Классификация слов таджикского языка для их генерации и определения в автоматической обработке текстов

Н.Ш. Мадибрагимов, А.В. Пруцков

***Аннотация*—Автоматическая обработка текстов остается актуальной, несмотря на появление больших языковых моделей. Автоматическая обработка текстов выполняется на морфологическом, синтаксическом и семантическом уровнях. На морфологическом уровне решаются задачи генерации и определения форм слов. Для этого используется метод генерации и определения форм слов. Было показана универсальность этого метода для языков различных семейств и групп. Для применения метода необходимо классифицировать слова естественного языка по типам формообразования. Целью работы является автоматизация генерации и определения форм слов таджикского языка с помощью этого универсального метода. Были классифицированы числительные, наречия и причастия таджикского языка. 46 числительных были разделены на 3 типа. Числительные имеют от 88 до 91 формы. 289 наречий классифицированы также на 3 типа. Наречия имеют от 94 до 97 форм. 21 причастие также разделены на 3 типа. Причастия имеют 80 или 81 форму. Каждый тип имеет особенности, позволяющие отнести слова к этому типу. Ранее были классифицированы существительные, глаголы, прилагательные и местоимения. Типы этих частей речи, их особенности и их показатели также приведены в статье. Статья содержит итоговый результат пятилетнего исследования типов формообразования таджикского языка. Классификация использовалась при разработке Интернет-приложения «Ибора» (http://ibora.su) для генерации словоформ.**

***Ключевые слова* — Автоматическая обработка текстов, таджикский язык, морфология, генерация, определение, тип формообразования, классификация.**

Введение

В последнее десятилетие сфера искусственного интеллекта и машинной обработки естественной языка, в том числе больших языковых моделей, значительно прогрессируют. Во многих сферах используются сервисы (главным образом, в сети Интернет), которые распознают речь, анализируют текст для извлечения смысла, для выявления полезной информации, и на основе их предлагают производные услуги. Такие возможности появились благодаря развитию компьютерной лингвистики и, в частности, методов автоматической обработки текстов (АОТ). Этому способствовали разработки для различных языков, особенно английского, но и русского языка.

I. Автоматическая обработка текстов

АОТ – это сфера компьютерной лингвистики, связанная с анализом, преобразованием и созданием текстов с помощью программного обеспечения. Обработка текстов применяется для решения различных задач: нахождения ключевых слов в тексте [1], извлечения смысла из текста [2], классификации текстов [3], перевода текста с одного языка на другой [4] и др.

В классической лингвистике [5] текст необходимо разбить его на составляющие единицы разного уровня: морфологического, синтаксического и семантического [6]. На каждом уровне решаются задачи анализа и синтеза.

При семантическом анализе связываются слова и словосочетания, выявляется их смысл. Разработаны универсальные формализмы представления смысла и алгоритмы преобразования текста [7]. Алгоритмы семантического анализа должны обеспечивать результаты максимально похожие на естественный человеческий язык [8]. При синтаксическом анализе устанавливаются отношения между словами в предложении [9]. Морфологический анализ связывает слова предложения и их грамматические значения [10]. На морфологическом уровне АОТ решаются задачи генерации и определения словоформ. Генерация состоит в получении требуемой формы слова, а определение – лемматизацию – нахождение основы и грамматического значения словоформы [11].

АОТ на таджикском языке продолжает развиваться [12], в основном, за счет исследований ученых из Таджикистана.

Основоположником компьютерной лингвистике в Таджикистане является академик З.Д. Усманов. Под его руководством исследована морфология таджикского языка [13]. Разработан «Морфораспознаватель» – полуавтоматическая итеративная процедура для определения корней и аффиксов словоформ таджикского языка [14]. В этой работе сформирована база морфов таджикского языка, классифицированы аффиксы и словоформы, разработано позиционное кодирование таджикских словоформ, предложено эквивалентное представление словосочетательных словоформ фрагментами предложения, создан морфологический анализатор словоформ таджикского

Статья получена 26 февраля 2026 г.

Мадибрагимов Навруз Шавкатович, Рязанский государственный радиотехнический университет имени В.Ф. Уткина (РГРТУ), 390005, Российская Федерация, Рязань, Гагарина, 59/1 (e-mail: navruzmadibragimov@gmail.com);

Пруцков Александр Викторович, Рязанский государственный радиотехнический университет имени В. Ф. Уткина, 390005, Российская Федерация, Рязань, ул. Гагарина, 59/1; Липецкий государственный педагогический университет имени П. П. Семенова-Тян-Шанского, 398020, Российская Федерация, Липецк, ул. Ленина, 42 (e-mail: mail@prutzkow.com).

языка.

В [15] созданы алгоритмы обнаружения и исправления орфографических ошибок в тексте на таджикском языке. Разработаны веб-приложения «Автоматические системы обработки информации на таджикском языке» (http://www.tajlingvo.tj) [16], создан тезаурус и словарь MultiGANJ, предназначенный для перевода слов с таджикского языка на русский и английский языки, а также в обратном направлении [17], автоматическая система TajSpell-2.0 проверки орфографии таджикского языка в офисном пакете программ Microsoft Office версий 2010-2019 [18], компьютерное озвучивание таджикского текста Tajik Text-to-Speech [19].

В [20] исследована методика распознавания автора текста на таджикском языке. Разработан метод классификации текстов с использованием нейронных сетей [21], предложена модель структуры простых предложений, на основе которого разработан алгоритм исправления порядка слов в простых предложениях таджикского языка [22].

II. Метод генерации и определения форм слов

В [11] описан метод генерации и определения форм слов, который не зависит от естественного языка – является универсальным.

В основе метода лежит модель формообразования естественных языков. Модель предполагает, что получение словоформы с данным грамматическим значением можно представить в виде последовательности конечного числа преобразований над основой. Несколько преобразований можно объединить в цепочку.

Слова характеризуются типом формообразования. Два слова относятся к одному типу формообразования, если формы этих слов образуются одинаковыми цепочками преобразований.

Показано, что цепочки преобразований образуют алгоритмическую систему. Цепочка преобразований является алгоритмом преобразования одной формы в другую форму.

Универсальность метода была показана теоретически и нуждается в практическом подтверждении. Для этого был выбран таджикский язык.

III. Цель работы и структура статьи

Целью работы является автоматизация генерации и определения форм слов таджикского языка с помощью метода [11]. Для автоматизации необходимо классифицировать слова языка по типам формообразования. Классификация основана на [[13].

Статья имеет следующую структуру. Сначала кратко охарактеризуем морфологию таджикского языка. Вновь представим результаты классификации существительных, глаголов, прилагательных и местоимений. Затем приведем новые результаты классификации числительных, наречий и причастий.

IV. Морфология таджикского языка

Таджикский язык является флективно-аналитическим, агглютинативным языком, поэтому форма слова может содержать несколько аффиксов и имеет большое количество грамматических значений [23].

В таджикском языке грамматические значения чаще всего выражаются при помощи флексии. Союзы и частицы могут присоединяться к словам как окончания и порождать новые формы. Например, *хуб* (хороший), *хубу* (хороший и …), *хубак, хубча* (хорошенький), *хубаку, хубчаю* (хорошенький и …).

В таджикском языке нет системы флективных форм, нет падежей, а грамматические значения образуются с помощью предлогов, послелогов, изафета (специальной конструкции), грамматические отношения выражаются с помощью порядка слов [24]. Исследовался литературный таджикский язык.

Имя существительное в таджикском языке не имеет категории грамматического рода, нет определенных артиклей, изменяется по числам и лицам. Все существительные обозначают лица мужского пола. Для указания женского рода используются вспомогательные слова. Существительные в единственном числе имеют одну форму, а во множественном числе по две формы, образуются с помощью суффиксов *-ҳо* или *-он*.

В глаголах таджикского языка нет грамматического рода и грамматической категории вида глагола [25]. Категории образуются порядком слов в предложении с помощью предлогов, послелогов, изафетом и глагольными связками. Различают простые и сложные глаголы. Простые глаголы включают в себя две основы: прошедшего и настоящего времени. Сложные глаголы образуются с помощью сочетания существительного и глагола. В нашем исследовании рассматривались словоформы простых глаголов. Глаголы бывают совершенного и несовершенного вида, но эти виды определяются через систему времен глаголов, которая включает три основных времени. В систему входят 11 времен [26]. Существует отдельный тип глагола прошедшего времени – аорист. Глаголы имеют следующие наклонения: изъявительное, сослагательное, повелительное, предположительное, аудитивное, а также активные и пассивные залоги.

Прилагательные в таджикском языке делятся на качественные и относительные [27]. Существуют две степени сравнения прилагательных: сравнительная и превосходная. Сравнительная степень образуются с помощью суффиксов, образуя новые словоформы. Превосходную степень образуется с помощью вспомогательных слов. По форме различаются простые, производные и составные прилагательные. Также используется интенсивная форма глагола для обозначения оценки качественных прилагательных [28].

При формообразовании используются местоименные суффиксы двух видов:

1-й вид – постфиксы *+ам, +ат, +аш, +амон, +атон, +ашон* используются для слов основ, заканчивающихся на гласную букву;

2-й вид – постфиксы *+ям, +ят, +яш, +ямон, +яшон, +ятон* чаще всего применяются к словам основам, заканчивающимся на согласную букву.

Для получения возможных форм слова, той или иной части речи, необходимо генерировать словоформы используя соответствующие префиксы и постфиксы. В [14] собраны аффиксы для каждой части речи таджикского

языка, но рассматривались только словоизменительные аффиксы, поэтому префиксы игнорировались. Чтобы отнести какой-либо аффикс из собранного множества к определенному слову, необходимо классифицировать слова.

V. Словарь исходных слов

Слова для классификации были взяты из словаря учебника [26] по следующим причинам:

- целью исследования было подтверждение универсальности метода генерации и определения форм слов [11], а не полная классификация слов литературного таджикского языка;
- слова классифицировал один исследователь.

Кроме слов исходного словаря, авторами были добавлены и другие слова, представляющие интерес.

Всего было классифицировано по типам формообразования 2 714 слов таджикского языка разных частей речи.

VI. Классификация существительных

Существительные таджикского языка классифицированы следующим образом [29] (указаны типы формообразования и их особенности):

N 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*. При формообразовании основа не меняется, а постфиксы и вспомогательные слова добавляются по общим правилам. Применяются местоименные суффиксы 1-го вида (постфиксы *+ат, +аш* и т. д.).

- **N 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* и на гласные буквы *а, я*. Множественное число образуется постфиксом *-ҳо*.
- **N 1.2.** Основы слов, множественное число которых образуется постфиксами *-ҳо* и *-он*.

N 2. Основы слов, заканчивающиеся на гласные буквы *а, е, ё, и, о, у, ӯ, э, ю, я*. Основа не меняется, а постфиксы и вспомогательные слова добавляются по общим правилам.

- **N 2.1.** Основы слов, заканчивающиеся на гласные буквы *е, ё, и, о, у, ӯ, э, ю, я*. Множественное число образуется постфиксом *-ҳо*. Применяются местоименные суффиксы 2-го вида (*+ят, +яш* и т. д.).
- **N 2.2.** Основы слов, заканчивающиеся на гласные буквы *а, я*. В отличие от N 1.1, множественное число образуется постфиксами *-ҳо* и *-гон*. Применяются местоименные суффиксы 1-го вида.
- **N 2.3.** Основы слов, заканчивающиеся на гласные буквы *о*. Множественное число образуется постфиксами *-ҳо* и *-ён*. Применяются местоименные суффиксы 2-го вида.
- **N 2.4.** Основы слов, заканчивающиеся на гласные буквы *у, ӯ*. Множественное число образуется постфиксами *-ҳо* и *-вон*. Применяются местоименные суффиксы 2-го вида.

N 3. Основы слов, заканчивающиеся на *й* (и краткое). При добавлении местоименных суффиксов буква *й* в конце слова удаляется и используются местоименные суффиксы 2-го вида. Остальные постфиксы и вспомогательные слова добавляются по общим правилам.

- **N 3.1.** Основы слов, множественное число которых образуется постфиксом *-ҳо*.
- **N 3.2.** Основы слов, множественное число которых образуется постфиксами *-ҳо* и *-ён*.

N 4. Основы слов, заканчивающиеся на безгласный ъ. Постфиксы и вспомогательные слова добавляются по общим правилам, не изменяя основу.

- **N 4.1.** Предпоследняя буква основы является согласной. Применяются местоименные суффиксы как в типе N 1.1.
- **N 4.2.** Предпоследняя буква основы является гласной. Применяются местоименные суффиксы как в типе N 2.1.

N 5. Основы слов, заканчивающиеся на букву *ӣ* (и заданок, и с макроном). При добавлении постфикса, буква *ӣ* в конце заменяется на *и*. Применяются местоименные суффиксы 2-го вида, как и слова, заканчивающиеся на гласную. Вспомогательные слова добавляются по общим правилам.

- **N 5.1.** Основы слов, множественное число которых образуется постфиксом *-ҳо*.
- **N 5.2.** Основы слов, множественное число которых образуется постфиксами *-ҳо* и *-ён*.

Результат классификации существительных имеет следующий вид (таблица 1).

Таблица 1. Типы формообразования существительных таджикского языка и их показатели

Тип формообразования	Количество слов	Количество словоформ одного слова
N 1.1.	792	1265
N 1.2.	398	1355
N 2.1.	56	1266
N 2.2.	37	1356
N 2.3.	10	1356
N 2.4.	4	1356
N 3.1.	10	1265
N 3.2.	3	1355
N 4.1.	3	1265
N 4.2.	8	1266
N 5.1.	36	1266
N 5.2.	24	1356
Всего	1381	

VII. Классификация глаголов

Морфологические преобразования глагола таджикского языка базируются на двух основах глагола: основе настоящего времени (ОНВ) и основе прошедшего времени (ОПВ) [30]. При преобразовании инфинитива и для получения аориста к ОНВ добавляются личные окончания аориста или местоименные суффиксы. Для временных преобразований постфиксы времен добавляются к ОПВ.

В таджикском языке личные окончания для аориста бывают двух видов, которые также совпадают с окончаниями настояще-будущего времени [27]:

1-й вид – постфиксы *+ам, +ӣ, +ад, +ем, +ед, +анд*; эти окончания чаще всего применяются к глаголам с окончанием ОНВ на согласную букву;

2-й вид – постфиксы *+ям, +ӣ, +яд, +ем, +ед, +янд*, которые используются для глаголов с окончанием ОНВ на гласную букву.

После исследования формообразования глаголов были выделены следующие типы и подтипы формообразования глаголов таджикского языка:

F 1. Основы настоящего времени, 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*. При морфологическом преобразовании и для получения аориста к ОНВ добавляются личные окончания аориста 1-го вида постфиксов (*+ам, +ӣ, +ад, +ем, +ед, +анд*).

Для создания каузативов двух степеней используются конструкции ОНВ+*он(и)дан,* ОНВ+*онон(и)дан* соответственно: *бастондан* (принудить завязать), *бастонондан* (принудить завязать через кого-то). Спряжение по временам, где не используется деепричастие, а также образование формы предположительного наклонения используют единые правила для всех типов.

F 2. ОНВ, заканчивающиеся на гласные буквы *ё, о, у, ӯ*. Аорист получается путем добавления местоименных суффиксов 2-го вида (*+ям, +ӣ, +яд, +ем, +ед, +янд*). Для создания каузативов двух степеней используются конструкции ОНВ+*ён(и)дан,* ОНВ+*ёнон(и)дан* соответственно: *намоёндан* (заставить появляться), *намоёнондан* (заставить появляться через кого-то). Некоторые ОНВ имеют 2 вида. Второй вариант можно отнести к типу F 3. В нашей базе приведен только 1 вариант и слово отнесено к типу F 2.

F 3. ОНВ, заканчивающиеся на гласные букву *й*. При создании аориста буква *й* в конце удаляется. Личные окончания применяются по принципу типа F 2, т. е. местоименные суффиксы из 2-го вида.

Для создания каузативов двух степеней используются конструкции ОНВ+*ён(и)дан,* ОНВ+*ёнон(и)дан* соответственно: *поёндан* (заставить стеречь), *поёнондан* (заставить стеречь через кого-то).

F 4. Сложные глаголы, образованные с помощью префиксов, такие как *бозгаштан* (возвращаться), *бархӯрдан* (биться, врезаться), *даргирифтан* (зажигать) и др. При спряжении сложных глаголов в настоящее-будущем времени префиксы прибавляются не к префиксу, входящему в состав производного глагола, а к глаголу [27]. При создании аориста личные окончания применяются по принципу типа F 1.

Для создания каузативов двух степеней используются конструкции ОНВ+*он(и)дан,* ОНВ+*онон(и)дан* соответственно: *даргирондам* (принудил зажигать).

Исключением являются сложные глаголы, после префиксов которых идет гласная: *баромадан* (выйти), *даромадан* (войти), *фуромадан* (спуститься) и другие. Они спрягаются как простые глаголы.

F 5. ОНВ, заканчивающиеся на гласную букву *ӣ*. При создании аориста буква *ӣ* в конце заменяется на *и*. Личные окончания применяются по принципу типа F 2, т. е. местоименные суффиксы из 2-го вида.

Для создания каузативов двух степеней используются конструкции ОНВ+*ён(и)дан,* ОНВ+*ёнон(и)дан* соответственно: *зиёндан* (заставить жить), *зиёнондан* (заставить жить через кого-то).

F 6. Недостаточные (безличные) глаголы.

F 6.1. Глагол *боистан.* Этот глагол не спрягается, есть всего несколько форм: *бояд*, *набояд*, *набояд*, *мебоист* и т. д.

F 6.2. Глагол *шоистан.* Этот глагол тоже не спрягается, есть только одна форма *шояд.*

F 7. Глаголы, имеющие два вида ОНВ и общий ОПВ. Словоформы образуются каждого вида ОНВ отдельно, личные окончания применяются по принципу типа F 1.

Для создания каузативов двух степеней используются конструкции ОНВ+*он(и)дан,* ОНВ+*онон(и)дан* соответственно: *деҳондан* или *диҳондан* (заставить отдать), *деҳонондан* или *диҳонондан* (заставить отдать через кого-то).

F 8. Глаголы, имеющие два вида ОНВ и общий ОПВ. При добавлении префикса *би*, буква «*о*» вначале основы заменяется на «*ё*».

F 8.1. Глагол *омадан* (прийти). При создании аориста личные окончания применяются по принципу типа F 2, т. е. местоименные суффиксы 2-го вида. ОНВ – *о*, *ой*, формы с префиксом – *биёям* (приду), *биёӣ* (придешь), *биё* (приди), *биёед* (придите) и т. д.

Для создания каузативов двух степеней используются конструкции *би*+ОНВ+*ён(и)дан, би*+ОНВ+*ёнон(и)дан* соответственно: *биёёндан* (заставить прийти), *биёёнондан* (заставить прийти через кого-то).

F 8.2. Глагол *овардан* (принести). При создании аориста личные окончания применяются по принципу типа F 1 (*+ам, +ӣ, +ад, +ем, +ед, +анд*). ОНВ – *ор*, *овар*, формы с префиксом – *биёрам*, *биёварам* (принесу), *биёр*, *биёвар* (принеси) и т. д.

Для создания каузативов двух степеней используются конструкции *би*+ОНВ+*он(и)дан, би*+ОНВ+*онон(и)дан* соответственно: *биёрондан* (заставить принести), *биёронондан* (заставить принести через кого-то).

F 9. Вспомогательный глагол *ҳаст* (быть) и его отрицание *нест* (не быть, нет). Есть всего несколько форм, а именно спряжение по лицам и числам: *ҳастам* (я есть), *хастӣ* (ты есть), *ҳаст* (он есть) и т. д.; *нестам* (я не есть, меня нет), *нестӣ* (ты не есть, тебя нет), *нест* (он не есть, его нет) и т. д. Нет каузативных форм.

В результате классификации глаголы были распределены по типам формообразования (таблица 2).

Таблица 2. Типы формообразования глаголов таджикского языка и их показатели

Тип формообразования	Количество слов	Количество словоформ одного слова
F 1	147	1503
F 2	13	1518
F 3	2	1518
F 4	10	1503
F 5	1	1518
F 6.1	1	8
F 6.2	1	1
F 7	2	2655
F 8.1	1	1518
F 8.2	1	2655
F 9	2	12
Всего	181	

VIII. Классификация прилагательных и местоимений

Выделены следующие типы формообразования прилагательных таджикского языка и их особенности [31]:

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*. Основа не меняется, а постфиксы и вспомогательные слова добавляются по общим правилам. Применяются местоименные суффиксы 1-го вида.

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* (союз *и* в русском языке).

S 1.2. Основы слов, заканчивающиеся на гласные буквы *а, я*. Применяются энклитические союзы *-ю, -ву*.

S 2. Основы слов, заканчивающиеся на гласные буквы *е, ё, о, у, ӯ*. Применяются местоименные суффиксы 2-го вида. К основам слов типа S 1.2, заканчивающихся на гласные буквы *а, я*, также могут применяться местоименные постфиксы 2-го вида. Можно было бы отнести такие слова к типу S 2, например *гандаам* = *гандаям* (мой плохой), *гушнаам* = *гушнаям* (мой голодный). Однако первый вариант (*гандаам, гушнаам*) на практике применяется чаще в письменной и в разговорной речи, поэтому было решено отнести слова, основы которых оканчиваются на гласные буквы *а, я,* к типу S 1.2.

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

S 4. Основы слов, заканчивающиеся на безгласный ъ. Постфиксы и вспомогательные слова добавляются по общим правилам, не изменяя основу. Применяются те же местоименные суффиксы, что и в типе S 1. Применяются энклитические союзы *-ю, -ву*.

S 5. Основы слов, заканчивающиеся на букву *ӣ*. При добавлении постфикса буква *ӣ* в конце заменяется на *и*. Применяются местоименные суффиксы 2-го вида. Вспомогательные слова добавляются по общим правилам.

Типы формообразования прилагательных имеют следующие показатели (таблица 3).

Таблица 3. Типы формообразования прилагательных таджикского языка и их показатели

Тип формообразования	Количество 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	

IX. Классификация числительных

Проанализированы 46 числительных таджикского языка. После классификации других частей речи авторы выявили их типы формообразований и сформировали определенный принцип классификации. Авторы применили этот принцип для образования словоформ числительных. В [14] для числительных перечислены 168 постфиксов. После анализа формообразования числительных таджикского языка были выделены следующие типы формообразования:

X 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*. Применяются местоименные суффиксы 1-го вида.

X 2. Основы слов, которые заканчиваются на гласные буквы *а, е, и, у*. Применяются местоименные суффиксы 2-го вида.

X 3. Основы слов, заканчивающиеся на букву *ӣ*. Применяются местоименные суффиксы 2-го вида.

При формообразовании основа не меняется, а постфиксы и вспомогательные слова добавляются по общим правилам.

Классифицированные числительные распределены по типам формообразования следующим образом (таблица 4).

Таблица 4. Типы формообразования числительных таджикского языка и их показатели

Тип формообразования	Количество слов	Количество словоформ одного слова
X 1	27	88
X 2	8	91
X 3	11	91
Всего	46	

X. Классификация наречий

Были классифицированы 289 наречий таджикского языка. В [14] для наречий получены 166 постфиксов. Используя эти постфиксы, были сгенерированы словоформы наречий. Были выделены 3 типа формообразования:

Z 1. Основы слов, которые заканчиваются на согласные буквы *б, г, д, з, й, к, қ, л, м, н, п, р, т, ф, ш,* а также на безгласный ъ. Применяются местоименные суффиксы 1-го вида.

Z 2. Основы слов, которые заканчиваются на гласные буквы *а, о, и, у*. Применяются местоименные суффиксы 2-го вида.

Z 3. Основы слов, заканчивающиеся на букву *ӣ*. Применяются местоименные суффиксы 2-го вида.

При формообразовании основа не меняется, а постфиксы и вспомогательные слова добавляются по общим правилам. После классификации наречий получены следующие результаты (таблица 5).

Таблица 5. Типы формообразования наречий таджикского языка и их показатели

Тип формообразования	Количество слов	Количество словоформ одного слова
Z 1	134	94
Z 2	83	97
Z 3	72	97
Всего	289	

XI. Классификация причастий

Были исследованы 21 причастие из словаря книги [26] и классифицированы. В работе [14] перечислены 157 постфиксов для причастий.

Были выделены 3 типа формообразования причастий таджикского языка:

L 1. Основы слов, которые заканчиваются на согласные буквы *д, з, н, р*. Применяются местоименные суффиксы 1-го вида.

L 2. Основы слов, которые заканчиваются на гласные буквы *а, о*. Применяются местоименные суффиксы 2-го вида.

L 3. Основы слов, заканчивающиеся на букву *ӣ*. Применяются местоименные суффиксы 2-го вида.

При формообразовании основа не меняется, а постфиксы и вспомогательные слова добавляются по общим правилам.

В результате классификации словоформ причастий таджикского языка получена следующая статистика (таблица 6).

Таблица 6. Типы формообразования причастий таджикского языка и их показатели

Тип формообразования	Количество слов	Количество словоформ одного слова
L 1	15	81
L 2	5	80
L 3	1	80
Всего	21	

XII. Заключение

В статье представлены новые результаты классификации числительных, наречий и причастий. Ранее были классифицированы существительные, глаголы, прилагательные и местоимения. Краткие результаты их классификации также представлены в статье.

Статья содержит окончательную классификацию слов таджикского языка изменяемых частей речи по типам формообразования в соответствии с методом генерации и определения форм слов.

Исследование подтвердило применимость метода генерации и определения форм слов для таджикского языка.

Генерация форм слов таджикского языка реализована в Интернет-приложении «Ибора» (http://ibora.su).

Библиография

[1] Hasan K.S., Ng V. Automatic Keyphrase Extraction: A Survey of the State of the Art // 52nd Annual Meeting of the Association for Computational Linguistics. 2014. Vol. 1: Long Papers. Pp. 1262–1273.

[2] Markowitz D.M. The Meaning Extraction Method: An Approach to Evaluate Content Patterns from Large-Scale Language Data // Frontiers in Communication. 2021. No. 6. P. 588823.

[3] Li Q. et al. A Survey on Text Classification: From Traditional to Deep Learning // ACM Trans. Intell. Syst. Technol. 2021. Vol. 37. No. 4. P. 111.

[4] Головко Д.Р. Особенности и виды машинного перевода // Вестник Московского информационно-технологического университета – Московского архитектурно-строительного института. 2020. № 4. С. 24–30.

[5] Муродов П.С., Пруцков А.В. Математическая модель нечеткого определения тематики научных статей с помощью синтаксически связанных слов // Вестник Таджикского национального университета. Серия естественных наук. 2024. № 2. С. 14–22.

[6] Большакова Е.И., Воронцов К.В., Ефремова Н.Э. Автоматическая обработка текстов на естественном языке и анализ данных. М.: ВШЭ, 2017. 269 с.

[7] Фомичев В.А. Математическая модель многообразия естественно-языковых семантических структур и ее значение для биомедицинских наук // Информационные технологии. 2025. Т. 31. № 8. С. 405-418.

[8] Chernenko O., Gordeeva O. Semantic Analysis of Text Data with Automated System // 3rd Information Technology and Nanotechnology. 2017. Pp. 72–76.

[9] Сак А.Н., Бессонова Е.В. Сравнение синтаксического анализа предложения естественного языка // Балтийский гуманитарный журнал. 2021. Т. 10. № 1 (34). С. 373–377.

[10] Кочконбаева Б.О. О морфологическом анализе в приложениях автоматической обработки текста // Бюллетень науки и практики. 2018. Т. 4. № 12. С. 608–612.

[11] 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.

[12] Мадибрагимов Н.Ш. Современные тенденции развития автоматического морфологического анализа таджикских словоформ // Современные технологии в науке и образовании – СТНО-2019: сб. тр. междунар. науч.-техн. конф.: в 10 т. Рязань: РГРТУ, 2019. Т. 1. С. 12–15.

[13] Усманов З.Д., Довудов Г.М. Морфологический анализ словоформ таджикского языка. Душанбе: Дониш, 2015. 143 с.

[14] Усманов З.Д., Довудов Г.М. Концептуальная модель автоматического морфологического анализа таджикских словоформ // Доклады Академии наук Республики Таджикистан. 2014. Т. 57. № 3. С. 205–209.

[15] Худойбердиев Х.А. Об алгоритме проверки орфографии на примере таджикского языка // Политехнический вестник. Серия Интеллект. Инновации. Инвестиции. 2021. № 3 (55). С. 58–63.

[16] Худойбердиев Х.А., Назаров А.А., Ашурова Ш.Н. Формирование электронного словаря для системы автоматического перевода текста с таджикского языка на русский // Информационный обмен в междисциплинарных исследованиях II: статьи Всерос. науч.-практ. конф. с междунар. участием. Рязань: Акад. ФСИН России, 2023. С. 227–231.

[17] Худойбердиев Х.А., Солиев О.М. Лингвистический тезаурус таджикского языка // Новые информационные технологии в автоматизированных системах. 2017. № 20. С. 103–105.

[18] Солиев О.М.О., Худойбердиев Х.А., Довудов Г.М. Система автоматической проверки орфографии таджикского языка – TajSpell // Вестник Технологического университета Таджикистана. 2021. № 3 (46). С. 188–194.

[19] Худойбердиев Х.А. О синтезаторе таджикской речи по тексту // Новые информационные технологии в автоматизированных системах. 2013. № 16. С. 273–276.

[20] Усманов З.Д., Косимов А.А. Разработка программного комплекса для распознавания автора незнакомого текста. Душанбе: Дониш. 2022. 105 с.

[21] Косимов А.А. Математический метод описания нейронных сетей для классификации свойств // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. 2025. № 3 (71). С. 63–67.

[22] Қосимов А.А., Шамсов С.М. Таҳияи алгоритми хулосабарорӣ барои моделҳо (амсилаҳо) оиди ба тартиби дуруст овардани чумлаи содаи тоҷикӣ ҳангоми навишти нодуруст // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. 2025. № 3 (71). С. 68–72.

[23] Истамкулов Х., Музафаров Д. Методы токенизации текста на таджикском языке с помощью языка Python // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. 2023. С. 78–82.
[24] Ниязмухаммадов Б., Бузург-зода Л. Морфология таджикского языка. Сталинабад: Таджикгосиздат, 1941. 325 с.
[25] Атамова С.М. Соотнесенность форм видо-временной системы таджикского глагола с формами времени русского глагола // Ученые записки Худжандского государственного университета им. академика Б. Гафурова. Гуманитарные науки. 2016. № 2 (47). С. 182–187.
[26] Арзуманов С.Д., Джалолов О.Д. Учебник таджикского языка для вузов. Душанбе: Ирфон, 1969. 320 с.
[27] Арзуманов С.Д., Сангинов А. Таджикский язык. Душанбе: Маориф, 1988. 416 с.
[28] Сорокина М.А. Разряды имен прилагательных в таджикском, английском и русском языках // Вестник Педагогического университета. 2015. № 6-1 (67). С. 117–131.
[29] 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.
[30] Мадибрагимов Н.Ш. Особенности машинного морфологического анализа и синтеза глаголов таджикского языка // International Journal of Open Information Technologies. 2023. Т. 11. № 1. С. 79–86.
[31] Мадибрагимов Н.Ш., Пруцков А.В. Типы прилагательных и местоимений таджикского языка и их использование для генерации словоформ // International Journal of Open Information Technologies. 2021. Т. 9. № 11. С. 85–89.

Classification of the Tajik Language Words for Their Generation and Recognition in Natural Language Processing

Navruz Madibrgimov, Alexander Prutzkow

***Abstract*—Natural language processing remains relevant despite the emergence of large language models. Natural language is processed at the morphological, syntactic, and semantic levels. At the morphological level, word-form generation and recognition tasks are solved. For this purpose, a method for generating and recognizing word-forms is used. The universality of this method has been demonstrated for languages of various families and groups. To apply the method, it is necessary to classify natural language words by their word-forming types. The purpose of the study is automation of Tajik word-form generation and recognition by the universal method. We classify numerals, adverbs, and participles of the Tajik language. 46 numerals fall into 3 types. Numerals have from 88 to 91 forms. 289 adverbs were also classified into 3 types. Adverbs have from 94 to 97 forms. 21 participles were also divided into 3 types. Participles have 80 or 81 forms. Each type has attributes that allow words to be classified into that type. Nouns, verbs, adjectives, and pronouns were previously classified. The types of these parts of speech, their characteristics, and their attributes are resumed in the article. The article contains the final results of a five-year study of word formation types in the Tajik language. The classification was used in the development of the Ibora internet application (http://ibora.su) for generating word-forms.**

***Keywords*—Natural language processing, Tajik language, morphology, generation, recognition, word-forming type, classification.**

REFERENCES

[1] Hasan K.S., Ng V. Automatic Keyphrase Extraction: A Survey of the State of the Art. In *52nd Annual Meeting of the Association for Computational Linguistics*, vol 1: Long Papers, pp. 1262–1273, 2014.

[2] Markowitz D.M. The Meaning Extraction Method: An Approach to Evaluate Content Patterns from Large-Scale Language Data. *Frontiers in Communication*, no. 6., p. 588823, 2021.

[3] Li Q. et al. A Survey on Text Classification: From Traditional to Deep Learning, *ACM Trans. Intell. Syst. Technol*., vol. 37, no. 4, p. 111, 2021.

[4] Golovko D.R. Osobennosti i vidy mashinnogo perevoda [Features and Types of Machine Translation], *Vestnik Moskovskogo informatsionno-tekhnologicheskogo universiteta – Moskovskogo arkhitekturno-stroitel'nogo instituta*, no. 4. pp. 24–30, 2020. [In Rus].

[5] Murodov P.S., Prutzkow A.V. Matematicheskaja model' nechetkogo opredelenija tematiki nauchnykh statej s pomosch'ju sintaksicheski svjazannykh slov [Mathematical Model of Fuzzy Recognizing of the Topics of Scientific Articles Using Syntactically Related Words], *Vestnik Tadzhikskogo natsional'nogo universiteta*. *Serija estestvennykh nauk*, no. 2, pp. 14–22, 2024. [In Rus].

[6] Bol'shakova E.I., Vorontsov K.V., Efremova N.E. *Avtomaticheskaja obrabotka tekstov na estestvennom jazyke i analiz dannykh* [Automatic Processing of Natural Language Texts and Data Analysis]. Moscow, VShE, 2017, 269 pp. [In Rus].

[7] Fomichev V.A. Matematicheskaja model' mnogoobrazija estestvenno-jazykovykh semanticheskikh struktur i ee znachenie dlja biomeditsinskikh nauk [Mathematical Model of the Diversity of Natural language Semantic Structures and Its Significance for Biomedical Sciences], *Informatsionnye tekhnologii*, vol. 31, no. 8, pp. 405-418, 2025. [In Rus].

[8] Chernenko O., Gordeeva O. Semantic Analysis of Text Data with Automated System. In *3rd Information Technology and Nanotechnology*, pp. 72–76, 2017.

[9] Sak A.N., Bessonova E.V. Sravnenie sintaksicheskogo analiza predlozhenija estestvennogo jazyka [Comparison of Syntactic Analysis of a Natural Language Sentence], *Baltijskij gumanitarnyj zhurnal*, vol. 10, no. 1 (34), pp. 373–377, 2021. [In Rus].

[10] Kochkonbaeva B.O. O morfologicheskom analize v prilozhenijakh avtomaticheskoj obrabotki teksta [On Morphological Analysis in Automatic Text Processing Applications] *Bjulleten' nauki i praktiki*, vol. 4, no. 12, pp. 608–612, 2018. [In Rus].

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

[12] Madibragimov N.Sh. Sovremennye tendentsii razvitija avtomaticheskogo morfologicheskogo analiza tadzhikskikh slovoform [Current Trends in the Development of Automatic Morphological Analysis of Tajik Word-Forms]. In *Sovremennye tekhnologii v nauke i obrazovanii – STNO-2019*: sb. tr. mezhdunar. nauch.-tekhn. konf.: v 10 t. Rjazan', RGRTU, 2019, vol. 1, pp. 12–15. [In Rus].

[13] Usmanov Z.D., Dovudov G.M. *Morfologicheskij analiz slovoform tadzhikskogo jazyka* [Morphological Analysis of Word-Forms of the Tajik Language]. Dushanbe, Donish, 2015, 143 pp. [In Rus].

[14] Usmanov Z.D., Dovudov G.M. Kontseptual'naja model' avtomaticheskogo morfologicheskogo analiza tadzhikskikh slovoform [Conceptual Model of Automatic Morphological Analysis of Tajik Word-Forms], *Doklady Akademii nauk Respubliki Tadzhikistan*, vol. 57, no. 3, pp. 205–209, 2014. [In Rus].

[15] Khudojberdiev Kh.A. Ob algoritme proverki orfografii na primere tadzhikskogo jazyka [On the Spell Checking Algorithm Using the Tajik Language As an Example], *Politekhnicheskij vestnik. Serija Intellekt. Innovatsii. Investitsii*, no. 3 (55), pp. 58–63, 2021. [In Rus].

[16] Khudojberdiev Kh.A., Nazarov A.A., Ashurova Sh.N. Formirovanie elektronnogo slovarja dlja sistemy avtomaticheskogo perevoda teksta s tadzhikskogo jazyka na russkij [Forming an Electronic Dictionary for the System of Automatic Translation of Text from Tajik into Russian]. In *Informatsionnyj obmen v mezhdistsiplinarnykh issledovanijakh II*: stat'i Vseros. nauch.-prakt. konf. s mezhdunar. uchastiem. Rjazan', Akad. FSIN Rossii, 2023, pp. 227–231. [In Rus].

[17] Khudojberdiev Kh.A., Soliev O.M. Lingvisticheskij tezaurus tadzhikskogo jazyka [Linguistic Thesaurus of the Tajik Language], *Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh*, no. 20, pp. 103–105, 2017. [In Rus].

[18] Soliev O.M.O., Khudojberdiev Kh.A., Dovudov G.M. Sistema avtomaticheskoj proverki orfografii tadzhikskogo jazyka – TajSpell [Automatic Spell Checking System for the Tajik Language – TajSpell], *Vestnik Tekhnologicheskogo universiteta Tadzhikistana*, no. 3 (46), pp. 188–194, 2021. [In Rus].

[19] Khudojberdiev Kh.A. O sintezatore tadzhikskoj rechi po tekstu [On a Tajik Speech Synthesizer from Text]. *Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh*, no. 16, pp. 273–276, 2013. [In Rus].

[20] Usmanov Z.D., Kosimov A.A. *Razrabotka programmnogo kompleksa dlja raspoznavanija avtora neznakomogo teksta* [Development of a Software Package for Recognizing the Author of an Text]. Dushanbe, Donish, 2022, 105 pp. [In Rus].

[21] Kosimov A.A. Matematicheskij metod opisanija nejronnykh setej dlja klassifikatsii svojstv [Mathematical Method for Describing Neural Networks for Classifying Properties], *Politekhnicheskij vestnik. Serija: Intellekt. Innovatsii. Investitsii*, no. 3 (71), pp. 63–67, 2025. [In Rus].

[22] Kosimov A.A., Shamsov S.M. Takhijai algoritmi khulosabarorj baroi modelkho (amsilakho) oidi ba tartibi durust ovardani chumlai sodai tochikj khangomi navishti nodurust [Developing an Inference Algorithm for Models (Examples) to Correct a Simple Tajik Sentence When It Is Misspelled] , *Politekhnicheskij vestnik. Serija: Intellekt. Innovatsii. Investitsii*, no. 3 (71), pp. 68–72, 2025. [In Taj].

[23] Istamkulov Kh., Muzafarov D. Metody tokenizatsii teksta na tadzhikskom jazyke s pomosch'ju jazyka Python [Methods of Text Tokenization in the Tajik Language Using the Python Language], *Sovremennaja*

nauka: aktual'nye problemy teorii i praktiki. Serija: Estestvennye i tekhnicheskie nauki, no. 4, pp. 78–82, 2023. [In Rus].

[24] Niyazmukhammadov B., Buzurg-zoda L. *Morfologiya tadzhikskogo yazyka* [Morphology of the Tajik language. In Tajik language]. Stalinabad, Tadzhikgosizdat, 1941, 325 pp. [In Rus].

[25] Atamova S.M. Sootnesennost' form vido-vremennoj sistemy tadzhikskogo glagola s formami vremeni russkogo glagola [Correlation of the Forms of the Aspect-Tense System of the Tajik Verb with the Tense Forms of the Russian verb], *Uchenye zapiski Khudzhandskogo gosudarstvennogo universiteta im. akademika B. Gafurova. Gumanitarnye nauki*, no. 2 (47), pp. 182–187, 2016. [In Rus].

[26] Arzumanov S.D., Jalolov O.D. *Uchebnik tadzhikskogo yazyka dlya vuzov* [Tajik Language Textbook for Universities]. Dushanbe, Irfon, 1969, 320 pp. [In Rus].

[27] Arzumanov S.D., Sanginov A. *Tadzhikskij jazyk* [The Tajik Language]. Dushanbe, Maorif, 1988. 416 pp. [In Rus].

[28] Sorokina M.A. Razrjady imen prilagatel'nykh v tadzhikskom, anglijskom i russkom jazykakh [Classes of adjectives in the Tajik, English and Russian languages], *Vestnik Pedagogicheskogo universiteta*, no. 6-1 (67), pp. 117–131, 2015. [In Rus].

[29] Madibragimov N.Sh., Prutzkow A.V. Klassifikatsija suschestvitel'nykh tadzhikskogo jazyka dlja avtomaticheskoj obrabotki tekstov [Classification of Nouns of the Tajik Language for Natural Language Processing], *Prikaspijskij zhurnal: upravlenie i vysokie tekhnologii*, no. 4, pp. 39–52, 2020. [In Rus].

[30] Madibragimov N.Sh. Osobennosti mashinnogo morfologicheskogo analiza i sinteza glagolov tadzhikskogo jazyka [Features of Computer Morphological Analysis and Synthesis of verbs of the Tajik language], *International Journal of Open Information Technologies*, vol. 11, no. 1, pp. 79–86, 2023. [In Rus].

[31] Madibragimov N.Sh., Prutzkow A.V. Tipy prilagatel'nykh i mestoimenij tadzhikskogo jazyka i ikh ispol'zovanie dlja generatsii slovoform [Types of Adjectives and Pronouns of the Tajik Language and Their Use to Generate Word-Forms], *International Journal of Open Information Technologies*, vol. 9, no 11, pp. 85–89, 2021. [In Rus].

About the authors

Navruz Madibragimov is with the Ryazan State Radio Engineering University, 390005, Gagarin str., 59/1, Ryazan, Russia (e-mail: navruzmadibragimov@gmail.com);

Alexander Prutzkow is with the Ryazan State Radio Engineering University, 390005, Gagarin str., 59/1 and with Lipetsk State Pedagogical University, 398020, Lenin str., 42, Lipetsk, Russia (e-mail: mail@prutzkow.com).