

РАЗРАБОТКА АЛГОРИТМА СТЕММАТИЗАЦИИ И ГЕНЕРАЦИИ СЛОВОФОРМ ДЛЯ УЗБЕКСКОГО ЯЗЫКА

Назирова Э.Ш

Ташкентский университет информационных технологий, Ташкент

Боймуродов Ф.Ф

Каршинский государственный университет, Карши

Аннотация: Алгоритмы синтеза и анализа словоформ узбекского языка разработаны на основе принципов деления существительных, прилагательных и глаголов на флективные классы исходя из поставленной задачи. Поскольку узбекский язык является агглютинативным, для автоматизации морфологического анализа нецелесообразно подключать словарь словоформ. Гораздо эффективнее использовать словарь аффиксов и набор правил. В процессе написания статьи была создана и алгоритмически реализована морфологическая модель узбекского языка. В частности, словари организованы по флективным классам существительных, прилагательных и глаголов.

Ключевые слова: генерация, стемматизация, Портера, аффикс, суффикс.

Annotation: Algorithms for the synthesis and analysis of word forms of the Uzbek language are developed on the basis of the principles of dividing nouns, adjectives and verbs into inflectional classes based on the task. Since the Uzbek language is agglutinative, it is not advisable to use a dictionary of word forms to automate morphological analysis. It is much more efficient to use an affix dictionary and a set of rules. In the process of writing the article, a morphological model of the Uzbek language was created and algorithmically implemented. In particular, the dictionaries are organized into inflectional classes of nouns, adjectives, and verbs.

Keywords: generation, stemmatization, Portera, affix, suffix.

Процесс информирования общественности становится требованием современности. Все сферы деятельности человека неразрывно связаны с процессами получения, обработки и усвоения информации. Сегодня компьютерные и информационные технологии становятся основой нашей повседневной жизни, поскольку они быстро обновляются, изменяются и развиваются. В то же время Узбекистан

постепенно входит в систему оцифровки. В связи с этим развитие компьютерной лингвистики является актуальной задачей.

Целью статьи является введение узбекского языка в компьютерную лингвистику, разработка эффективного способа усвоения основных понятий, морфологического образования узбекского языка.

Морфологическая основа узбекского языка характеризуется большим разнообразием аффиксов, добавляющихся к основе слова в определенной последовательности и придающих ему самые разные характеристики, а также огромным количеством глагольных форм, позволяющих создавать до 150-200 вариантов, отличающихся по смыслу, порой, кардинально, но при этом образованных от одного глагола. Программные средства, которые автоматически находят нужные словоформы в изучаемых текстах, являются важным помощником в проведении лингвистических исследований.

Диалект - помогает людям, живущим в определенном районе, взаимодействовать друг с другом. В узбекском языке диалект отличается от литературного языка более или менее фонетически, лексически и грамматически.

Есть только устная форма народных диалектов.

Набор тесно связанных диалектов называется диалектом (диалект - это сочетание диалектов и диалектов).

В узбекском языке три диалекта:

1. Карлукский диалект (юго-восточная группа).
2. Кипчакский диалект (юго-западная группа).
3. Огузский диалект (северо-западная группа).

Узбекский - тюркский язык карлукской группы, который относится к категории синтетических агглютинативных языков и имеет богатую и сложную морфологию. Как и в других тюркских языках, слово состоит из основы, к которой прикреплены суффиксы, обозначающие различные грамматические особенности.

Карлукский диалект в основном включает городские диалекты (Ташкентский, Андижанский, Ферганский, Бухарский). Важными фонетическими и морфологическими особенностями этих диалектов являются:

1) Буква "k" в конце слова произносится как "y" : *elak* - *elay*, *terak* - *teray*.

2) есть деление: *aka* - *oka*, *Akram* - *Akrom*.

3) в форме винительного падежей, " -ni " : *ukamni daftari* - тетрадь моего брата.

Узбекский язык включает ряд диалектов, такие диалекты с этнолингвистической точки зрения он объединен в три основных диалекта - карлукский, кипчакский и огузский. Естественно, что эти три диалекта приняли участие в формировании современного узбекского литературного языка.

При изучении языка большое значение имеет способность правильно понимать структуру слова и определять его компоненты. Слово отражает структурные особенности языка, его лексико-семантические и функционально-грамматические законы. По своей типологии и морфологической структуре узбекский язык значительно шире обычных выражений и отличается относительной регулярностью, позиционной и грамматической устойчивостью морфологической структуры различных словоформ. Слова образуются путем добавления грамматических частиц к основанию слов – аффиксов. [Барахнин, 2017, с. 435].

Результатом этой статье является электронная система для изучения и обучения узбекскому языку, которая может работать в локальных и глобальных сетях и должна выполнять следующие задачи:

- Знакомство с курсами узбекской грамматики;
- Разработать системную базу данных на основе собранных данных;
- Разработка интерфейса пользователя и менеджера;
- Разработка пользовательских и управляющих модулей;
- Организация системного управления;
- Разработка системного дизайна;

Когда мы говорим, что программа для создания слов на узбекском языке является научной новинкой, мы имеем в виду, что она была создана с использованием новых технологий, а также имеет следующие преимущества при использовании морфологических систем. Система обладает следующими удобствами и преимуществами:

- Любой пользователь системы может пользоваться системой через Интернет;
- Не требует программного обеспечения от пользователя, использующего систему;
- удобный интерфейс, позволяющий пользователям без специальной подготовки;
- Предоставить любому пользователю хотя бы некоторые навыки узбекской грамматики;

Вид приложения показан на рисунке ниже. На разделе нашего меню посвящен созданию словоформ имени существительные (Генератор существ-ных). Пользователи должны ввести слово при использовании приложения. Наша программа предлагает все формы слова, в зависимости от значения слова. В основе нашей программы - суффиксы, составляющие целую словосочетание узбекского языка. Это наиболее распространенные случаи на узбекском языке. На узбекском языке большинство имен существительных создаются в основном с использованием определенных суффиксов.

Окончания имени существительные.

Конфигурация слово образования	Формы окончания
Начальный падеж,	-
Притяжательный падеж	-ning
Винительный падеж неод	-ni
Направительный падеж	-ga
Местно-временной падеж	-da
Исходный падеж	-dan
Множественное число	-lar
1 лицо. ед.число	-im
1 лицо. мн.число	-imiz
2 лицо. ед. число	-ing
2 лицо. мн. Число	-larning
Притяжательные окончания 2-ая форма	-niki
Множественное число + Исходный падеж	-lardan
Множественное число + Местный падеж	-larda
Множественное число + Дательно-направит. Падеж	-larga
Множественное число + Родительный падеж	-larning
Множественное число + Винительный падеж	-larni
...	

Система проверяет в базе основ корень слова, полученного из предыдущего шага, если результат успешный, то алгоритм заканчивает свою работу. А если нет, то он сохраняет эту комбинацию аффиксов и возвращается к шагу №2. В шаге №2 он уже не будет использовать те аффиксы, то были сохранены в шаге №3.

После нормализации для каждого слова осуществляется вычисление его морфологических характеристик и определяется морфологический класс основы и представляет пользователю.

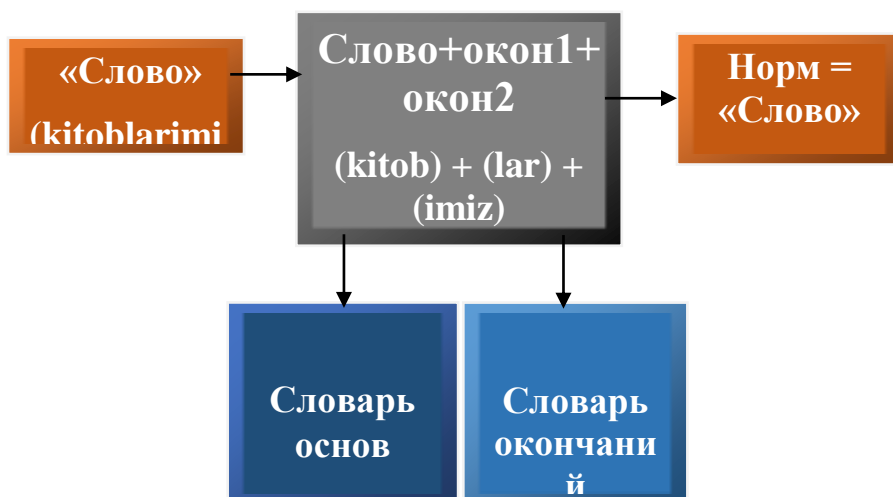


Рис: Процесс определения нормальной формы слова и его морфологических параметров.

На вход подается словоформа, происходит поиск в словарях аффиксов *-lar*, *-imiz* и основы слова - *kitob*. На основе морфологического класса основы (существительное) и аффиксов вычисляем морфологическую информацию: *-lar* «множественное число», *-imiz* «притяжательное окончание».

Пример работы приложения:

1. Генератор существительных: В этом разделе мы можем создавать словоформы существительных в узбекском языке. От пользователя требуется ввести корневую форму существительных в соответствующем месте.

давайте попробуем ввести существительное, “kitob” – “книга”:

[Генератор глаголов](#)
[Стемматизатор глаголов](#)
[Генератор существ-ных](#)
[Стемматизатор существ-ных](#)
[Прилагательные](#)

Конфигурация слово образования	Форма слова
Начальный падеж	kitob
Приказательный падеж	kitobning
Винительный падеж неод	kitobni
Направительный падеж	kitobga
Место-временной падеж	kitobda
Исходный падеж	kitobdan
Множественное число	kitoblar
1 лицо, ед. число	kitobim
1 лицо, мн. число	kitobimiz
2 лицо, ед. число	kitobing
2 лицо, мн. число	kitoblaring
2 лицо, (уз) ед. число	kitobingiz
2 лицо, (уз) мн. число	kitoblaringiz

2. *Стемматизатор существительных*: в этом разделе можно определить основную часть существительных, то есть корневую форму существительных.

давайте попробуем ввести существительное, “bolalarimizning” – “наших детей”. коренное слово, “bola” – “ребёнок”:

[Генератор глаголов](#)
[Стемматизатор глаголов](#)
[Генератор существ-ных](#)
[Стемматизатор существ-ных](#)
[Прилагательные](#)

Конфигурация слово образования	Форма слова
stem	bola

3. *Генератор глаголов*: В этом разделе мы можем создавать словоформ глагола в узбекском языке. От пользователя требуется ввести инфинитивную форму глагола в соответствующем месте:

Морфологический генератор / стемматизатор

[Генератор глаголов](#)
[Стемматизатор глаголов](#)
[Генератор существ-ных](#)
[Стемматизатор существ-ных](#)
[Прилагательные](#)

например, мы можем увидеть результаты, введя следующий пример:

[Генератор глаголов](#) | [Стемматизатор глаголов](#) | [Генератор существ-ных](#) | [Стемматизатор существ-ных](#) | [Прилагательные](#)

Конфигурация слово образования	Форма слова
Страдательное	ketma
Вопрос	ketmasi?
Результативно прошедшее время	ketgan
Конкретное настоящее время	ketgapti
Переходное время	ketadi
Прошедшее время	ketdi
Будущее продолжительное время	ketadi
Будущее время намеренно	ketmoqishi
Переходное прошедшее время	ketgan edi
Демонстративное время	kettdi
Результативно прошедшее время + Личное оконч. 1п ед.ч	ketganman
Результативно прошедшее время + Личное оконч. 2п ед.ч	ketgansan
Результативно прошедшее время + Личное оконч. 3л ед.ч	ketgansiz

4. *Стемматизатор глаголов*: в этом разделе можно определить основную часть глагольных, то есть инфинитивную форму глагола. давайте попробуем ввести глагол, “ketmoqchisiz” – “Вы хотите уйти”:

[Генератор глаголов](#) | [Стемматизатор глаголов](#) | [Генератор существ-ных](#) | [Стемматизатор существ-ных](#) | [Прилагательные](#)

Конфигурация слово образования	Форма слова
stem	ketmoq

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Барахнин В.Б., Федотов А.М., Бакиева А.М., Бакиев М.Н. Тажибаева С.Ж., Батура Т.В., Кожемякина О.Ю., Тусупов Д.А., Самбетбаева М.А., Лукпанова Л.Х. Алгоритмы генерации и стемматизации словоформ казахского языка // Cloud of Science. – 2017. – Т. 4. – № 3. – С. 434-449.

<https://cyberleninka.ru/article/n/algoritmy-generatsii-i-stemmatizatsii-slovoform-kazahskogo-yazyka/viewer>

2. Боймуродов.Ф., Algorithm development and implementation stemmatization and generation of word forms of uzbek language for automatic processing systems texts, https://www.researchgate.net/publication/347995316_ALGORITHM_DEVELOPMENT_AND_IMPLEMENTATION_STEMMATIZATION_AND_GENERATION_OF_WORD_FORMS_OF_UZBEK_LANGUAGE_FOR_AUTOMATIC_PROCESSING_SYSTEMS_TEXTS

3. Эшкараева Н.Г, Боймуродов Ф.Ф Разработка и реализация алгоритма стемматизации и генерации словоформ для глаголов

узбекского языка. http://qarshidu.uz/source/JURNAL/Aniq_fanlar_5-son_2022.pdf

4. Боймуродов.Ф., Algorithm development and implementation stemmatization and generation of word forms of uzbek language for automatic processing systems texts, https://www.researchgate.net/publication/347995316_ALGORITHM_DEVELOPMENT_AND_IMPLEMENTATION_STEMMATIZATION_AND_GENERATION_OF_WORD_FORMS_OF_UZBEK_LANGUAGE_FOR_AUTOMATIC_PROCESSING_SYSTEMS_TEXTS

5. Эшкараева Н.Г, Боймуродов Ф.Ф. Создание словоформ существительных в узбекском языке с использованием алгоритма портера. <https://scienceweb.uz/publication/14977>

6. Nazirova E.SH, Boymurodov F.F. Tabiiy tilli matnlarni avtomatik qayta ishlash tizimlari tahlili. <https://scienceweb.uz/publication/14975>

7. Боймуродов Ф.Ф. Стемматизация и генерация словоформ в узбекском языке для систем автоматической обработки текстов. https://api.scienceweb.uz/storage/publication_files/5482/14977/6561770b28541_%D0%A4.%D0%91%D0%BE%D0%B9%D0%BC%D1%83%D1%80%D0%BE%D0%B4%D0%BE%D0%B2%2014-16%20%D0%A2%D0%90%D0%A2%D0%A3.pdf