен будет послужить инструментом для обработки данных на таджикском языке и оказать помощь носителям этого языка и исследователям этой научной сферы.

Библиография

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

[2] Мадибрагимов Н.Ш., Пруцков А.В. Классификация существительных таджикского языка для автоматической обработки текстов // Прикаспийский журнал: управление и высокие технологии. – 2020. – № 4. – С. 39–52.

[3] Prutskov, A.V. Algorithmic provision of a universal method for word-form generation and recognition. In Automatic Documentation and Mathematical Linguistics, 2011, 45(5):232–238.

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

[5] Арзуманов С.Д., Сангинов А. Таджикский язык. – Душанбе: Маориф, 1988 – 416 с.

[6] Иванов В.Б., Семёнова Е.В., Хушкадамова Х.О. Учебник таджикского языка для стран СНГ: в 2 ч. Ч. 1 / Мос .гос. ун-т им. М.В. Ломоносова, Ин-т стран Азии и Африки. – М.: Ключ-С, 2009. – 232 с.

Types of adjectives and pronouns of the Tajik language and their use to generate word-forms

Navruz Madibragimov, Alexander Prutzkow

***Abstract* — Despite the informatization of all spheres of people's life, the computational linguistics of the Tajik language suffers from a lack of development. The reason is the lack of research done on this topic. Within the project of formalizing the inflection of natural languages for automatic processing of texts in the Tajik language, a classification of adjectives and pronouns of this language according to the types of morphogenesis is proposed. The classification is based on a universal morphogenesis model, which assumes that inflection can be represented as a chain of transformations of finite length. For 694 adjectives of the Tajik language, 5 types and 2 subtypes of morphogenesis are distinguished. For 32 words related to the pronouns of the Tajik language, 5 types of form formation have been identified. One type of shaping includes words, the receipt of forms of which is described by the same chains of transformations. For the selected types and subtypes, the distinctive features are described, the types of conversions used in the chains are indicated. The classification carried out continues the research begun by the classification of nouns in the Tajik language. The classification was used to fill in the linguistic knowledge base of an Internet application that is available to other researchers and people studying this language in different parts of the world. Using this knowledge base, an Internet application generates the forms of words in the Tajik language. The classification of the remaining parts of speech of the Tajik language continues.**

***Keywords*—computational linguistics, automatic text processing, Tajik language, morphology of the Tajik language, morphogenesis model, generation of word forms, determining word forms, Internet application.**

REFERENCES

[1] Dovudov G.M. Komp'yuternyy morfologicheskiy analiz tadzhikskikh slovoform [Computer morphological analysis of Tajik word forms]. Dushanbe, 2018. 161 p.

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

[3] Ivanov V.B., Semyonova E.V., Khushkadamova Kh.O. Textbook of the Tajik language for countries: in 2 parts, Part 1 / Moscow State University named after M.V. Lomonosov, Institute of Asian and African countries. - M.: Klyuch-S, 2009. – 232 c. – ISBN 978-5-93136-078-2.

[4] Madibragimov N.Sh., Prutskov A.V. Klassifikatsiya sushchestvitel'nykh tadzhikskogo yazyka dlya avtomaticheskoy obrabotki tekstov [Classification of nouns of the Tajik language for natural language processing] // Caspian journal: management and high technologies. – 2020. – № 4. – С. 39–52.

[5] 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.

[6] Madibragimov N. Komp'yuternyye 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.

[7] 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.