SCIENCE AND INNOVATION
INTERNATIONAL SCIENTIFIC JOURNAL VOLUME 2 ISSUE 2 FEBRUARY 2023
UIF-2022: 8.2 | ISSN: 2181-3337 | SCIENTISTS.UZ

DISTRIBUTION OF WORD STEMS FOR OLD UZBEK WRITING

[1]Iskanderova S.N., [2]Begimkulova Parizod

[1]Associate Professor of Tashkent University of Information Technologies named after Muhammad al-Khorazmi, Ph.D.

[2]Graduate student of Samarkand branch of Tashkent University of Information Technologies named after Muhammad al-Khorazmi

https://doi.org/10.5281/zenodo.7622029

Abstract. *The rules for creating the tajweed of the old Uzbek language, the steps of the algorithm for creating the dictionary and creating the tajweed, and their descriptions are highlighted. The effectiveness of using Hamming distance and Levenshtein distance in word segmentation and word recognition is highlighted.*

Keywords: *old uzbek language, tajweed, electronic dictionary, electronic translation.*

I. INTRODUCTION

The direct replacement of letters is not enough for the old Uzbek language, one of the main reasons for this is the richness of our language. For this reason, we should separate the vocabulary and put the appropriate words in Persian and Arabic from our existing dictionaries.

One of the problems encountered during the preparation of an electronic dictionary is, of course, how to search for a word. A user may enter a word incorrectly while searching for a word, or may miss a few letters while typing. We will consider the following algorithms to avoid them [1-2].

II. MAIN PART

Hamming distance. The Hamming distance is measured by the number of different characters of two strings of equal length. In other words, it is the number of minimum character permutations of the second word to form one word.

For two binary strings a and b, the Hamming distance is equal to the number of ones of the value produced by the operation a {XOR b}. The Hamming distance generated from these binary strings is also known as the Hamming cube.

In the 1950s, after Richard Hamming invented the process of working with Hamming codes, this value was coined by his name, the Hamming distance.

Levenshtein distance. Levenshtein distance means the amount of mutual differences between two lines. In other words, it means creating the second word by minimally changing (adding, deleting, replacing) the first word. The Levenstein distance was calculated by Vladimir Levenstein in 1965 and is named after him. This method is also called "edit distance".

Mathematically, the Levenshtein distance is found as follows. To us given the strings a and b, $lev_{a,b}(|a|,|b|)$ is equal to:

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

If $a_i = b_j$ equality holds, $1_{(a_i \neq b_j)}$ is equal to 0 (zero), otherwise, is equal to 1 [3-4].

It should be kept in mind that in finding minima, the first element means to remove one character from the string a, the second element means to add one character, and the third one means to match or not.

The electronic translator consists of the following modules:

1. A module that takes into account prefixes and suffixes and the word stems in the old Uzbek script.
2. Search module for the words entered in the Arabic, Persian, Tajik language dictionary, taking into account the prefix and suffix and stem word.
3. There are modules that edit, delete and add new words in the dictionary.
4. A module that takes into account prefixes and suffixes and the root word.

In general, there are the following ways to refer to tables in the relational data base:

- • Sequential - in this method, all the given table are viewed in sequence and the desired one is searched;
- • True - in this method, the required data are selected from the table based on the key or index;
- • Indexed sequence - this method includes both of the above methods and is used to search a given group [5-7].

We use the first method in the search.

For example, we take the following word:

The word "school" is a 6-sound (phonemic) word, but in writing it is represented by 4 consonants in the form هکزت (м+к+т+б). The word "country" has 8 phonemes, and in writing it is written with 5 letters in the form هولکذ (м+м+л+к+т). These are Arabic words.

Turkish-Uzbek words such as "qoshiq", "qiziq", "chiziq", "bildi" have 5 phonemes, which are also expressed by 5 letters in writing: ٱبّٰی چیيٰی ثِللی

ىيٰى .

to find the Tajweed translation of this word, the program does the following:

• first of all, by reading from right to left, each word in the given sentence is separated;

• each word is checked for prefixes and suffixes and the root;

• extracted chunks are first checked for condition, then searched from given repository;

• the corresponding translation of the fragment in the base is obtained and replaced by the fragment;

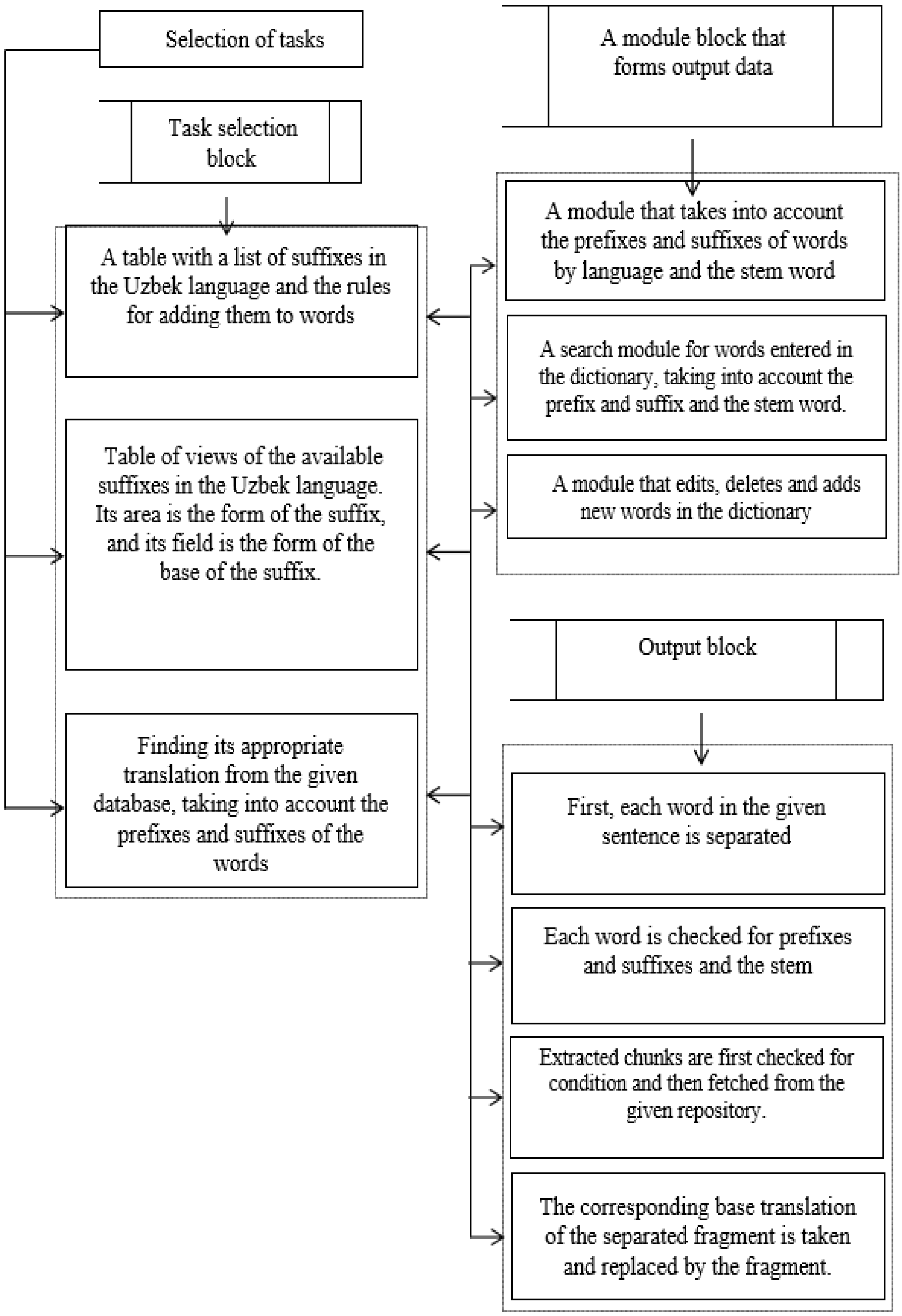
• direct letter replacement of words that are not in the dictionary;

• all pieces are combined and printed from left to right to get the result.

Picture 1.

Functional diagram of software operation

A graphical representation of the above process is depicted in Picture 1. The diagram of the electronic translator is shown in Picture 1.



In it, C is the text to be translated. We extract the given text words as elements of the Soz[i] array.

ثبی‌اهکّه ثباه آثبککّه کاهثبی

When parsing, we read the string from right to left and first check and pars each element of the resulting array for the prefix.

The text given above does not contain a prefix. If present, it is taken as an element of the array x[j]. Then the suffix is checked and extracted in the same way.

If the suffix is present, it is taken as an element of the array s[j]. In some cases, there can be more than one suffix in one word. The rest remains in place as an array element.

The extracted prefix, stem and suffix are checked against the given ones respectively. After verification, a new word is generated and printed. If a word is not found in the table, it will be printed as garbled text.

III. CONCLUSION

In this article, the formalization of programming processes and the model of the software complex and the scheme of the electronic translator have been developed based on the modular analysis of the electronic translator calculation algorithms.

REFERENCES

1. Назиров Ш.А., Кадиров У.Э. Давлат тилидаги элементар саволларнинг лексик-морфологик анализаторини яратиш.«Современные проблемы алгоритмизации и программирования». 5-7 сентября 2001 г. Ташкент. –С.413-414.
2. Назиров Ш.А., Урынбаев С.К. Электронный морфологический и двуязычный словари наиболее употребляемых слов узбекского языкапо экологии и кибернетике. // Уз. журн. «Проблемы информатики и энергетики». №2. – Ташкент, 2002. –С.62-66.
3. Назиров Ш.А., Хомидов Х.Х., Алниязов А.И. и др. Формализация конструкций предложений узбекского, турецкого и каракалпакского языков. Труды I Международной конференции «Компьютернаяобработка тюркских языков». – Астана: ЕНУ им. Л.Н. Гумилева, 2013.С. 33-47.
4. Назиров Ш.А., Рахманов К.С., Махмудов А.З. Классификация и построения базы данных словарей по тюркских языков //Международная конференция «Актуальные проблемы прикладной математики, информатики и механики». – Воронеж, 17-19 сентября 2012. С. 110-116.
5. Назиров Ш.А. Туркий тиллар учун электрон таржиманинг дастурий таъминоти яратиш муаммолари // ТАТУ хабарлари №4/2011 Тошкент.57-60 бет.
6. Назиров Ш.А., Рахманов К.С., Махмудов А.З. Туркий тиллар учун электрон таржимани яратишда алфавитларнинг турлича бўлиши масаласи // Фан, таълим ва ишлаб чиқариш интеграциясини ахборот коммуникация технологиялари асосида ривожлантириш муаммолари. Республика илмий – амалий анжуман материаллари тўплами. Қарши 2012й. 71-74б.
7. Назиров Ш.А., Рахманов Қ.С., Махмудов А.З. Икки тилли электрон таржима дастурининг берилганлар базаси структураси // Ахборот технологиялари ва телекоммуникация муаммолари. Республика илмий-техник конференцияси. Тошкент 14-15 март 2013 й. 131-133 б.