**РАЗРАБОТКА АЛГОРИТМА ПО НОРМАЛИЗАЦИИ ИМЕНИ СУЩЕСТВИТЕЛЬНЫХ НА ЯЗЫКЕ PYTHON**

**Жакыпова Ж.Ж.**

Жакыпова Жанар Жакыпкызы – магистрант,

кафедра «Информатика и информационная безопасность»,

Факультет информационных технологий,

Евразийский национальный университет имени Л.Н. Гумилева,

г.Астана, Казахстан

[z.zhakypova@gmail.com](mailto:z.zhakypova@gmail.com)

***Аннотация***

В данной статье рассматривается общий жизненный процесс обработки текстовой информации, описываются этапы по приведению текста в нормальную словоформу для применения их при реализации графематического анализа в казахском языке. Представлены правила по разбору словоформ в казахском языке, алгоритм по разбору текстовой информации и в частности, имен существительных и извлечение из них основной формы словоформы. Подробно описан процесс по нормализации имен существительных через отсечение аффиксов и нахождения корня слова. Также представлены модель образования словоформ и возможные комбинации аффиксов при образовании составных слов.

***УДК 004.912***

***Ключевые слова*:** казахский язык, стэмминг, морфологический анализ, токенизация, нормальная форма слова, аффиксы, словоформа.

**РАЗРАБОТКА АЛГОРИТМА ПО НОРМАЛИЗАЦИИ ИМЕНИ СУЩЕСТВИТЕЛЬНЫХ НА ЯЗЫКЕ PYTHON**

**Zhakypova Zh.Zh.**

Zhakypova Zhanar Zhakypkyzy – master student,

COMPUTER SCIENCE AND INFORMATION SECURITY DEPARTMENT

FACULTY OF INFORMATION TECHNOLOGY,

L.N.GUMILYOV EURASIAN NATIONAL UNIVERSITY,

ASTANA, KAZAKHSTAN

[z.zhakypova@gmail.com](mailto:z.zhakypova@gmail.com)

***Abstract***

The article considers the general life process of processing textual information, describes the steps of processing the plain text into a normal form (normalization) to use in the morphological analysis. There are presents rules for normalization word forms in the Kazakh language, an algorithm for intellectual text analysis and particularly parsing nouns and extracting the basic form of a word. The process of normalization of nouns by cutting off affixes and finding the root of the word is described in detail. Also there are considered the model of building word forms and all possible combinations of affixes during formation of composite words.

***Keywords*:** Kazakh language, stemming, morphological analysis, tokenization, normalization, affixes, word form.

**Введение**

На данный момент нас окружают огромное количество информационных потоков. Развитие информационных технологий способствует извлечению полезных данных и получению новых интерпретаций. Обработка текста предусматривает собой вероятностное рассмотрение контекста с применением различных инструментов. Выделяются 7 основных этапов жизненного цикла обработки информации [3]:

1. Сбор, получение данных (извлечение исходных текстов из источников и их загрузка)
2. Разбор данных (очистка лишних слов[стоп-слова] и разбор извлеченных текстов [графематический, морфологический, синтаксический])
3. Фильтрация данных (отбор пригодных текстов для дальнейшего анализа)
4. Интеллектуальный анализ текста
5. Представление данных (выбор основной модели для визуализации информации)
6. Улучшение представления (улучшение представления информации на основе анализа ее свойств и особенностей)
7. Взаимодействие интерактивного управления представлениям информации

Казахский язык считается агглюнативным языком и входит в группу языков тюркского мира. В казахском языке форма слова образуется путем прибавления суффиксов и окончаний к основе слова. При присоединении определенных окончаний строго учитываются грамматические признаки и особенности морфологических форм и законы сингармонизма. Под агглютинативностью понимается образование в языке грамматических форм и производных слов путем присоединения к корню аффиксов в определенном порядке. В казахском языке словоформы образуются путем конкатенации корня и аффиксов (суффиксов и окончаний). При этом каждый аффикс связан с наборами морфологических признаков, и порядок добавления аффиксов строго определен. Например, для имен существительных к основе слова сначала добавляется суффикс и далее окончание множественного числа, затем притяжательное окончание, далее следует падежное окончание и только после него — личное окончание.

Новые словоформы образуются с учетом морфологических и семантических признаков начальных форм следующим образом: сначала к начальной форме слова добавляются суффиксы; затем, двигаясь слева направо, определяется категория (глухие, звонкие и т. п.) последней буквы (последнего звука) начальной формы слова для добавления того или иного окончания.

Принципиальным отличием морфологии казахского языка от морфологии, например, русского является наличие в казахском языке (как и в других тюркских языках) закона сингармонизма, в соответствии с которым аффиксы слова полностью определяются звуковым составом его основы.

Разбор словоформ происходит по следующей схеме:

1. Последовательно отсекаются возможные окончания длиной от 0 до 𝑛 букв, таким образом слово разбивается на основу и флексию;

2. Для полученного окончания находится его словоизменительный класс;

3. Проверяется наличие полученной основы в словаре основ и находится номер её словоизменительного класса;

4. В случае совпадения словоизменительных классов выбираются соответствующие морфологические характеристики и строится лемма, которые и являются результатом анализа.

При разборе словоформ необходимо учитывать морфологические, синтаксические и фонетические особенности языка. Выделяются морфологические признаки при агглюнитации языка. В центре разбора выходит корень слова в именительном падеже, который служит основой для дальнейшего изменения(склонения, спряжения). Согласно принципам образования морфологических единиц существует определенная последовательность по присоединению аффиксов (т.е. суффиксов и окончаний).

Стэмматизация это центрообразующая технология в автоматической обработке текста на естественном языке по нахождению основы слова. Согласно заранее подготовленными лингвистическими разметками корпуса текста выполняется поиск словоформ. Необходимо правильно выполнять морфологический анализ слова, определить образующие его компоненты. Под словоформой мы будем понимать группу (кортеж), состоящую из токена, связанной с ним начальной формы и множества грамматических параметров.

Правило присоединения окончаний в казахском языке осуществляется правилу (Рис.1)[6] :

**С = ОС + МО + ПО + ППО + СО**

где **С** – словоформа, **ОС**- основа слова, **МО**- окончание множественного числа; **ПО** – притяжательное окончание; **ППО** – падежное окончание; **СО** – окончание формы спряжения.

Общая морфологическая форма определения состава выглядит так:

түбір (корень) +жұрнақ (суффикс) + жалғау (окончание).

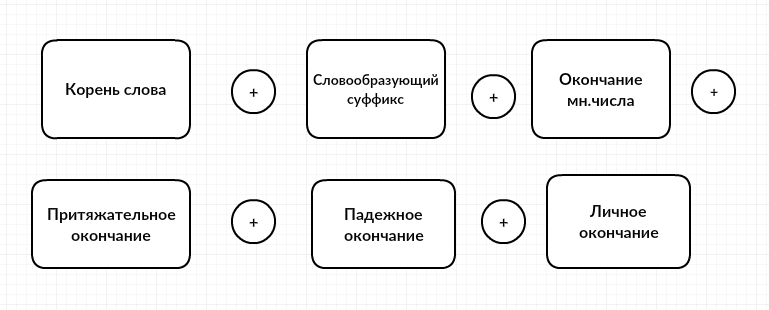


Рис 1. Правило присоединения окончаний

Согласно грамматическим правилам, для существительных строится следующая модель образования словоформ. Обозначим через *J* следующие виды окончаний (аффиксов)[7]:

1. *J*1 — окончание множественного числа.

2. *J*2 — притяжательное окончание.

3. *J*3 — падежное окончание.

4. *J*4 — личное окончание.

Возможны несколько комбинаций окончаний существительных:

а) окончание множественного числа + притяжательное окончание (*J1J2*);

б) окончание множественного числа + падежное окончание (*J1J3*);

в) окончание множественного числа + личное окончание (*J1J4*);

г) окончание множественного числа + притяжательное окончание + падежное

окончание (*J1J2J3*);

5) окончание множественного числа + притяжательное окончание + личное

окончание (*J1J2J4*);

6) притяжательное окончание + падежное окончание (*J2J3*);

7) притяжательное окончание + личное окончание (*J2J4*);

8) падежное окончание + личное окончание (*J3J4*).

Алгоритм разбора имен существительных выглядит следующим образом:

1. В качестве входных данных подается текстовый файл. Текст разделяется на предложения, а из предложений соответственно выделяются словоформы, которые будут входными данными для морфологического анализатора
2. Из цепочки слов определяется слово.
3. Слово проверяется в словаре основ слов.
4. Слово разбирается справа налево, начиная с последней буквы словоформ. Если аффикс найден в словаре окончаний, то он отсекается и данная проверка повторяется до тех пор, пока не будет найдена основа слова. При условии если оставшаяся часть слова больше не содержит окончаний, т.е после отсечения всех видов окончаний, оставшаяся часть считается корнем слова.

Проверка осуществляется по следующему принципу (Личное окончание->Падежное окончание-> Притяжательное окончание ->Окончание мн.числа)

Основная проблема описываемого алгоритма — наличие в казахском языке слов, в которых последние буквы основы совпадают с одним из аффиксов. В этом случае, алгоритм может отсечь больше, чем нужно. Единственный возможный механизм предотвращения таких ошибок — составление словаря основ, последние

буквы которых совпадают с одним из аффиксов.

Отсечение аффиксов и вывод основы слова осуществляется через метод

str.endswith(suffix[, start[, end]])

На начальном этапе проверялись имена существительные. Имена существительные имеют 4 вида окончаний[5]:

1. **Личные окончания**: 'мын', 'мін', 'бын', 'бін', 'пын', 'пін', 'мыз', 'міз', 'пыз', 'піз', 'быз', 'біз', 'сың', 'сің', 'сыңдар', 'сіңдер', 'сыз', 'сіз', 'сыздар','сіздер'.
2. **Падежные окончания**: 'ның', 'нің', 'тың', 'тің', 'дың', 'дің', 'ға', 'ге', 'қа', 'ке', 'на', 'не', 'а', 'е', 'ны', 'ні', 'ды', 'ді', 'ты', 'ті', 'н’, 'да', 'де', 'та', 'те', 'нда', 'нде', 'дан', 'ден', 'тан', 'тен', 'нан', 'нен', 'мен', 'бен', 'пен'.
3. **Притяжательные окончания**: 'м', 'ым', 'ім', 'ң', 'ың','ің', 'ңыз', 'ңіз', 'ыңыз', 'іңіз', 'сы', 'сі', 'ы', 'і', 'мыз', 'міз', 'ымыз', 'ларың', 'лерің', 'тарың', 'терің', 'дарың', 'дерің', 'ларыңыз', 'леріңіз', 'дарыңыз', 'деріңіз', 'тарыңыз', 'теріңіз', 'лары', 'лері', 'тары', 'тері', 'дары', 'дері'.
4. **Окончания множественного числа:** 'лер','лар','дар','дер','тар','тер'.

Как мы видим, на выходе при отсечении мы получаем смешанный набор отсеченной основы и нормальной формы слова. Но в некоторых случаях, все равно при отсечении окончаний остается нормальная форма слова. Эту проблему можно обойти посредством наличия словаря терминов.

**Заключение.** Алгоритм по нормализации имен существительных в казахском языке основан по принципу отсечения аффиксов и выявлению в результате основы слова. В ходе проведения экспериментов с текстами, выяснилось, что некоторые основы слов (лексемы) совпадают с аффиксами и вызывают конфликт при выявлении исходного корня. Рекомендуется составить словарь исключений и специальных слов для повышения эффективности обработки текста.

**© Ж.Ж. Жакыпова, 2018**

***Список литературы/References***

1. Stiven B., Ewan K., Edward L. Natural Language Processing with Python // O'Reilly, 2009.
2. P.Jackson., I.Moulinier., Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization //John Benjamins Publishing Company, 2002.
3. Кривальцевич Е.В., Обработка естественного языка (Natural Language Processing) при использовании технологии NLTK (natural language toolkit) на базе языка программирования Python // БСУ, 2015.
4. Jacob Perkins ., Text Processing with NLTK 3//Birmingham, Packt Publishing, 2014.
5. Рысбаева Г.К., Казахский язык: грамматический справочник //Сөздік-словарь, 2000.
6. Джумамуратов Р.А., Разработка средств создания морфологических словарей казахского языка на основе корпуса размеченных текстов // Новосибирск, 2013.
7. В. Б. Барахнин, А.М. Федотов, А.М. Бакиева, М. Н. Бакиев, С.Ж.Тажибаева, Т.В. Батура,О. Ю. Кожемякина, Д. А. Тусупов, М. А. Самбетбаева, Л.Х.Лукпанова, Алгоритмы генерации и стемматизации словоформ казахского языка // Cloud of Science, 2017.