ХУДЖАНДСКИЙ ПОЛИТЕХНИЧЕСКИЙ ИНСТИТУТ
ТАДЖИКСКОГО ТЕХНИЧЕСКОГО УНИВЕРСИТЕТА
ИМЕНИ АКАДЕМИКА М.С. ОСИМИ

УДК 519.25+81’322+811.222.8

На правах рукописи

Довудов Гулшан Мирбахоевич

КОМПЬЮТЕРНЫЙ МОРФОЛОГИЧЕСКИЙ АНАЛИЗ ТАДЖИКСКИХ СЛОВОФОРМ

А В Т О Р Е Ф Е Р А Т
диссертации на соискание учёной степени кандидата технических наук
по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Душанбе – 2018

Научная работа выполнена в Политехническом институте Таджикского технического университета имени академика М.С. Осими.

Научный руководитель: Усманов Зафар Джураевич,
доктор физико-математических наук, академик АН РТ, профессор, заведующий отделом математического моделирования Института математики АН РТ

Официальные оппоненты: Мирзоахмедов Фахриддин,
доктор технических наук, профессор, заведующий отделом инновационного развития науки Центра инновационного развития науки и новых технологий АН РТ

Зарипов Саидахмад Асрорович,
кандидат физико-математических наук, и.о. доцента кафедры «Программирования и компьютерной инженерии» Технологического университета Таджикистана

Оппонирующая организация: Межгосударственное образовательное учреждение высшего профессионального образования «Российско-Таджикский (Славянский) университет»

Защита состоится 06 июля 2018 г. в 14:00 часов на заседании диссертационного совета 6D.KOA-032 при Таджикском техническом университете имени академика М.С. Осими, г. Душанбе, проспект академиков Раджабовых, 10.

С диссертацией можно ознакомиться в библиотеке Таджикского технического университета имени академика М.С. Осими и на официальном сайте университета: http://ttu.tj/ru/2018/04/06/dovudov_g_m/

Автореферат разослан «____» _____ 2018 года

Отзывы на автореферат в двух экземплярах, заверенные печатью учереждения, просим направлять по адресу: 734042, г.Душанбе, пр. акад. Раджабовых, 10, тел.: (+992 37) 227-37-81, e-mail: saidaliev.ss@mail.ru

Ученый секретарь
диссертационного совета  Ш.С. Саидалиев

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Автоматизация обработки текстовой информации относится к числу главных проблем современной цивилизации. В своей основе она опирается на морфологический анализ слов, который используется в компьютерном переводе, проверке орфографии, анализе и синтезе речи, диалоге с компьютером, индексировании, аннотировании, реферировании, классификации, рубрикации документов, извлечении ключевых слов и во многом другом.

Актуальность развертывания исследований по автоматизации морфологического анализа (*МА=морфоанализ,*) слов таджикского языка подчеркнута в Постановлении Правительства Республики Таджикистан «Об утверждении программы применения и развития информационных технологий в таджикском языке» от 06.06.2005 № 188. Реферируемая диссертация направлена на решение именно этой проблемы. В ней используются работы Ниёзмуҳаммадова Б., Рустамова Ш., Ниязи Ш., Розенфельда А.З., Расторгуевой В.С., Камолиддинова Б. Амоновой Ф.Р., Ализода С., Мирзоева Г.,Таджиева Д.Т. и других в области морфологии таджикского языка. В ней также учтены пионерские исследования Усманова З.Д., Исмоилова М.А., Исмоиловой Р.М., Абдуллаева Ф.А., Назарова Р.С. и Гращенко Л.А. по изучению отдельных задач проблемы автоматизации морфоанализа таджикских словоформ. Одной из причин, по которой первопроходцам не удалось продвинуться к окончательному решению проблемы, явилось отсутствие достаточно представительной базы морфов.

Воспользоваться опытом иранских компьютерных лингвистов по решению аналогичной проблемы оказалось не возможно. Несмотря на родственность таджикского и персидского языков, серьезной преградой тому явилось принципиальное различие в графических системах письма: у иранцев – на основе арабской графики, у таджиков - на основе расширенного кириллического алфавита.

Грамматика таджикского языка смоделирована по подобию русской грамматики. Поэтому в диссертации используется опыт решения проблемы для русского языка, представленный в работах Белоногова Г.Г., Мальковского М.Г., Андреева А.М., Демьянкова В.З., Шуклина Д.Е. и Ножова И.М.

Связь работы с научными программами. Диссертационная работа выполнялась в рамках государственной программы «Разработка и исследование математических моделей для решения прикладных и практических задач» ГР 0116TJ00533 в составе раздела «Разработка математических моделей, алгоритмов и программ для автоматизации обработки текстовой информации».

Цель и задачи исследования – алгоритмизировать процесс морфолоанализа таджикских словоформ и реализовать его в виде программного комплекса.

Для достижения цели решаются следующие задачи:

1. сформировать коллекцию текстов таджикского языка;

2. создать полуавтоматическую итеративную процедуру, названною *морфо-распознавателем*, для вычленения корней и аффиксов из словоформ коллекции текстов и сформировать по возможности наиболее полную их базу;

3. разработать структурные модели таджикских словоформ; предложить кодирование их на основе системного описания словоизменительных категорий и граммем частей речи таджикского языка;

4. разработать алгоритмическое обеспечение автоматического МА словоформ и реализовать его в виде программного комплекса.

Объект исследования – грамматические закономерности таджикского языка.

Предмет исследования – систематизация статистических закономерностей образования таджикских словоформ для целей автоматизации МА.

Методы исследования. Обоснованность результатов, полученных в диссертации, базируется на развитии и применении методов:

- комбинаторно-статистических и итерационных вычислительных процедур обработки коллекции текстов с целью разложения словоформы на морфы и выявления и формирования базы морфов;

- теории множеств, системного анализа и кодирования для классификации типов таджикских аффиксов и словоформ;

- математического моделирования для разработки алгоритмического обеспечения процесса морфологического анализа;

- математической статистики для изучения полноты баз морфов и выявления статистических закономерностей;

- объектно-ориентированного программирования для разработки программных средств.

Научная новизна. Основные результаты диссертации являются новыми и заключаются в следующем:

- путем обработки коллекции текстов объемом в 59 344 883 словоупотреблений, сформирована обширная база морфов таджикского языка, содержащая 81 префикс, 76 539 корней и 128 760 постфиксов. Статистическими методами показано, что состав префиксов – окончательный, состав постфиксов в дальнейшем может несколько расшириться, а база корней необозримо далека от своего предельного значения;

- с учетом специфики таджикского языка предложена классификация типов аффиксов (словоизменительных, словообразовательных и словосочетательных) и соответствующая ей аналогичная классификация словоформ;

- разработано позиционное кодирование таджикских словоформ;

- разработано эквивалентное представление словосочетательных словоформ фрагментами предложения;

- разработано алгоритмическое обеспечение автоматического морфологического анализа таджикских словоформ.

- разработан комплекс программ автоматического МА таджикских словоформ.

Теоретическая ценность диссертации состоит в том, что разработанные в ней методы морфологического анализа словообразовательных и словосочетательных словоформ, путем специального преобразования их к словоизменительным словоформам, могут быть использованы, прежде всего, для языков иранской группы.

Практическая ценность работы. Разработанный в диссертации компьютерный морфологический анализатор зарегистрирован Национальным патентно-информационным центром Министерства экономического развития и торговли Республики Таджикистан (МЭРиТ РТ) в качестве информационного ресурса под индексом ЗИ-03.2.220TJ от 20.12.2011 года. Он предоставляет широкие возможности для решения самых разнообразных проблем автоматической обработки текстов на таджикском языке.

В частности, на основе предложенного в диссертации морфораспознавателя созданы языковые пакеты для проверки таджикской орфографии в OpenOfficeOrg и Microsoft Office. Они зарегистрированы в качестве информационных ресурсов под индексами ЗИ-03.2.222TJ от 11.01.2012 г. и № 4201200235 от 04.10.2012 г. соответственно. Эти пакеты получили широкое применение в практической деятельности организаций и учреждений Республики Таджикистан.

С помощью программного комплекса автоматического МА сформирован грамматический словарь словоизменений основ для 243758 таджикских словоформ.

В свою очередь, этот грамматический словарь использован для морофологической разметки корпуса таджикского языка размером более 100 млн. токенов (ske.fi.muni.cz/open/), смотрите [146, 148, 149].

Положения, выносимые на защиту:

- создание *морфораспознавателя* - полуавтоматической итеративной процедуры для определения корней и аффиксов словоформ таджикского языка;
- формирование базы морфов таджикского языка;
- классификации типов аффиксов и словоформ таджикского языка;
- разработка позиционного кодирования таджикских словоформ;
- эквивалентное представление словосочетательных словоформ фрагментами предложения;
- создание автоматического *морфологического анализатора* (МА) таджикских словоформ.

Достоверность и обоснованность главного результата диссертации - создание *автоматического морфологического анализатора* таджикских словоформ - гарантированы

*применением и развитием общепризнанного высокоэффективного гибридного метода проведения лингвистических исследований;

** подтверждением вычислительными экспериментами его высокой, на

уровне 92,02 %, способностью правильно разделять таджикские словоформы на морфы.

Личный вклад автора. Постановка задач осуществлялась совместно с научным руководителем. Основные результаты диссертационной работы получены автором самостоятельно.

Апробация результатов работы. Эффективность морфоанализатора обсуждалась на научных конференциях:

- “Recent Advances in Slavonic Natural Language Processing”. Университет Масарик, г. Брно, Чехия 2010-2012;

- “Прикладные информационные системы: проблемы моделирования, разработки и применения в развивающихся странах”. Худжанд, 29-30 июня 2012;

- “Language Technology for Normalisation of Less-Resourced Languages SALTMIL 8 - AFLAT 2012”. Стамбул, Турция, 22.05.2012;

- на научно-исследовательских семинарах по компьютерной лингвистике при Математическом институте АН РТ, Технологическом университете Таджикистана и Российско-Таджикском Славянском Университете в 2013-2017 гг.

- на международной научно-практической конференции «Роль ИКТ в инновационном развитии Республики Таджикистан». Душанбе, 17-18 ноября 2017;

Основные публикации. По теме диссертации опубликовано 19 работ: 2 монографии, [138, 139], 13 статей, [140-152] и 4 свидетельства о государственной регистрации информационного ресурса, [153-156]. Из них 10 наименований в изданиях, рекомендованных ВАК при Президенте Республики Таджикистан.

Структура и объем диссертации. Диссертация состоит из введения, пяти глав, заключения и списка литературы из 156 наименований. Основная часть диссертации изложена на 120 страницах. Диссертация содержит 38 таблиц и 41 рисунков.

СОДЕРЖАНИЕ ДИССЕРТАЦИИ[1]

Во введении изложены основные структурные элементы текста диссертации в соответствии с ГОСТ Р 7.0.11-2011.

Глава 1 посвящена построению с помощью морфораспознавателя (MR), по возможности, наиболее полных баз морфов на основе обработки текстового файла. *MR, обозначаемый через $MR(n_1, n_2, n_3)$, – это заключённые в единое целое, с одной стороны, полуавтоматическая итеративная процедура, предназначенная для формирования базы морфов на основе обработки текстового файла и, с другой, – "ручная экспертиза" разделения словоформы на морфы*. Числа n_1, n_2 и n_3 показывают, какие количества соответственно префиксов, корней и постфиксов уже содержатся среди данных MR перед выполнением очередной итерации.

Автоматическая часть распознавания возлагается на компьютер. По заданному алгоритму компьютерная программа пытается самостоятельно разделить слово на морфы. В прямой зависимости от того, какая фиксированная база морфов используется, на выходе автоматического MR *произвольное слово либо будет разделено на три части, либо такое разделение окажется невозможным*.

Распознавание морфов слова может завершиться отказом. Так как мы считаем, что анализируемое слово написано правильно, то появление отказа объясняется тем, что в базах данных отсутствует либо соответствующий корень, либо соответствующий аффикс. В таких случаях слово отправляется на экспертный анализ для того, чтобы уже "вручную" разделить слово на морфы.

Часть работы – анализ слов, не поддавшихся автоматическому разделению, возлагается на эксперта. Эксперт-лингвист, прежде всего, определяет принадлежность слова таджикскому языку. Если так, то он производит разделение слова на морфы, которые в дальнейшем используются для расширения исходных баз данных путем добавления к ним новых морфов. Если же слово оказывается иностранного происхождения, то оно пропускается (далее не анализируется)

Для проведения статистических исследований в § 1.2 сформирована коллекция текстовых материалов на таджикском языке размером в 59 344 883 словоупотреблений (из них 273 734 словоформ).

Для формирования базы префиксов в §§ 1.3 и 1.4 диссертации предложен *комбинаторно-статистический (К-С) метод*, суть которого состоит в следующем. Множество префиксов $\{PR\}$ таджикского языка представляется в виде:

$$\{PR\} = \bigcup_{k=1}^{k=3} \{PR(k)\}, \qquad (1.2)$$

то есть как теоретико-множественное объединение трех подмножеств $\{PR(k)\}$, $k = 1, 2, 3$. Элементами подмножества $\{PR(1)\}$ являются *односложные (простые)*

[1] В автореферате используются нумерация параграфов, формул, таблиц, рисунков и т.п. в соответствии с обозначениями, принятыми в диссертации.

префиксы, а подмножеств $\{PR(2)\}$ и $\{PR(3)\}$ − двойные и тройные префиксы, которые назовем *составными префиксами*.

Примем следующие предположения:

− *список, составленный из простых префиксов* ***ба, бар, бе, би, бо, боз, бу, во, дар, ма, ме, на, но, то, фар, фур, ҳам, ҳаме, ҳа*** является исчерпывающим, то есть кроме этих 19 префиксов других простых префиксов в таджикском языке нет;

− *любой* ***составной префикс*** *является сочетанием различных простых префиксов*: двух – для двойных и трех – для тройных, при этом среди двойных и тройных префиксов нет повторяющихся простых префиксов.

Подмножества $\{PR(2)\}$ и $\{PR(3)\}$ могут содержать, соответственно, не более 342 (= 19·18) двойных и 5814 (= 19·18·17) тройных префиксов. Полученные формальным образом элементы подмножеств $\{PR(2)\}$ и $\{PR(3)\}$ названы ***виртуальными префиксами***. Их полные списки используются для последующего различения и составления списков *реальных* и *фиктивных* составных префиксов.

Далее с помощью компьютерной программы с использованием подмножества $\{PR(1)\}$ простых и подмножеств $\{PR(2)\}$ и $\{PR(3)\}$ виртуальных префиксов каждая словоформа исследовалась на предмет возможности выделения в ней реального префикса, при этом использовалось следующее

Правило. *Префикс признавался таковым, если после его отделения от словоформы оставшаяся цепочка букв представляла собой слово.*

Процедура распознавания префикса выполнялась экспертом “вручную”. В результате обработки коллекции текстов с помощью MR выявлены 53 двусложных (барма-, барме-, барна-, ..., ҳамефар-, ҳамефур-) и 9 трехсложных (барнаме-, бознаме-, вонаме-, дарнаби-, дарнаме-, намебар-, намедар-, намефар-, намефур-) префиксов с их частотностями. Итак, К-С методом выявлены *81 реальных префиксов* таджикского языка (19+53+9).

Полученная база префиксов позволила в § 1.4.1 установить *распределение префиксов по длине, по начальным биграммам, с фиксированной первой буквой и с фиксированными биграммами*.

База постфиксов. К-С метод, успешный при построении базы префиксов, оказался непригодным для формирования базы постфиксов (число простых постфиксов – 113, максимальный уровень сложности – 8, потребность компьютерной памяти для хранения комбинаций из 2-х, 3-х, … и 8 простых постфиксов оценивается величиной порядка $4.16\cdot10^{8}$ Гб). В этой связи были применены *итерационные процедуры* (пошаговое расширение базы постфиксов) для обработки поначалу небольшой коллекции текстов объёмом в 3800 страниц с общим количеством в 1 540 019 словоупотреблений (§ 1.5.). Такое расширение начал выполнять $MR(66, 6000, 270)$, база данных которого содержала 66 префиксов (окончательное число 81 префиксов было получено позднее), 6000 корней и 270 постфиксов. В число последних вошли 113 простых постфиксов: **ӣ, ад, ам, амон, анд, аст, ат,**

атон, аш, ашон, ву, ед, ем, етон, и, ро, ӣ, ст, у, ю, яд, ям, ямон, янд, яст, ят, ятон, яш, яшон, а, ак, ан, анда, андар, бон, бор, вӣ, ванд, вар, ваш, ввум, вода, вон, вор, вот, вум, гӣ, гар, гаҳ, гин, гон, гона, гор, гоҳ, гун, гуна, дон, е, ё, ев, ева, евич, евна, ён, ёнд, зор, ид, ин, ина, истон, иш, када, лох, мӣ, манд, монанд, навард, но, нок, о, ов, ова, ович, овна, ок, он, она, онд, ор, осо, от, пона, сон, сор, стон, тар, то, тоб, ум, фом, ҳак, ҳо, хон, чӣ, ча, чак, ҷон, ҷот, юм, шан, як, ян, янда, а также некоторые уже известные сложные постфиксы.

После 10 итераций, в каждой из которой обрабатывалось примерно 150000 словоупотреблений, число постфиксов возросло до 2533 и число корней до 25323. Число префиксов не изменилось. Получена новая версия $MR(66, 25323, 2533)$.

Парадигмообразующие (п-о) постфиксы. С целью описания огромного многообразия постфиксов таджикского языка оказалось целесообразным представлять произвольный постфикс PS в виде конкатенации

$$PS = PS^1 \oplus PS^2 \tag{1.3}$$

суффикса PS^1 и окончания PS^2, то есть в виде п-о и п-ф постфиксов. Установлено, что п-о постфиксы PS^1 состоят не более чем из 5 простых постфиксов.

Парадигмоформирующие (п-ф) постфиксы. Дальнейший анализ показал, что из 2533 постфиксов можно выделить 127 таких, которые участвуют в формировании парадигм постфиксов, принадлежащих множеству PS^1. При анализе самого списка обнаружилось его неполнота и возможность расширения до 344 элементов за счёт п-ф постфиксов, не вошедших в число 127.

Модифицированный *MR*. Изучение структуры 2533 постфиксов последней версии *MR*, способствовало выявлению 657 элементов из PS^1 и 127 элементов из PS^2, далее расширенного до 344 элементов. На основе этих двух множеств удалось сгенерировать 46263 постфикса, признанные экспертом как реальные.

Выявленная база постфиксов оказалась, однако, неудобной для непосредственного использования в новой версии $MR(66, 25323, 46263)$, поскольку явилась причиной заметного уменьшения скорости автоматической обработки информации. В этой связи более подходящим оказалось решение о сохранении в базе постфиксов 657 элементов множества PS^1 и расширении функций MR за счёт включения в его состав модуля автоматического опознавания постфиксов PS^2, основанного на данных о 344 элементах из PS^2 и правилах присоединения простых постфиксов.

Обновлённая версия MR была обозначена через $MR^*(66, 25323, 657)$ и названа *модифицированным* MR, нацеленным на дальнейшее формирование исчерпывающей коллекции морфов таджикского языка.

Статистические закономерности множества постфиксов. Излагаемые в § 1.5.1 статистические закономерности получены на основе анализа множества постфиксов, извлечённых $MR^*(66, 25323, 657)$ из коллекции текстов размером в

55184508 словоупотреблений. Обработка информации производилась следующим образом. Исходная коллекция была разделена на 10 подколлекций объёмами в 5–5,1 млн. словоупотреблений. Подколлекции обрабатывались поочерёдно. Вначале текстовая информация преобразовывалась в частотный словарь словоформ. Затем каждая словоформа разделялась с помощью MR^* на морфы. Если процедура разделения завершалась успешно, MR^* переходил к рассмотрению следующей словоформы. В случае отказа (невозможности автоматического разделения на морфы) словоформа обосабливалась и лишь по окончании процесса обработки подколлекции наряду с другими отказными словоформами поступала на анализ экспертам, которые «вручную» осуществляли их разделение и полученные префиксы, корни и постфиксы, если они не содержались в базе, в качестве новых элементов добавляли в морфораспознаватель MR^*.

В следующей итерации обновлённая версия морфораспознавателя применялась к анализу очередной подколлекции. После выполнения 10-й итерации список простых постфиксов расширился до 125 (за счёт обнаруженных 12 новых элементов ***-акак, -ас,-вона, -вора, -гора, -ёна, -ол, -ос, -со, -ур, -ҳакак, -якак***) и был получен итоговый $MR^*(81,65422,2847)$, который позволил составить список 122 729 реальных постфиксов. Затем после обработки коллекции текстов объемом 55 184 508 словоупотреблений (240 208 различных словоформ) установлено, что более 51 % словоупотреблений и 21 % словоформ обходятся без постфиксов. Обнаружено также, что слова с постфиксами не менее шестого уровня сложности встречаются не чаще одного раза среди 100 000 слов. Отметим, что к восьмому уровню сложности отнесены нами постфиксы типа **ишмандтаринҳояшонрову** с учетом его разложения в виде **иш-манд-тар-ин-ҳо-яшон-ро-ву.**

База корней. С помощью MR^* в **§** 1.6 удалось выявить внушительный список из 65 422 корней таджикского языка.

Задача о полноте базы морфов. Для рассмотрения вопроса была привлечена выборка объемом в 877 страниц из газет и журналов (503 725 словоупотреблений и 47 224 словоформы), которая была обработана с помощью $MR^*(81,65422,2847)$.

Установлено, что 0.356 % словоупотреблений и 1.8 % словоформ не поддались автоматическому распознаванию. Причина отказов обуславливалась присутствием в текстах *слов иностранного происхождения, собственных имен и орфографических ошибок в написаниях слов*. В последнем случае отказные словоформы оправлялись на анализ эксперту, который выправлял допущенные ошибки, затем выполнял разделение словоформ «вручную». Появление за этот счёт новых корней и постфиксов составило незначительную долю от общего числа отказов. Итак, был получен более усовершенствованный образец $MR^*(81,66108,2862)$. Расширившийся до 2862 список постфиксов из множества PS^1 позволил ему распознавать уже 125 020 постфиксов.

Представление о специфике проблемы получено путем обработки случай-

ной выборки, представленной 10 извлеченными из газетных статей и журналов текстовыми файлами объемами около 305000-496000 слов. Случайный взятый первый из 10 файлов обрабатывался с помощью $MR^*(81,66108,2862)$. Обнаруженные морфы (корни, префиксы и постфиксы множества PS^1) в качестве новых элементов добавлялись в уже имеющиеся базы исходного распознавателя, который тем самым преобразовывался в более усовершенствованную версию. Далее вновь полученный MR^* применялся для обработки следующего случайно выбранного одного из 9 оставшихся файлов, и, опять-таки, новые морфы использовались для создания следующей версии MR^*.

Результаты обработки всех 10 файлов показаны в таблице 1.28. Следует обратить внимание на то, что в столбцах 9 и 12 выдаются сведения лишь о постфиксах из множества PS^1.

Таблица 1.28. - Частота появления новых морфов

№ файла	Число словоупотреблений	Число Словоформ	Число отказов	Число новых корней	однокорневых	многокорневых	Число новых префиксов	Число новых постфиксов	% префиксов	% корней	% постфиксов
1	496724	49637	800	600	150	450	0	18	0	1,208	0,036
2	424712	48762	1016	829	62	767	0	11	0	1,7	0,022
3	310511	12812	716	525	63	462	0	8	0	4,098	0,062
4	342303	11688	545	381	49	332	0	5	0	3,26	0,043
5	343562	12407	531	381	60	321	0	6	0	3,07	0,048
6	307297	13084	488	386	43	343	0	3	0	2,95	0,023
7	328127	13034	315	226	20	206	0	4	0	1,73	0,03
8	320400	13478	205	133	16	117	0	5	0	0,99	0,037
9	305674	12225	253	176	13	163	0	6	0	1,44	0,049
10	477340	37396	910	742	12	730	0	7	0	1,984	0,018
	3656650	224523	5779	4379	488	3891	0	73	0		

В таблице в 1-м столбце даются порядковые номера 10 обработанных файлов; во 2-м столбце указывается количества словоупотреблений, содержавшихся в файлах; в 3-м столбце − число различных словоформ среди общего количества проанализированных словоупотреблений; в 4-м столбце приводятся данные о числе отказов, то есть о тех случаях, в которых морфораспознавателю не удается разложить слово на морфы. Это может происходить по причине того, что слово содержит новый морф (префикс, корень или постфикс) или же является многокорневым, на анализ которого изначально не настроен морфораспознаватель. Именно такие отказные слова передаются эксперту-лингвисту, который, разделяя слова на морфы, заполняет столбцы 5 − 9 результатами своей работы. В столбцах 10 – 12 приводятся выраженные в процентах доли вновь выявленных морфов по отношению к числу различных словоформ, содержавшихся в тестовом файле. Последняя, 11-я строка таблицы информирует об окончательных результатах обра-

ботки случайной выборки общим объемом в 3 656 650 словоупотреблений. Из этого количества слов не удалось извлечь ни одного нового префикса, зато объявилось 4379 новых корней и 73 новых PS^1 - постфиксов. В итоге была получена новая версия $MR^*(81, 70487, 2935)$, контролирующая 128 760 постфиксов.

Итак, исходная база из 81 префикса не изменилась. В этой связи сделано предположение о том, что база префиксов *достаточно полная*. Что касается вновь выявляемых PS^1 - постфиксов, то по мере увеличения объема обрабатываемых текстов их число не проявляет тенденции к сходимости, см. столбцы 9 и 12. Относительно корней картина аналогичная: столбцы 5 и 11 не позволяют получить хотя бы приближенную оценку мощности словаря корней, обеспечивающих полноту автоматического морфораспознавания таджикских слов.

Результаты, излагаемые далее, относятся, прежде всего, к однокорневым словам, которые могут принадлежать одной из четырех возможных структур R, $PR \oplus R$, $PR \oplus R \oplus PS$, $R \oplus PS$, обозначающих, что слово состоит соответственно из: одного корня R; префикса PR и корня R; префикса PR, корня R и постфикса PS и, наконец, корня R и постфикса PS.

Таблица 1.31. - Распределение частот встречаемости словных структур в %

Число	R	$PR \oplus R$	$PR \oplus R \oplus PS$	$R \oplus PS$	Всего
Словоформ	18,87	1,58	8,41	71,14	100
Словоупотреблений	49,99	1,07	4,27	44,67	100

Из этой таблицы следует, что подавляющая масса таджикских словоформ (94.66 % - среди общего числа словоупотреблений и 90.01% - среди всех словоформ) имеет беспрефиксную структуру.

В **главе 2** формируются необходимые представления о структуре таджикской словоформы и наборе её граммем, характеризующих конкретные значения грамматических категорий словоформы, что составляет основу для развития алгоритмического обеспечения морфоанализа.

Описанный в главе 1 морфораспознаватель позволяет разложить любую словоформу на корень, основу и аффиксы (префиксы и постфиксы). Однако этого оказалось не достаточно: для целей морфоанализа требуется знание частей речи корня и основы, а для самой словоформы – соответствующий ей набор граммем.

Объективные потребности морфоанализа обусловили необходимость использования вместо десяти общепринятых частей речи ***девять самостоятельных частей речи*** таких, как имя существительное (*исм*), имя прилагательное (*сифат*), имя числительное (*шумора*), местоимение (*ҷонишин*), глагол (*феъл*), инфинитив (*масдар*), причастие (*сифати феълӣ*), деепричастие (*феъли ҳол*), наречие (*зарф*), и ***семь служебных частей речи*** таких, как предлог (*пешоянд*), послелог (*пасоянд*), союз (*пайвандак*), частица (*ҳиссача*), междометие (*нидо*), звукоподражательное слово (*калимаи тақлиди овозӣ*), нумератив. В главе 2 даётся систематическое

описание частей речи с их словоизменительными грамматическими категориями, а последних – с их списком допустимых грамматических значений.

Самостоятельные части речи имеют развитую систему грамматических категорий. Каждая категория представляется совокупностью её граммем, то есть словоизменительных значений. Граммемы выражаются посредством словоизменительных морфем, которые, будучи присоединенными к корню (основе) слова, образуют словоформы. Применяемое здесь понятие "словоизменительный" означает, что и корень (основа) слова и построенные из него словоформы принадлежат одной и той же части речи. Иными словами, часть речи инвариантна относительно присоединения к корню словоизменительного аффикса.

Семь других частей речи (предлог, послелог, союз, частица, междометие, звукоподражательное слово и нумератив) – служебные. Среди них последние две мы также относим к словоизменительным частям речи. Остальные пять частей речи не имеют словоизменительных категорий[2]. В словаре (словнике) и текстах (в словоупотреблениях) отдельные слова, принадлежащие к таким частям речи, встречаются в неизменном виде за редким исключением. По этой причине их анализ оказывается тривиальным и сводится к идентификации с соответственными элементами в базе данных. В связи со сказанным в этой главе, мы рассматриваем словоизменительные категории только одиннадцати отмеченных частей речи.

Во всех последующих параграфах настоящей главы, описывая грамматические категории частей речи, мы, по существу, излагаем порядок присоединения к *нормальной форме слова* словоизменительных аффиксов, что приводит к построению словоформ, названных *словоизменительными*. ***При этом, в силу специфики таджикского языка, в качестве нормальной формы слова выступают либо его корень, либо его основа.*** Из этих двух понятий более общим является основа, что она и будет использоваться в качестве нормальной формы слова.

В этой главе вводится кодирование словоформ из различных частей речи. Кодирование – позиционное. Первые две позиции кода словоформы отводится для распознавания части речи, так что для существительного – 01, прилагательного – 02, числительного – 03, местоимения – 04, глагола – 05, инфинитива – 06, причастия – 07, деепричастия – 08, наречия – 09, предлога – 10, послелога – 11, союза – 12, частицы – 13, междометия – 14, звукоподражательного слова – 15 и нумератива – 16. Таким образом, по первым двум цифрам кода словоформы распознаётся её принадлежность к соответствующей части речи. Последующие цифры – это коды значений грамматических признаков (грамемм).

В § 2.1 для кодирования словоформ-существительных предложен шаблон:

01	α_1	α_2	α_3	α_4

[2] Исключение составляет предлог барои, который может встречатся в тексте в виде бароям, бароят и т.д.

Символам $\alpha_k \left(k = \overline{1,4}\right)$ присваиваются коды 0 или 1, которым соответствуют определённые граммемы и реализующие их постфиксы по следующим правилам:

$$\alpha_1 = \begin{cases} 0 - \textit{единственное число}\ (\varnothing); \\ 1 - \textit{множественное число} \\ (\textit{вон}, \textit{гон}, \textit{он}(\textit{ён}), \textit{хо}, \textit{чот}). \end{cases} \qquad \alpha_3 = \begin{cases} 0 - \textit{нет изафета}\ (\varnothing); \\ 1 - \textit{есть изафет}\ (\textit{и}). \end{cases}$$

$$\alpha_2 = \begin{cases} 0 - \textit{определенность}\ (\varnothing); \\ 1 - \textit{неопределенность}\ (\textit{е}). \end{cases} \qquad \alpha_4 = \begin{cases} 0 - \textit{нет послелога}\ (\varnothing); \\ 1 - \textit{есть послелог}\ (\textit{ро}). \end{cases}$$

В этих правилах использован символ ∅, обозначающий пустой постфикс (постфикса нет). Примеры кодирования словоформ-существительных.

Пример 1. китобҳоеро = китоб ⊕ ҳо ⊕ е ⊕ ∅ ⊕ ро = 011101.

Пример 2. китобҳои = китоб ⊕ ҳо ⊕ ∅ ⊕ и ⊕ ∅ = 011010.

*Пример 3***.** китоб = китоб ⊕ ∅ ⊕ ∅ ⊕ ∅ ⊕ ∅ = 010000.

В этих примерах вначале приводится словоформа, затем она представляется расчленённой на морфемы и в конце выписывается её код. Символ ⊕ обозначает приклеивание.

В §§ 2.2 – 2.11 соответствующие шаблоны предложены для кодирования словоформ, принадлежащим другим частям речи (прилагательным, числительным, местоимениям, глаголам с основами настоящего и прошедшего времени, инфинитивам, причастиям, деепричастиям, наречиям, звукоподражательным словам и нумеративам), приводятся примеры кодирования словоформ.

В главе 1, в результате применения морфораспознавателя к полуавтоматической обработке коллекции текстов получено 79702 нормальных форм таджикских слов, из которых 50984 являются корнями и 28718 основами слов. В главе 2, по существу, описаны математические модели формирования слов из нормальных форм 16 частей речи. Модели представлены в форме позиционного кодирования словоформ, что позволяет путем присвоения кодам допустимых значений, соответствующих конкретным грамматическим значениям словоизменительных категорий, сконструировать исчерпывающие многообразия парадигм, которые можно извлечь из нормальных форм, принадлежащих различным частям речи.

В **главе 3** мы обращаемся к довольно обширной базе аффиксов таджикского языка, которые в словоформах встречаются в виде комбинаций их элементарных представителей. К ныне принятой кластеризации аффиксов на *словоизменительные* и *словообразовательные* мы присоединяем дополнительный тип – *словосочетательный*, объективно востребованный спецификой таджикского словообразования. Основное содержание главы - описание базы данных о трансформации части речи словоформы вследствие присоединения к ней простого аффикса. Полученные “в ручную” для всех основных частей речи эти результаты создали платформу для реализации автоматического морфоанализа словоформ.

Использованное ранее представление произвольной словоформы в виде $WF = PR \oplus R \oplus PS$ можно интерпретировать как отображение

$$\boldsymbol{R} \xrightarrow{f} \boldsymbol{WF} \qquad (3.1)$$

множества корней $\boldsymbol{R} = \{R\}$ в множество $\boldsymbol{WF} = \{WF\}$ словоформ таджикского языка, осуществляемое f - оператором, который выступает как:

тождественный оператор $f = E$, если $WF = R$;

префикс-оператор $f = PR$, если $WF = PR \oplus R$;

постфикс-оператор $f = PS$, если $WF = R \oplus PS$;

префикс $\wedge$ постфикс-оператор $f = PR(\cdot)PS$, если $WF = PR \oplus R \oplus PS$.

В такой интепретации WF выступает как результат конструирования словоформы путем присоединения к корню словоизменительных, словообразовательных или словосочетательных аффиксов, при этом словоформу WF естественно называть *образом* корня R, а корень R – *прообразом* словоформы WF.

Несмотря на то, что идея конструирования словоформы с целью описания её свойств представляется вполне понятной, её реализация на практике оказывается не простым делом. Главная причина кроется в громоздских структурах присоединяемых аффиксов – префиксов (до 3 уровней) и постфиксов (до 8 уровней). В этой связи исследование словоформы с фиксированным набором аффиксов развёртывается постепенно, начиная с корня, к которому присоединяется ближайший аффикс, затем к полученному фрагменту присоединяется ещё один аффикс, и т.д. вплоть до построения самой словоформы. На каждой итерации определяется часть речи фрагмента и его семантика, что позволяет получить окончательное представление о грамматических категориях и граммемах словоформы в целом.

Типы словоформ (§ 3.1). Некоторым словоформам таджикского языка не удается приписать часть речи. Например, словоформу "*бӯямат*" не удается отнести к какой-либо части речи, поскольку она по смыслу эквивалентна словосочетанию "*ман туро бӯям*". Такого типа словоформы назовём *словосочетательными*.

Статистические исследования на коллекции текстов в 59344883 словоупотреблений показали, что словосочетательные словоформы встречаются 27.25 % среди словоформ и 4.78% среди словоупотреблений. Из этих данных следует необходимость учёта третьего типа словоформ.

Определение 3.3. *Словосочетательная словоформа – результат отображения* (3.1)*, при котором образ (то есть получаемая словоформа) семантически эквивалентен словосочетанию* (*фрагменту предложения*).

Примеры: словоформа «*дидамат*» эквивалентна словосочетанию «*ман туро дидам*»; «*гуфтамаш*» - словосочетанию «*ман ба ӯ гуфтам*».

В § 3.2 объясняется система обозначений, используемая для описания результатов применения аффикс-операторов к самостоятельным частям речи. Там же вводятся коды для трёх типов словоформ и соответствующих им аффиксов (1-

словоизменительный, 2-словообразовательный, 3- словосочетательный).

В § 3.3 рассматривается результат присоединения аффиксов к прообразу, принадлежащему имени существительному, см. таблицу 3.4. В ней содержатся 5 столбцов. В 1-м столбце латинскими буквами обозначаются группы аффиксов (префиксов или постфиксов), элементы которых выписываются во 2-м столбце. В 3-м столбце указывается число аффиксов рассматриваемой группы. В 4-м столбце указывается часть речи, к которой относится образ существительного после присоединения к существительному любого из списка аффиксов столбца 2. В 5-м столбце отмечается тип полученной словоформы (тип присоединенного аффикса).

Таблица 3.4. - Присоединение аффиксов к имени существительному

№	Префикс	Число	Код части речи	Код типа словоформы
a	ҳам-, бар-,	2	01	2
b	ба-, бе-, беҳам-, бо-, но-, ноҳам-, ноба-, то-, бар-, баҳам-, ноҳам-.	11	02	2
c	бано-,	1	09	2
d	би-, ма-, ме-, на-, наме-, ҳаме-, бу-, мена-, наби-, меби-, ҳамена-, ҳамеби-, ноби-,	13	05	2
	Постфикс			
e	+ӣ, +ӣ-ву, +ӣ-ю, +ам, +ям, +ам-ӣ, +ям-ӣ, +ам-ӣ-ву…	166	ℵ	3
f	+и, +ро,	2	*01*	1
g	бон, … гар, … гор,.. истон,…манд,…навард,..сор…,	1389	*01*	2
h	ваш, … ворона, … гин, гун, … ин,… манд, … монанд, …нок…,	88	*02*	2
i	ид, онд, ёнд, онид, ёнид	5	*05*	2
j	идан, ондан, ёндан, онидан, ёнидан	5	*06*	2
k	анда, янда, ида, идагӣ, онда, ёнда, ондагӣ, ёндагӣ, онида, ёнида, онидагӣ, ёнидагӣ, иданӣ,	13	*07*	2
l	ан, ян, гарона, мандона, мандонатар, сон, сонӣ,ангӣ, янгӣ	9	*09*	2
m	ӣ, акӣ, бор, борӣ, вона, вор, ворӣ, гӣ, гинӣ, гоҳӣ, инӣ, ина, инча, манд, мандӣ, нокӣ, оӣ, онӣ, отӣ, сор, сорӣ, якӣ	22	*01/02*	2
n	истонӣ, нгӣ, стонӣ, шанӣ	4	*01/09*	2
o	вон, ... гон, ... е, ... он, ... ён, ... ҳак, ...ҷот,...	57	*01*	1

В приложениях 3.1-3.8. представлены результаты присоединения аффиксов к основным частям речи, а также к звукоподражательному слову и нумеративу.

Глава 4 посвящена алгоритмизации процесса морфоанализа таджикских словоформ, состоящего из следующих четырёх процедур:

1. *разложение словоформы WF на морфы, что представляется в виде*

$$WF = PR \oplus R \oplus PS^1 \oplus PS^2, \qquad (4.1)$$

где R *– корень и* PR *– префикс. Суффикс* PS^1 *и окончание* PS^2 *выступают в качестве парадигмообразующего и парадигмоформирующего постфиксов;*

2. *определение типа заданной словоформы;*

3. *распознавание части речи и граммем корня R;*

4.1. определение основы, части речи и граммем словоформы, если она является словоизменительной или словообразовательной;

4.2. представление словоформы в виде сочетания словоформ, если она является словосочетательной, и определение основы, части речи и граммем каждой словоформы.

Разработанный в главе 1 морфораспознаватель позволяет выполнить первые два пункта морфоанализа. Существенную часть системы морфоанализа составляют алгоритмы обработки словоформ, образованных из корней тех или иных частей речи. Эти алгоритмы, несмотря на их разнонаправленность, демонстрируют единый порядок анализа словоформ различных типов, суть которого поясняется на рисунке 4.1.

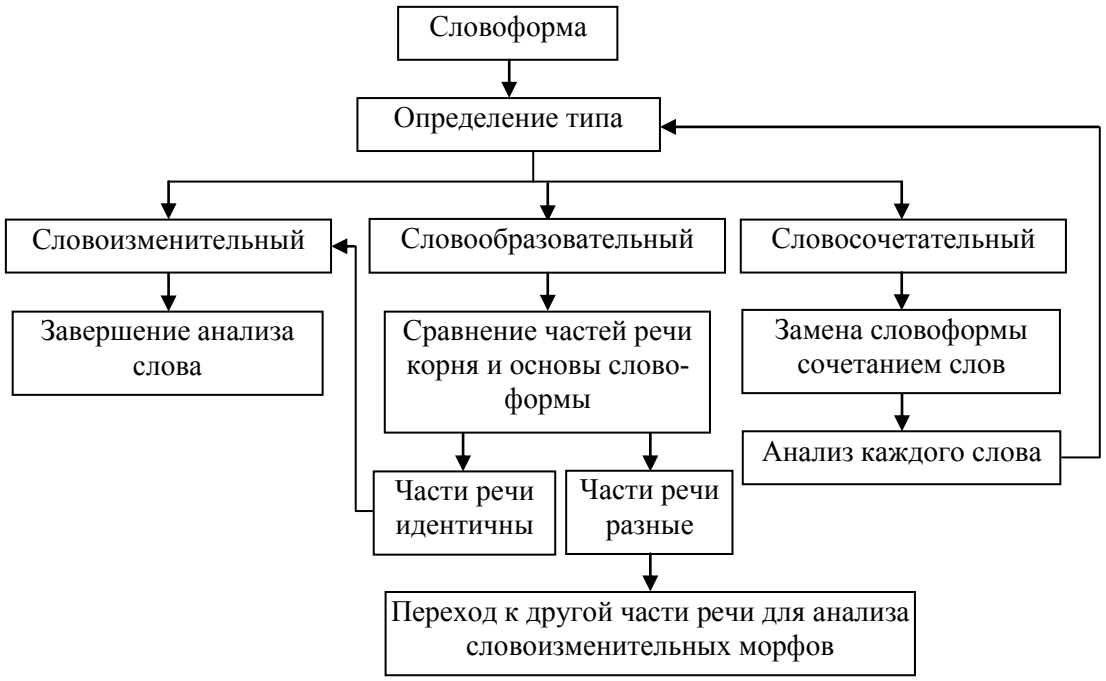

Рисунок 4.1. - Порядок обработки словоформ

Сходство агоритмов обнаруживается в обработке разных типов словоформ. Если *WF* – словоизменительная, то её анализ завершается полностью в рамках алгоритма, предназначенного для обработки словоформы из данной части речи.

Если словоформа – словосочетательная, то она заменяется на сочетание слов и последующая обработка сводится к морфоанализу каждого слова в отдельности.

Наконец, если словоформа – словообразовательная, то производится анализ её основы, в общем случае имеющей представление $PR \oplus R \oplus PS^1$. В зависимости от результата возможны два случая:

– корень *R* и основа $PR \oplus R \oplus PS^1$ принадлежат одной части речи; тогда словоформа рассматривается как словоизменительная по отношению к своей основе,

и её анализ также завершается полностью;

– корень R и основа $PR \oplus R \oplus PS^1$ принадлежат различным частям речи; тогда алгоритм анализа словоформы с корнем, принадлежащим конкретной части речи, завершает свою работу и предоставляет дальнейший анализ другому алгоритму, который настроен на обработку словоформы с той частью речи корня, которая совпадает с частью речи основы $PR \oplus R \oplus PS^1$ изучаемой словоформы.

Для морфоанализа словоформы разработан компьютерный *морфоанализатор* в составе морфораспознавателя и 16 частных подсистем. Агрегированная модель морфоанализатора представлена на рисунке 4.2.

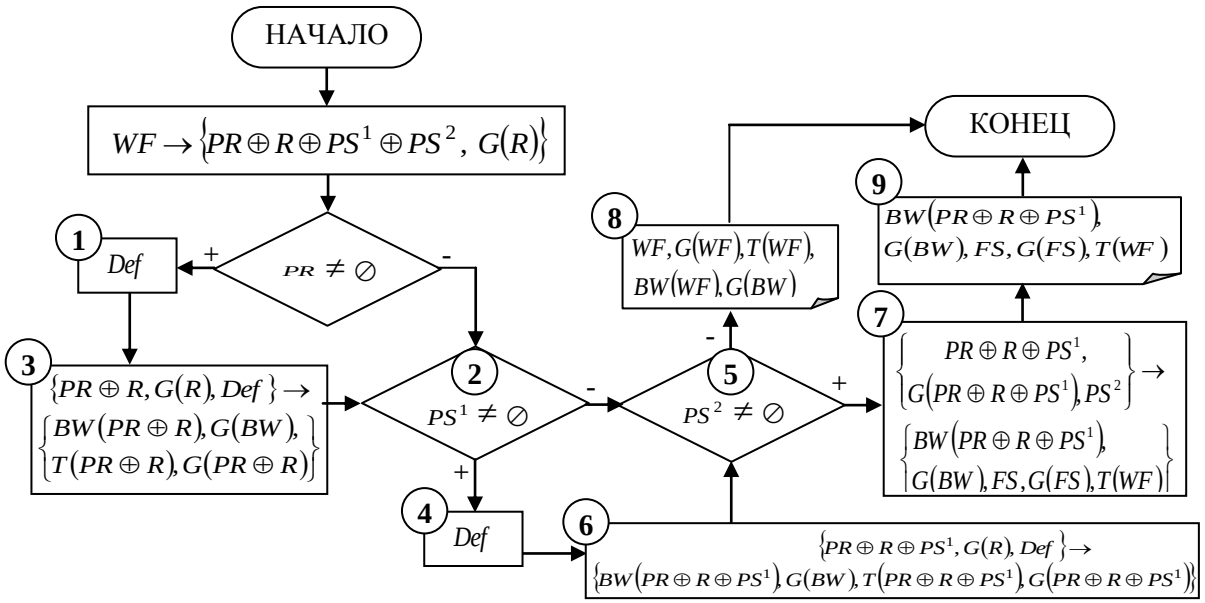

Рисунок 4.2. - Агрегированная модель морфоанализа таджикских словоформ

Её агрегированность выражается в том, что в ней одновременно заключены концепции функционирования, с одной стороны, каждой из 16 частных подсистем и, с другой стороны, системы в целом.

На этом рисунке помимо уже известных WF, PR, R, PS^1 и PS^2 приняты следующие обозначения: BW – основа словоформы, FS – словосочетание, эквивалентое по смыслу сочетанию словоформ (фрагмент предложения); $G(f)$, $BW(f)$, $T(f)$ - функции для определения соответственно граммем, основы и типа (словоизменительный, словообразовательный или словосочетательный) аргумента f, причём f может принимать значения $PR \oplus R$, R, $PR \oplus R \oplus PS^1$, $R \oplus PS^1$. Кроме того, тип префикса (постфикса) обозначен через Def. Если морф – словоизменительный, то $Def = 1$, иначе $Def = 0$.

Согласно рисунка 4.2 морфоанализ начинается с ввода словоформы *WF*. Морфораспознаватель главы 1 выполняет две процедуры задачи морфоанализа:

- раскладывает *WF* на морфы,
- определяет часть речи и список граммем корня R.

Далее решается вопрос, содержит ли анализируемая словоформа *префикс*. Если да, то выполняется операция пункта 1, если нет, то пункта 2.

В п.1 символу *Def* присваивается значение 1 или 0, является ли префикс словоизменительным или нет. После чего происходит переход к п.3, в котором по фрагменту $PR \oplus R$ словоформы *WF*, граммемам корня R и информации *Def* о типе префикса определяются основа $BW(PR \oplus R)$ фрагмента $PR \oplus R$, её граммемы $G(BW)$, тип $T(PR \oplus R)$ и граммемы $G(PR \oplus R)$ фрагмента $PR \oplus R$. Отметим, что фрагмент $PR \oplus R$ вовсе не обязан совпадать с основой. Если PR - словоизменительный, то $BW(PR \oplus R) = R$. Например, в словоформе "*мерафт*" PR = "*ме*" является словоизменительным. Следовательно, *BW(ме ⊕ рафт) = рафт*.

Если PR - словообразовательный, то $BW(PR \oplus R) = PR \oplus R$. Например, в словоформе "*бооодоб*" *PR* = "*бо*" является словообразовательным. Следовательно, *BW(бо ⊕ одоб)* = "*бооодоб*", то есть основа совпадает с фрагментом.

В п.2 при наличие *постфикса* PS^1 осуществляется переход к п.4, иначе в п.5.

В п.4 в зависимости от того, является ли постфикс словоизменительным или нет, символу *Def* присваивается значение 1 или 0. После чего происходит переход к п.6, в котором по фрагменту $PR \oplus R \oplus PS^1$ словоформы *WF*, граммемам корня *R* и информации *Def* о типе постфикса определяются основа $BW(PR \oplus R \oplus PS^1)$ фрагмента $PR \oplus R \oplus PS^1$, её граммемы $G(BW)$, тип $T(PR \oplus R \oplus PS^1)$ и граммемы $G(PR \oplus R \oplus PS^1)$ фрагмента $PR \oplus R \oplus PS^1$. Отметим, что фрагмент $PR \oplus R \oplus PS^1$ вовсе не обязан совпадать с основой. Если PS^1- словоизменительный, то $BW(PR \oplus R \oplus PS^1) = (PR \oplus R) \wedge R$ (символ $\wedge$ обозначает логическое "или"). Например, в словоформе "*ҳампешагон*" постфикс PS^1 =*«гон»* является словоизменительным. Следовательно, $BW(хам \oplus пеша \oplus гон) = хампеша$. Если PS^1- словообразовательный, то $BW(PR \oplus R \oplus PS^1) = (PR \oplus R \oplus PS^1) \wedge R \oplus PS^1$. Так в словоформе "*ҳампешагӣ*" постфикс PS^1 ="*гӣ*" является словообразовательным. Потому $BW(хам \oplus пеша \oplus гӣ) = хампешагӣ$, т. е. основа совпадает с фрагментом.

В п.5 проверяется наличие PS^2. Если есть, то переход к п.7, иначе к п.8.

В п.8. по фрагменту $PR \oplus R \oplus PS^1$ словоформы *WF*, граммемам $G(PR \oplus R \oplus PS^1)$, фрагменту $PR \oplus R \oplus PS^1$ и постфиксу PS^2 определяются основа $BW(PR \oplus R \oplus PS^1)$ фрагмента $PR \oplus R \oplus PS^1$, её граммемы $G(BW)$, тип $T(WF)$ и граммемы $G(PR \oplus R \oplus PS^1)$ фрагмента $PR \oplus R \oplus PS^1$. Далее с помощью специального алгоритма, который предложен в [9] словосочетательная словоформа разворачивается в виде сочетания словоформ FS с присоединением граммем каждой словоформы $G(FS)$.

В п.8 и п.9 представлены окончательные итоги морфоанализа словоформы.

В §§ 4.2-4.11 представлены алгоритмы морфоанализа словоформ с корнем основных частей речи. В § 4.12 приведен алгоритм для представления таджикских словосочетательных словоформ фрагментами предложений.

В **главе 5** приводится описание программного комплекса, реализующего 4 этапа морфоанализа. В агрегированном виде комплекс программ МА представлен на рисунке 5.1.

Рисунок 5.1. - Структурная схема программного комплекса

Результаты анализа словоформы выводятся на интерфейс пользователя, смотрите рисунок 5.2.

В § 5.4 тестирование программного комплекса на предмет оценки его эффек-

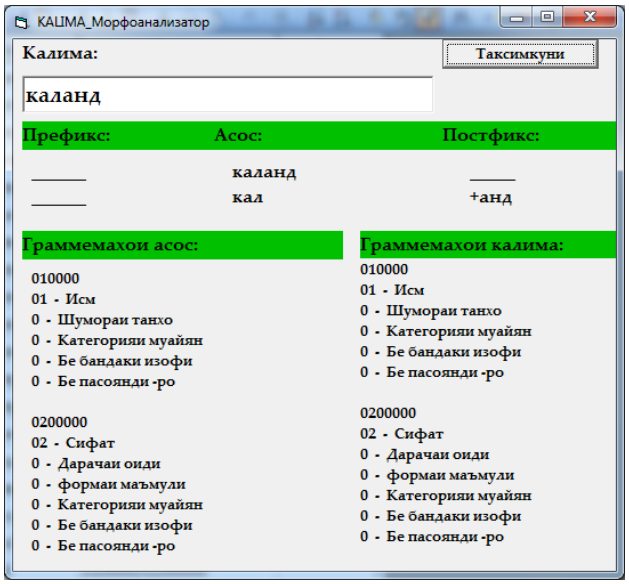

Рисунок. 5.2. - Интерфейс программного комплекса

тивности производился на случайной выборке – текстовом файле объемом в

314134 словоупотреблений. Из 41824 словоформ, предъявленных программному комплексу, успешный морфологический анализ преодолели 38487 словоформ, т.е. 92,02 % от общего количества.

Таким образом, построенный в диссертации морфоанализатор с эффективностью автоматического морфоанализа в 92,02 % можно считать в настоящее время вполне удовлетворительным для практического и теоретического применения.

ЗАКЛЮЧЕНИЕ

Основными результатами диссертации являются:

1. Создание морфораспознавателя – соединения в одно целое полуавтоматической компьютерной итеративной процедуры формирования базы морфов путём обработки текстового файла и “ручной экспертизы” разделения отказных словоформ на морфы.

2. Создание базы морфов таджикского языка из 82 префиксов, 76 539 корней и 128 760 постфиксов (посредством формирования коллекции текстов объемом в 59 344 883 словоупотреблений и применения к ней морфораспознавателя). Доказательство статистическими методами полноты базы префиксов, вероятности в дальнейшем некоторого увеличения списка постфиксов и неограниченного расширения численности корней.

3.Разработка “словоизменительно-словообразовательно-словосочетательной” классификации аффиксов и словоформ таджикского языка.

4. Разработка позиционного кодирования таджикских словоформ различных частей речи на основе системного описания словоизменительных категорий и граммем корней и основ слов.

5. Формализация правил словообразования таджикского языка.

6. Разработка комплекса программ автоматического морфологического анализа таджикских словоформ.

Результаты диссертации в сочетании с методикой Hunspell послужили основой для создания языковых пакетов проверки таджикской орфографии в OpenOfficeOrg и Microsoft Office. Они зарегистрированы в качестве интеллектуальных продуктов ЗИ-03.2.222TJ от 11.01.2012 г. и № 4201200235 от 04.10.2012г. Национальным патентно-информационным центром Министерства экономического развития и торговли РТ и получили широкое применение в работе свыше 30 организаций и учреждений Таджикистана.

Система компьютерного морфологического анализа составляет основу для разработки разнообразных задач автоматической обработки текста. Своим появлением она предоставила широкие возможности для достижения успехов в различных направлениях вычислительной лингвистики таджикского языка, при разработке которых огромный объём необходимых предварительных технических процедур берёт на себя система компьютерного морфологического анализа.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ В ИЗДАНИЯХ ИЗ ПЕРЕЧНЯ ВАК ПРИ ПРЕЗИДЕНТЕ РТ

1. Довудов, Г.М. О формировании базы префиксов таджикского литературного языка [Текст] / Г.М. Довудов, З.Д. Усманов // Доклады АН РТ. - 2009. - Том 52, №6. - С. 431-436.

2. Довудов, Г.М. О множестве постфиксов таджикского литературного языка [Текст] / Г.М. Довудов, З.Д. Усманов, О.М. Солиев // Доклады АН РТ. - 2010. - Том 53, №2. - С. 99-103.

3. Довудов, Г.М. О статистических закономерностях морфемной базы таджикского языка [Текст] / Г.М. Довудов, З.Д. Усманов // Доклады АН РТ. -2010. - Том 53, №3. - С. 188-191.

4. Довудов, Г.М. Частотный морфемный словарь таджикского литературного языка [Текст] / Г.М. Довудов, З.Д. Усманов // Доклады АН РТ. - 2010. - Том 53, №4. - С. 257-262.

5. Довудов, Г.М. Концептуальная модель автоматического морфологического анализа таджикских словоформ [Текст] / Г.М. Довудов, З.Д. Усманов // Доклады АН РТ. - 2014. - Том 57, №3. - С. 205-209.

6. Довудов Г.М. Алгоритм автоматического морфологического анализа таджикских слов [Текст] / Г.М. Довудов // Известия АН РТ. -2010. - №2 (139). - С. 22-26.

7. Довудов, Г.М. Статистика частей речи таджикского языка [Текст] / Г.М. Довудов // Известия АН РТ. - 2012. - №3 (148). - С.54-56.

8. Довудов, Г.М. Алгоритм представления таджикских словосочетательных словоформ фрагментами предложений [Текст] / Г.М. Довудов, З.Д. Усманов // Известия АН РТ. - 2013. - №4 (153). - С. 69-75.

9. Довудов, Г.М. Позиционное кодирование таджикских словоформ [Текст] / Г.М. Довудов, З.Д. Усманов // Известия АН РТ. - 2015. - №1(158). - С.58-66.

10. Довудов, Г.М. Формирование базы морфов таджикского языка [Текст]: монография / Г.М. Довудов, З.Д. Усманов. - Душанбе: Дониш, 2014. - 109 с.

11. Довудов, Г.М. Морфологический анализ словоформ таджикского языка [Текст]: монография / Г.М. Довудов, З.Д. Усманов. - Душанбе: Дониш, 2015. - 132 с.

В других изданиях опубликованы 4 свидетельства о государственной регистрации информационных ресурсов в Национальном патентно-информационном центре Министерство экономического развития и торговли РТ и 4 статьи в трудах Masaryk University, Brno и European Language Resources Association (ELRA).

АННОТАЦИЯ

диссертации Довудова Гулшана Мирбахоевича на тему «Компьютерный морфологический анализ таджикских словоформ» на соискание ученой степени кандидата технических наук по специальности 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Ключевые слова: Таджикский язык, автоматический морфологический анализ.

Объект исследования: Грамматические закономерности таджикского языка.

Цель работы: Алгоритмизировать процесс морфологического анализа таджикских словоформ и реализовать его в виде программного комплекса.

Методы исследования: Теория множеств, системный анализ, элементы теории кодирования, математическое моделирование, математическая статистики, объектно-ориентированное программирование.

Полученные результаты и их новизна: Результаты диссертации являются новыми и заключаются в следующем:

- сформирована обширная база морфов таджикского языка, содержащая 81 префикс, 76 539 корней и 128 760 постфиксов;
- предложена классификация типов аффиксов (словоизменительных, словообразовательных и словосочетательных) и аналогичная классификация словоформ;
- разработано позиционное кодирование таджикских словоформ;
- разработано эквивалентное представление словосочетательных словоформ фрагментами предложения;
- разработано алгоритмическое обеспечение морфологического анализа таджикских словоформ.

Степень использования: на основе морфологического анализатора созданы языковые пакеты для проверки таджикской орфографии в OpenOfficeOrg и Microsoft Office, зарегистрированные в виде информационных ресурсов ЗИ-03.2.222TJ от 11.01.2012 г. и № 4201200235 от 04.10.2012 г. Пакеты используются во многих организациях и учреждениях РТ.

Область применения: решение задач автоматической обработки текста, компьютерный перевод, проверка орфографии, анализ и синтез речи, диалог с компьютером, автоматизация процессов индексирования, аннотирования, реферирования, классификации и рубрикации документов, извлечения ключевых слов, морофологической разметки корпуса таджикского языка.

ШАРҲИ МУХТАСАР

ба рисолаи диссертатсионии Довудов Гулшан Мирбахоевич дар мавзӯи «Таҳлили морфологии компютерии калимаҳои забони тоҷикӣ» барои дарёфти дараҷаи илмии номзади илмҳои техникӣ аз рӯйи ихтисоси 05.13.11 - Таъминоти математикӣ ва барномавии мошинҳои ҳисоббарор, муҷтамаъҳо ва шабакаҳои компютерӣ

Калимаҳои калидӣ: забони тоҷикӣ, таҳлили автоматии морфологӣ.

Объекти таҳқиқот: қонуниятҳои грамматикии забони тоҷикӣ.

Мақсади кор: Алгоритмиронии раванди таҳлили морфологии калимаҳои забони тоҷикӣ ва татбиқи он дар намуди комплексӣ барномавӣ.

Методҳои таҳқиқот: назарияи маҷмӯъҳо, таҳлили системавӣ, элементҳои назарияи кодиронӣ, моделиронии математикӣ, омори математикӣ, барномарезии баобъектнигаронидашуда.

Натиҷаҳои бадастомада ва навгонии онҳо: Натиҷаҳои асосии диссертатсия нав буда, чунин ифода мегарданд:

- базаи бузурги морфҳои забони тоҷикӣ ташкил карда шуд, ки аз 81 префикс, 76 539 реша ва 128 760 постфикс иборат аст;
- таснифи намуди аффиксҳо (шаклсоз, калимасоз ва иборасоз) ва калимаҳо пешниҳод шуд;
- кодиронии мавқеии калимаҳои забони тоҷикӣ коркард шуд;
- алгоритми табдили калимаҳои иборасоз ба ибораҳо пешниҳод шуд;
- таъминоти алгоритмии таҳлили морфологии калимаҳои забони тоҷикӣ коркард шуд.

Аҳмияти амалии таҳқиқот: дар асоси таҳлилгари морфологӣ пакетҳои забони барои тафтиши имлоии калимаҳои забони тоҷикӣ дар OpenOfficeOrg ва Microsoft Office тартиб дода шудааст, ки ҳамчун захираҳои иттилоотии ЗИ-03.2.222TJ аз 11.01.2012 ва № 4201200235 аз 04.10.2012 ба қайд гирифта шудаанд. Пакетҳои мазкур дар аксари муассисаҳо ва корхонаҳои Ҷумҳурии Тоҷикистон истифода мешаванд.

Соҳаи истифодабарӣ: ҳалли масъалаҳои коркарди автоматии матн, тарҷумаи компютерӣ, тафтиши имло, таҳлил ва синтези нутқ, муколама бо компютер, автоматизатсияи равандҳои индексиронӣ, шарҳдиҳӣ, гуруҳбандӣ, фишурданамоӣ, ҷудо намудани калимаҳои калидӣ ва тафсили матн.

ANNOTATION

on the dissertation of Dovudov Gulshan Mirbahoevich on the theme «Computer morphological analysis of Tajik word form (slovoform)» candidate for a degree of technical sciences on a specialty

05.13.11 – Mathematical and software of computers, complexes and computer networks

Key words: Tajik, automatically morphological analysis.

Research object: Grammatical rules of Tajik.

Objectives: To algorithmize the process of morphological analysis Tajik word form (slovoform) and to implement it as a software package

Research methods: Set theory, system analysis, elements of coding theory, math modeling, math statistics, object-oriented programming.

The obtained results and novelty: The novelty of the dissertation results are:

- the wide base of Tajik morph is formed, which contains 81 prefixes, 76539 roots and 128760 postfixes;
- the classification of types of affixes (inflectional, derivational and word- combinations) is suggested and similar classification of word form (slovoform) also is given;
- the positional coding of Tajik word form (slovoform) is developed;
- an equivalent representation of word form is developed with fragments of the sentence;
- algorithmic support of the morphological analysis of the Tajik word forms is developed.

Usage degree: Based on of morphological analyzer language packs for Tajik spellchecking in OpenOfficeOrg and Microsoft Office are created, which are registered as informational resources ЗИ-03.2.222TJ from 11.01.2012 and № 4201200235 from 04.10.2012. The packs are used in many companies and offices of Tajikistan

Application: Solving the tasks of automatic text processing, computer translation, spellchecking, analysis and synthesis of speech, dialogue with computer, automation of indexing processes, annotation, abstracting, classification and rubric of documents, extracting keywords, marking of the hull of the Tajik language.

Подписано к печати 14.05.2018
Формат 6х84/16. Бумага офсетная
Тираж 100 экз. Объём 1,3 п.л. Заказ №162
Отпечатано в типографии «Мехвари дониш»
г. Худжанд, ул. Ленина, 226