ОТЗЫВ

научного руководителя о работе ДОВУДОВА Гулшана Мирбахоевича “Компьютерный морфологический анализ таджикских словоформ”, представляемой на соискание учёной степени кандидата технических наук по специальности 05.13.11 – “Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей”

Работа Г.М.Довудова посвящена автоматизации морфологического анализа слов таджикского языка – центральной проблеме автоматической обработки информации на таджикском языке. От успешного решения именно этой проблемы зависит, по существу, прогресс в вопросах автоматизации процедур компьютерного перевода, проверки орфографии, анализа и синтеза речи, диалога с компьютером, индексирования, аннотирования, реферирования, классификации, рубрикации документов, извлечения ключевых слов и во многом другом. Актуальность темы диссертации особо подтверждается в том, что она вошла в перечень первостепенных задач, записанных в Постановлении Правительства Республики Таджикистан «Об утверждении программы применения и развития информационных технологий в таджикском языке» от 06.06.2005 № 188.

Цель и задачи исследования – алгоритмизировать процесс морфологического анализа таджикских словоформ и реализовать его в виде программного комплекса. Для достижения цели решены следующие задачи:

1. сформировать коллекцию текстов таджикского языка;
2. создать базу словоформ с их частотностью;
3. разработать морфораспознаватель (полуавтоматическую итеративную процедуру) для вычленения корней и аффиксов из словоформ;
4. сформировать базу префиксов, корней и постфиксов таджикского языка;
5. предложить системное описание словоизменительных категорий и граммем частей речи таджикского языка;

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

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

- комбинаторно-статистического для разложения словоформы на морфы;
- 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 префиксов, 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.

Практическая значимость. 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 морфораспо-знавателя 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.

Структура диссертации. Работа Г.М.Довудова состоит из введения и 5 глав. Материал введения преподносится стандартным образом: вначале дан обзор исследований, имеющих отношение к работе, затем в соответствии с требованиями ВАК представлены ответы на вопросы по достижениям диссертанта.

В главе 1 **“Формирование базы морфов таджикского языка”** центральное место занимает м*орфораспознаватель* - инструмент для формирования упомянутых баз. Это суть - компьютерная программа, которая по мере увеличения объема обрабатываемых текстов помогает эксперту выявлять всё новые и новые морфы и получать статистические оценки мощности баз морфов.

В главе 2 **“Словоизменительные категории и граммемы частей речи таджикского языка. Кодирование словоформ”** на основе предварительных детальных исследований впервые для таджикского языка предложена система позиционного кодирования словоформ различных

частей речи, которая позволила путём приписывания по особым правилам корню слова соответствующего цифрового кода однозначно определять порождаемую из него словоформу. Именно это достижение способствовало алгоритмизации большого круга лингвистических процедур морфологического анализа.

Основное содержание главы 3 **“Таджикские словоформы и аффиксы”** - 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 морфологического анализа таджикских словоформ. Здесь же следует отметить предложенную кластеризацию аффиксов на *словоизменительные, словообразовательные* и присоединённый к ним новый тип – *словосочетательный*, объективно востребованный спецификой таджикского словообразования.

Глава 4 **“Морфологический анализ словоформы”** благодаря результатам предыдущих глав представляется в виде стройной системы, смысл которой наглядно развертывается с помощью концептуальной модели морфоанализа.

Глава 5 **“Программный комплекс автоматического морфологического анализа таджикских словоформ** посвящена описанию реализации 4-х этапов морфологического анализа, указанных во введении главы 4. Первые 3 из них выполняется морфораспознавателем, четвертый этап – собственно морфологическим анализатором. На случайно выбранных текстах получены статистические оценки эффективности работы морфоанализатора. Масштабность проведенной работы косвенно можно уяснить по списку компьютерных программ, вошедших в состав программного комплекса.

Публикации по теме диссертации. Список опубликованных работ Г.М.Довудова содержит 19 наименований (10 из них – в изданиях, рекомендованных ВАК Таджикистана)**:** 2 монографии и 7 статей в соавторстве с научным руководителем, 4 статьи в соавторстве с словацкими

специалистами (из них 3 имеют отношение к созданию корпуса таджикского языка и одна – к частному вопросу морфологии) и 4 свидетельств о регистрации интеллектуальных продуктов, подтверждающих их внедрение, совместно с соавторами и одна статья без соавторов.

Как научный руководитель подтверждаю весомый вклад диссертанта во все наши совместные исследования. Считаю, что за прошедшее время Г.М.Довудов существенно повысил свою научную квалификацию, поднявшись до уровня самостоятельно мыслящего, инициативного исследователя.

По моему глубокому убеждению, работа Г.М.Довудова отвечает всем требованиям ВАК как в теоретическом отношении, так и практической направленности и вполне готова к представлению в качестве научного доклада для государственной итоговой аттестации на предмет присуждения ему учёной степени кандидата технических наук по специальности 05.13.11 – "Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей"

Научный руководитель УСМАНОВ Зафар Джураевич
доктор ф.-м.н., академик АН РТ, профессор

Дата: «7» февраля 2018 г. Подпись: _____

Подпись Усманова З.Д. заверяю:

Ученый секретарь
Института математики АН РТ,
к. ф.-м. н. **Н. Назрублоев**

«08» 02 2018 г.