1.1.6.[01.01.07] МАТЕМАТИКАИ ҲИСОББАРОР
1.1.6.[01.01.07] ВЫЧИСЛИТЕЛЬНАЯ МАТЕМАТИКА
1.1.6.[01.01.07] COMPUTATIONAL MATHEMATICS

ТДУ 004.934.2

ОИД БА МАСЪАЛАИ ТАШКИЛИ КОРПУСИ НУТҚИ ТОҶИКӢ

Мақсудов Анвар Темурович, *н.и.физ-мат., дотсент, директори Маркази илмии Хуҷандии Академияи милли илмҳои Тоҷикистон;* ***Худойбердиев Хуршед Атохонович,*** *н.и.физ.- мат., дотсенти кафедраи барномарезӣ ва низомҳои иттилоотӣ, ДПДТТХ;* ***Ашурова Шабнам Нуруллаевна,*** *омӯзгори калони кафедраи барномарезӣ ва низомҳои иттилоотӣ, ДПДТТХ (Тоҷикистон, Хуҷанд)*

О ПРОБЛЕМАХ ФОРМИРОВАНИЯ ТАДЖИКСКОГО РЕЧЕВОГО КОРПУСА

Максудов Анвар Темурович, *к.физ.-мат.наук, доцент, директор Худжандского научного центра Национальной академии наук Таджикистана;* ***Худойбердиев Хуршед Атохонович,*** *к.физ.-мат.наук, доцент кафедры программиро-вания и информационных систем, ХПИТТУ;* ***Ашурова Шабнам Нуруллаевна,*** *старший преподаватель кафедры программирования и информационных систем, ХПИТТУ, (Таджикистан, Худжанд)*

ABOUT THE PROBLEMS OF CREATING THE TAJIK SPEECH CORPUS

Maqsudov Anvar Temurovich, *Candidate Of Physical and Mathematical Sciences, Director of Khudjand Scientific Center of the National Academy of Sciences of Tajikistan,* ***E-mail:*** *atmaxudov@mail.ru;* ***Khudoyberdiev Khurshed Atokhonovich,*** *candidate Of Physical and Mathematical Sciences, Associate Professor of the Department of Programming and Information Systems, KPITTU;* ***E-mail:*** *tajlingvo@gmail.com;* ***Ashurova Shabnam Nurullaevna,*** *Senior Lecturer of the Department of Programming and Information Systems, KPITTU (Tajikistan, Khujand),* ***E-mail:*** *sh.nurulloevna@gmail.com*

Калидвожаҳо: *лингвистикаи компютерии забони тоҷикӣ, амсилаи математикӣ, тарҳрезии компютерӣ, балоиҳагирии сохтори мантиқӣ, корпуси нутқ, технологияи коркарди нутқ.*

Дар мақола ҳалли масъалаи ташкили корпуси нутқ ва аҳамияти он дар раванди коркарди низомҳои шинохт ва синтези нутқ бо забони тоҷикӣ мавриди омӯзиш ва таҳлил қарор дода шудааст. Инчунин, усулҳо ва ташкили корпуси нутқ барои забони тоҷикӣ шарҳ дода шудааст. Дар қисми аввали мақола оид ба корпусҳои нутқ дар мисоли забонҳои англисӣ, русӣ ва форсӣ сухан меравад, дар қисми дуюм бошад сохтори мантиқии корпуси нутқ барои забони тоҷикӣ пешниҳод карда шудааст, қисми сеюм оид ба амсилаи математикии корпуси нутқ бахшида шудааст ва дар қисми ҷамъбастӣ воситаҳои компютерӣ барои коркарди корпуси нутқ пешниҳод карда шудаанд. Натиҷаҳои баррасишуда дар доираи иҷрои лоиҳаи буҷавии «Коркарди корпуси овози забони тоҷикӣ барои ҳалли масъалаҳои лингвистикаи компютерӣ» таҳти рақами № 0123TJ1547 тасдиқ шуда, ҳамчун як қисми тадқиқоти давомдор ба даст оварда шудаанд.

Ключевые слова*. Компьютерная лингвистика таджикского языка, математическая модель, компьютерное моделирование, проектирование логической структуры, корпус речи, технология обработки речи.*

В данной статье приведены результаты изучения и анализа по решению проблем организации речевого корпуса и его значения в процессе разработки систем распознавания и синтеза речи на таджикском языке. Также описаны методы и формирование речевого корпуса таджикского языка. В первой части статьи рассмотрены речевые корпусы на примере английского, русского и персидского языков, во второй части представлена логическая структура речевого корпуса таджикского языка, в третьей части рассмотрена математическая модель речевого корпуса, а в заключительной части представлены компьютерные средства разработки речевого корпуса. Обсуждаемые результаты получены в рамках продолжающегося исследования и выполнения

бюджетного проекта «Обработка корпуса таджикского языка для решения задач компьютерной лингвистики» утвержденной по номеру № 0123TJ1547.

Keywords: *Computer linguistics of Tajik language, mathematical model, computer modelling, design of logical structure, speech corpus, speech processing technology.*

The given article studies and analyses the solution to the problem of speech corpus organization and its importance in the process of developing speech recognition and synthesis systems in the Tajik language. The problems and processing of the speech corpus of the Tajik language are also described. In the first part of the article speech corpus on the example of English, Russian and Persian languages are considered, in the second part the logical structure of the speech corpus of the Tajik language is presented, in the third part the mathematical model of the speech corpus is considered, and in the final part the computer means of speech corpus development are presented. The discussed results were obtained within the framework of the ongoing research and implementation of the budget project «Processing the Tajik language corpus for solving computational linguistics problems» approved under number 0123TJ1547.

Муқаддима. Дар шароити имрӯзаи рушди технологияи иттилоотӣ масъалаҳои шинохт ва синтези худкори нутқ аз матни додашуда яке аз масъалаҳои маъмултарини низомҳои идоракунӣ ва технологияҳои овозӣ ба ҳисоб мераванд. Дар амал, низомҳои шинохт ва синтези нутқ бо худ масъалаи босамар ва зуд идора кардани равандҳои корӣ бо имконияти мусоҳибаи овозӣ байни корбар ва таҷҳизотро фароҳам меоранд. Барои коркарди низомҳои овозӣ, дар мадди аввал, масъалаи коркард ва ба истифода омода кардани корпуси нутқ дар забони табиӣ ҷой дорад. Аз ин сабаб дар мақола масъалаи ташкили корпуси нутқ бо забони тоҷикӣ, тарҳрезии амсилаи математикӣ ва сохтори мантиқии он мавриди муҳокима қарор дода шудааст.

Корпусҳои нутқ аз манбаи додаҳои мултимедиа, аудио ва воҳидҳои матнии ба онҳо мувофиқ иборат мебошанд. Додаҳои аудиовӣ бо ишораҳои элементҳои нутқ, яъне садоҳо, ҳиҷоҳо, калимаҳо, ибораҳо нигоҳ дошта мешаванд ва воҳидҳои матнӣ бошанд, транскрипсияҳои мувофиқро дар бар мегиранд. Барои коркард ва ташкили корпуси нутқ воситаҳои ёрирасон иборат аз барномаҳои компютерӣ, ки имконияти кор бо додаҳои аудиовӣ ва транскрипсияи худкори онҳо истифода мешаванд. Инчунин, дар манбаи додаҳо барои таъмини иттилоотӣ ҷамъорваии маълумот оид ба хусусиятҳои аудио муҳим аст.

Корпуси нутқ аз нигоҳи технологияи компютерӣ аз маҷмуи порчаҳои нутқ иборат мебошад, ки ба талаботи сохтори махсуси додаҳои компютерӣ ҷавобгӯ мебошанд. Сохтори додаҳо вобаста аз қоидаҳои рақамикунонии нутқ дар хотираи компютер бо барномаҳои махсус муайян карда мешавад. Порчаи нутқ ҳамчун воҳиди асосии корпус рақамӣ карда шуда, ба воҳиди матнии ба он алоқаманд рост меояд [1].

1. Корпуси нутқ барои забонҳои хориҷӣ. Дар ин қисми мақола оид ба таҷрибаи таҳия ва истифодаи амсилаҳо, усулҳо ва талабот оид ба ташкили корпусҳои нутқ дар мисоли забонҳои англисӣ, русӣ ва форсӣ сухан меравад.

Корпуси нутқ барои забони англисӣ. Дар тули зиёда аз 20 соли охир, низоми FESTIVAL барои шинохт ва синтези нутқи забони англисӣ ва испанӣ, ки аз тарафи олимони Донишгоҳи Эдинбург таҳия шудааст, ба стандарти воқеӣ табдил ёфтааст. Кормандони Донишгоҳи амрикоии Карнеги Меллон дар лоиҳаи FESTIVAL ширкат варзида, низоми нави FestVox-ро барои тарҳрезӣ ва сохтани пойгоҳи додаҳои дифонӣ пешниҳод намуданд. Вақти ташкили пойгоҳи додаҳои нутқ дар асоси корпуси нутқ аз чанд моҳ то чанд ҳафта расонида шудааст. Ширкатҳои ҷаҳонӣ, ки дар самти технологияи овозӣ фаъолият мебаранд, аз ин низоми FESTIVAL васеъ истифода мебаранд [2].

Дар кишварҳои Аврупо ва Амрико яке аз маъмултарин корпуси нутқ TIMIT мебошад, ки ҳам барои таҳқиқоти фонетикӣ ва ҳам барои таҳияву санҷиши низомҳои шинохти пайвастаи нутқ бо забони англисӣ пешбинӣ карда шудааст. Дар таҳияи корпуси TIMIT (Texas Instruments/Massachusetts Institute of Technology) якчанд ташкилот ва марказҳои илмӣ-тадқиқотии маъруф ширкат варзиданд. Манбаи матнии TIMIT дорои 6300 ҷумла, 10 ҷумла барои ҳар як 630 сухангӯ сабтшуда, 70% мардон ва 30% занҳо рост меояд. Ҳангоми интихоб ва сабти сухангӯён синну сол, ҷинс, қад, нажод, сатҳи маълумот ва вақти сабти нутқ ба назар гирифта шудааст. Дар корпуси TIMIT, файлҳои овозии аз баландгӯякҳои гуногун сабтшуда ба қисмҳои омӯзишӣ ва санҷишӣ тақсим карда мешаванд. Бояд қайд кард, ки то имрӯз корпуси нутқи TIMIT бо илова кардани манбаи додаҳои забонҳои нав васеъ шуда, аз тарафи ширкатҳои барномасозии ҷаҳон истифода карда мешаванд [3].

Корпуси нутқ барои забони русӣ. Дар Федератсия Русия дар масъалаҳои коркарди низомҳои синтез ва шинохти нутқ бо забони русӣ якчанд корпусҳои миллии русӣ таҳия карда шудаанд. Шумораи зиёди олимон дар ин самт ба дастовардҳои бузург ноил шудаанд. Яке аз онҳо «Маркази технологияҳои нутқ» дар Санкт-Петербург ба ҳисоб меравад. Аз тарафи мутахассисони ширкат низоми VitalVoice (овози зинда) - амалӣ карда мешавад, ки дар асоси технологияи интихоби воҳидҳо, якҷоя кардани он бо синтези аллофонӣ амал мекунад. Дар натиҷа, овози 10 шахс, 4 мард ва 6 зан, дар корпусҳои нутқи дарозиаш гуногун, аз 1,5 то 8 соат истифода карда шуданд [4].

Дар Институти таҳлили системавии Академияи илмҳои Русия барои забони русӣ дар асоси корпуси TIMIT аввалин корпуси нутқи ҳамгирои ISABASE таҳия карда шудааст. Корпуси ISABASE дорои 4653 порчаҳои нутқ аст, яъне дар он ҷумлаҳои аз ҷиҳати фонетикӣ мутавозин, ки аз ҷониби 20 мард ва 18 зан хонда шудааст, сабт шудааст. Дар мавриди транскрипсия 110 монофон истифода шудааст. Шумораи зиёди ташкилоти барномасозӣ ва технологияи коркарди овоз дар Русия аз корпуси ISABASE васеъ истифода мебаранд ва он бо технологии нави шабакаи нейронӣ рушд карда истодааст [5].

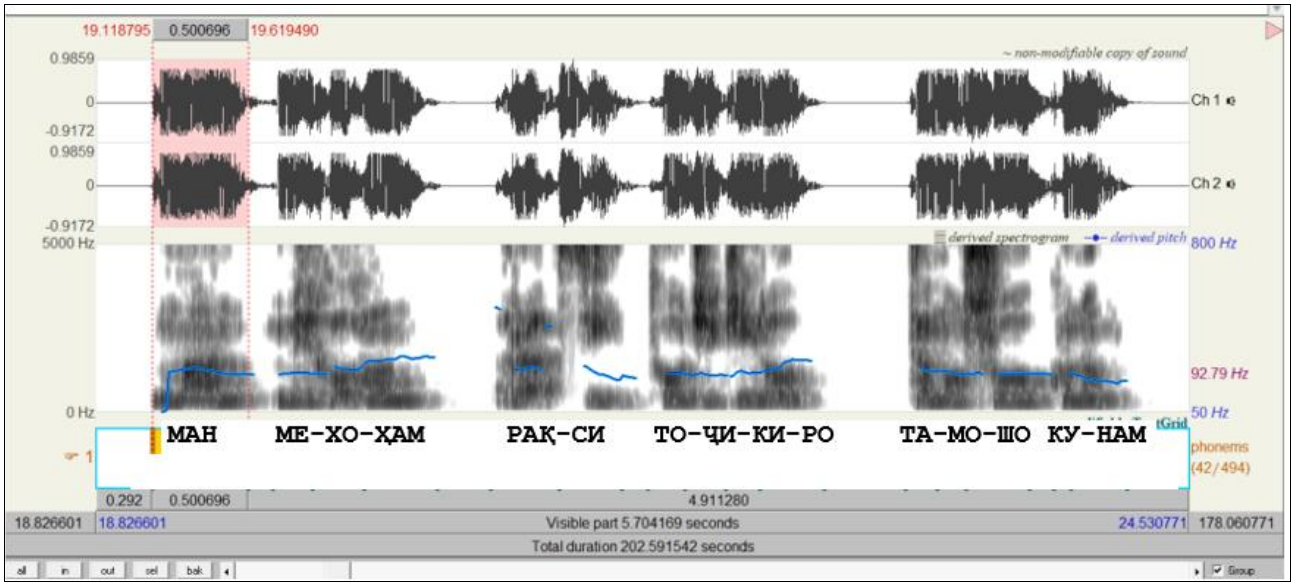
Ҷадвали 1. Таркиби корпусҳои нутқ ва самти истифодаи онҳо

№	Номи корпус	Забон	Миқдори элементи овозӣ	Самти истифодабарӣ
1	FESTIVAL	англисӣ, испанӣ		Синтез
2	TIMIT	англисӣ	6300 ҷумла	Синтез ва шинохт
3	VITALVOICE	русӣ		Синтез
4	ISABASE	русӣ	110 монофон	Синтез ва шинохт
5	FARSDAT	форсӣ	1000 ҷумлаҳо	Синтез ва шинохт

Корпуси нутқ барои забони форсӣ. Бо дарназардошти он, ки забони тоҷикӣ бо забони форсӣ қоидаҳои фонетикӣ ва хусусиятҳои табии хос дорад, дар марҳилаи тадқиқот якчанд корпуси нутқ бо забони форсӣ мавриди таҳлил қарор дода шудаанд. Корпуси нутқи FARSDAT аз 6000 воҳиди матнӣ бо забони форсӣ иборат аст, ки аз ҷониби 20 шахси гуногун 300 воҳиди матнӣ ташкил карда шудааст. Дар мадди аввал, маҷмуи ҷумлаҳо иборат беш аз 1000 калимаи форсӣ дар рӯзномаҳои ҳаррӯзаи форсӣ нашр шуда ҷамъоварӣ карда шуданд. Дар асоси тасдиқи муаллифон корпуси нутқи FARSDAT тамоми аллофонҳои забони форсиро дар бар мегирад ва дар раванди коркарди низомҳои коркарди овоз босамар истифода карда мешавад [6].

Аз маълумоти дар ҷадвали 1 оварда шуда чунин бармеояд, ки барои ташкили корпуси нутқ масъалаи якум, ҷамъоварии воҳидҳои матнӣ ва сипас сабти талаффузи онҳоро дар бар мегирад [7]. Ба сифати воҳидҳои матнӣ ҷумлаҳо, ибораҳо ва калимаҳои калидии забон истифода карда шудаанд. Аз ин рӯ, барои ташкили корпуси нутқи забони тоҷикӣ нақшаи корӣ бо таъмини иттилоотии он, тадқиқоти амсилаи математикӣ ва сохтори мантиқии компютерии он ба роҳ мондан лозим аст, ки дар қисмҳои поён баррасӣ карда шудаанд [8].

2. Амсилаи математикии корпуси нутқ. Барои татбиқи низомҳои дар қисмҳои аввал тавсияшуда, бояд хусусиятҳои амсилаи акустикӣ ва фонетикии онҳо баҳо дода шаванд. Азбаски амсилаҳо барои ифода кардани нутқи равон барои ҳар як шахс пешбинӣ шудаанд, омӯхтани онҳо дар алоҳидагӣ зарурият надорад. Инчунин, аз сабаби андозаи бузург доштани амсилаи забон, шумораи хусусиятҳо ва нишонаҳои нутқ аз ҷониби коршинос ва ё лингвист ба қисмҳо ҷудо кардан ғайриимкон мебошад. Бо ин мақсад амалҳои қисман худкор вуҷуд доранд, ки ҳадди ақал дахолати коршиносро талаб мекунанд. Яъне, танҳо дар марҳилаи муайян кардани транскрипсия ва баҳогузории дурустии он коршинос бо муайян кардани хусусиятҳои фонетикии корпуси нутқ ҳиссаи арзандаи худро мегузорад (расми 1).



Расми №1. Раванди баҳодиҳӣ ва транскрипсияи воҳидҳои корпус

Истифодаи алгоритми Баум. Барои намуди амсилаҳое, ки дар ифодаҳои интегралӣ истифода мешаванд, алгоритми Баум барои баҳодиҳии хусусиятҳои пайдарпайии воҳидҳои овозӣ истифода мешавад. Маълумоти омӯзишӣ ҳамчун маҷмуи зиёди воҳидҳои овозӣ, яъне сигнал тахмин карда мешавад, {xi(t)}N i=1. Ҳар як сигнал, xi(t) ба як ҷумлаи Wi рамзгузоришуда рост меояд ва аз калимаҳои w(i,1), w(i,2), ..., w(i,n) иборат аст. Ҳар як w(i,j) дар махзани W ҷойгир аст, ки аз он талаффузи ҳама гуна воҳидҳои матниро аз рӯи рӯйхати пешакӣ муайяншудаи воҳидҳои овозӣ ба даст орем, оварда мешавад.

$$w(i,j) = \varphi k(1), \varphi k(2) \ldots \varphi k(m)$$

Бо ҳар як воҳиди хурди овоз ба сифати фонемаҳо, φk(j-1), φkj, φk(j+1) дар тарафи рост дар асоси модели махфии Марков вектори хусусиятҳои λ(k,j) тавсиф мешавад. Вобаста аз ин, барои ҳар як ҷумлаи W(i) тавассути пайваст кардани ҳамаи воҳидҳо, дар пайдарпаӣ дуруст ҷойгир шудани тартиби калимаҳо ва талаффузи ҳар яки онҳо баҳогузорӣ карда мешавад. Бо дарназардошти мундариҷаи фонемаҳо барои ҳар як воҳиди овозӣ, аз ҷумла фонемаҳои дар ҳудуди калимаҳо ҷойгир шуда, транскрипсия муайян карда мешавад. Пас алгоритми Баум барои баҳодиҳии ҳама λ(k,j) якҷоя аз пайдарпаии мушоҳидаи коршинос гирифташуда, самаранок истифода мешавад [9,с.10].

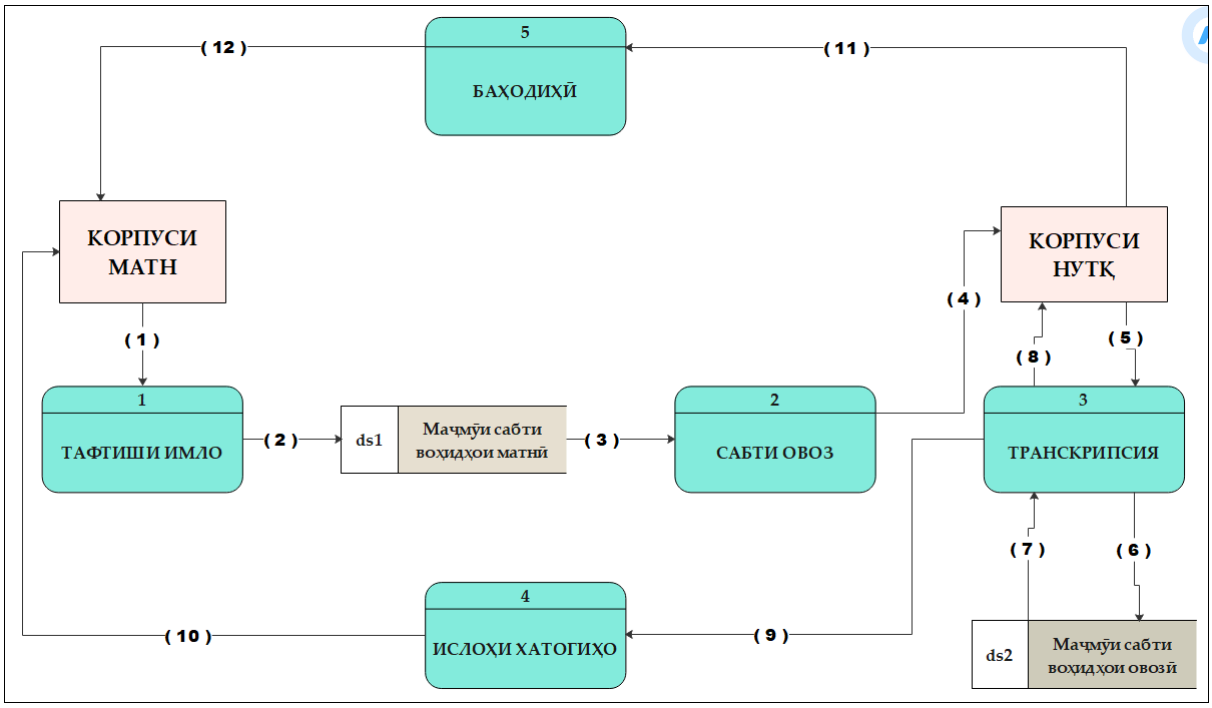
Ин раванд барои ҳар як W(i), i = 1, 2, ... , N амалӣ карда мешавад. Ҳар гуна қиматҳои барои λ(k,j) аз аъзои W(i) ба даст оварда, аз нав барои аъзои W(i+1) дар асоси алгоритми Баум муайян карда мешаванд. Аз ин рӯ, зермаҷмуаҳои гуногуни хусусиятҳо барои ҳар як аъзои W(i) дубора аз нав ҳисоб карда мешаванд. Ин ҷараён барои ба даст овардани қиматҳои нави воҳидҳои овозӣ ва ба қадри кофӣ таъмин кардани саҳеҳии амсилаи забон мусоидат менамояд. Барои ба даст овардани саҳеҳии ҳадди ақал аз 75% зиёд садҳо соат маълумоти овозӣ лозим аст, аммо раванди баҳодиҳӣ худкор аст ва танҳо вақти компютерро талаб мекунад.

3. Сохтори мантиқии корпуси нутқ. Дар асоси амсилаи математикӣ ва тарҳрезии низоми иттилоотӣ оид ба ташкили корпуси нутқ марҳилаҳои ғоявӣ ва сохтори мантиқии таъминоти барномавӣ муайян карда шуданд. Дар мадди аввал таҳқиқот панҷ марҳилаи асосиро муайян сохт, ки дар расми 2 тасвир карда шудаанд.

Ҷараёни асосии аз корпуси матн ба даст овардани корпуси нутқ аз 5 марҳила, 2 манбаи додаҳо ва 12 амалиёти коршиносон бо низоми иттилоотӣ иборат мебошад.

1. ТАФТИШИ ИМЛО – дар ин марҳила аз тарафи коршиносон ислоҳи имлои калимаҳо ва таҳлили синтаксисии ибора ва ҷумлаҳо ба роҳ монда мешавад. Коршиносон барои ислоҳи хатоҳои имлоӣ ва синтаксисӣ амалиёти (1), (2) ва (3)-ро иҷро мекунанд, ки дар натиҷа байни онҳо манбаи додаҳои ds1- «Маҷмуи сабти воҳидҳои матнӣ» ба даст оварда мешаванд [11].
2. САБТИ ОВОЗ – таҳти назорати коршиносон ва як муҳандиси садо дар студияи сабт раванди сабт, коркард ва нигоҳдории воҳидҳои талаффузшуда ба КОРПУСИ НУТҚ бо амалиёти (4) равона карда мешаванд.

3. ТРАНСКРИПСИЯ - генератори фонетикии транскрипсия ва ё бо дигар мазмун «фонетизатор» ба таври худкор ва назорати коршиносон пайдарпайии фонемаҳоро барои ҳар як воҳиди овозӣ омода мекунад. Дар ин маврид, якчанд талаффузҳои имконпазир ва ба воҳиди матнӣ хос ба КОРПУСИ НУТҚ равона карда мешавад. Сабтҳо ва транскрипсия барои ҳар як воҳиди фонетикӣ бо истифода аз модели махфии Марков мувофиқ карда мешаванд ва дар манбаи додаҳои ds2- «Маҷмуи сабти воҳидҳои овозӣ» нигоҳ дошта мешаванд. Дар ин марҳила, бо дарназардошти пайдо нашудани ягон хатогиҳои фонетикӣ, имлоӣ ва синтаксисӣ пайдарпаии амалиёти (5), (6), (7) ва (8) иҷро карда мешаванд.
4. ИСЛОҲИ ХАТОИҲО - дар мавриди қарор доштани хатоӣ аз тарафи коршиносон ва муҳандисон бо мақсади ҳамвор кардани воҳидҳои матнӣ ва овозӣ амалиёти (9) ва (10) иҷро карда мешаванд. Бояд қайд кард, ки дар ин марҳила танҳо ба намуди хатоиҳои фонетикӣ, имлоӣ ва синтаксисӣ аҳамият дода мешаванд. Илова бар ин, дар ҳолати ба чашм расидани хатоиҳое, ки дар мавриди сабти овоз иҷро мешаванд, дар марҳилаи баҳодиҳӣ ислоҳ карда мешаванд.
5. БАҲОДИҲӢ - бо мақсади муайян кардани мувофиқатии воҳидҳои овозӣ бо воҳидҳои матнӣ коршиносон баҳогузорӣ мекунанд. Дар ин марҳила алгоритми Бауми дар қисми дуюми мақола баррасӣ шуда, истифода карда мешавад. Дар натиҷа, бо иҷрои амалиёти (11) ва (12) корпуси мутавозин байни матн ва овоз ба даст оварда мешавад.



Расми №2. Сохтори мантиқии низоми ташкили корпуси нутқ

Дар умум, коркарди корпуси нутқ як раванди мураккаби технологӣ мебошад ва аз амалиёти зерин иборат аст:

1. Ҷамъоварии воҳидҳои матн (калимаҳо, ибораҳо ва ҷумлаҳои сода).
2. Омода намудани воҳидҳои матнӣ бо ислоҳ кардани хатоиҳои имлоӣ ва синтаксисӣ. Омода намудани қоидаҳои фонетикии забони тоҷикӣ.
3. Интихоби дикторҳо, талаффузи воҳидҳои матн ва сабти порчаҳои нутқ.
4. Санҷиши фонетикии порчаҳои нутқ ва аломатгузории онҳо.
5. Таҳияи стандартҳои транскрипсияи сигналҳои нутқи сатҳҳои гуногун.
6. Санҷиши фонетикии порчаҳои нутқ ва аломатгузории онҳо.
7. Коркарди модулҳои барномавии компютерӣ барои ташхиси худкори натиҷаи транскрипсия.
8. Ташхиси сифати сабти порчаҳои нутқ.
9. Коркарди қоидаҳои махсус ба аломатгузорӣ ва тафсири фонетикии порчаҳои нутқ.

10.Ҳамвор кардани воҳидҳои матнӣ ва овозӣ.

11.Баҳодиҳӣ, муайян кардани мувофиқатии воҳидҳои овозӣ ва матнӣ.

12.Ворид намудани маълумоти иловагӣ оид ба порчаи нутқ дар корпус.

Дар асоси таҳқиқот, марҳилаҳо ва амалиёти номбуда нақшаи коркарди корпуси нутқ барои забони тоҷикӣ таҳти Tajik-Speech-Corpus муайян карда шуда, дар вақти навиштани ин мақола дар марҳилаи сеюм, яъне «Транскрипсия» қарор дорад. Натиҷаи корҳои тадқиқотӣ оид ба марҳилаҳои «Ислоҳи хатоиҳо» ва «Баҳодиҳӣ» дар корҳои илмии оянда баррасӣ карда хоҳанд шуд.

Хулоса. Дар ин мақола усулҳое, ки барои беҳтаргардонии раванди ташкили корпуси нутқ барои забони тоҷикӣ истифода бурда мешаванд, муҳокима карда шуданд. Таҷрибаи таҳия ва истифодаи амсилаҳо, усулҳо ва талабот оид ба ташкили корпусҳои нутқ дар забонҳои англисӣ, русӣ ва форсӣ таҳқиқот карда шуданд. Бо мақсади содагардонии технологияи ташкили корпуси нутқ амсилаи математикӣ, як қатор марҳилаҳо ва амалиёти махсуси появӣ муайян ва асоснок карда шуданд. Барои сабти воҳидҳои нутқ барномаи компютерии Adobe Audition ва барои таъмини раванди транскрипсия маҷмуи барномаи Praat (Phonetic Analyzer) мавриди истифода қарор дода мешаванд. Дар умум, аз тарафи гурӯҳи корӣ веб-замимаи Taj-Speech-Corpus бо мақсади ташкили махзани додаҳои нутқ барои корпуси забони тоҷикӣ, аз ҷумла ҷамъоварӣ, нигоҳдорӣ ва ифодаи матнии воҳидҳои нутқ коркард карда шудааст. Веб-замима дар шабакаи интернет дар суроғаи https://tajlingvo.tj ҷойгир карда мешавад [12].

АДАБИЁТ:

1. Худойбердиев, Х. А. Разработка таджикского звукового корпуса для решения некоторых задач компьютерной лингвистики / Х. А. Худойбердиев, Д. З. Музафаров, Ш. Н. Ашурова // Вестник ПИТТУ имени академика М.С. Осими. – 2023. – № 2(27). – С. 7-14. – EDN FMCKBZ.
2. Продеус, А. Н. Речевые корпуса: создание и проблемы / А. Н. Продеус // Электротехнические и компьютерные системы. – 2013. – № 9(85). – С. 118-126. – EDN QAGMTL.
3. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, The Trustees of the University of Pennsylvania. [Маҳзани электронӣ]. Суроғаи дастрасӣ: https://catalog.ldc.upenn.edu/LDC93S1 [Санаи муроҷиат - 08.08.2024]
4. Соломенник, А. И. Автоматизация процедуры подготовки нового голоса для системы синтеза русской речи/А.И.Соломенник,П.Г Чистиков, С В Рыбин [и др.] // Известия высших учебных заведений. Приборостроение. – 2013. – Т. 56, № 2. – С. 29-32. – EDN PYBEDH.
5. Кривнова, О. Ф. Речевые корпуса на новом технологическом витке / О. Ф. Кривнова // Речевые технологии. – 2008. – № 2.– С. 13 – 23.
6. Kermanshahi M. A., Akbari A., Nasersharif B. Transfer learning for end-to-end ASR to deal with low-resource problem in Persian language //2021 26th International Computer Conference, Computer Society of Iran (CSICC). – IEEE, 2021. – С. 1-5.
7. Усманов, З. Д. Алгоритм представления таджикских словосочетательных словоформ фрагментами предложений / З. Д. Усманов, М. Довудов // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2013. – № 4(153). – С. 69-76. – EDN SADBON.
8. Худойбердиев, Х. А. Моделирование системы автоматической обработки текста на таджикском языке/Х.А.Худойбердиев// International Journal of Open Information Technologies. – 2023. – Т. 11, № 3. – С. 27-33. – EDN KRBOBH.
9. Савин, А.Н. Разработка системы распознавания речи на основе скрытых марковских моделей отдельных слов/А.Н.Савин,Н.Е.Тимофеева,А.С.Гераськин,Ю.А.Мавлютова // Известия Саратовского университета. Новая серия. Серия: Математика. Механика. Информатика. – 2017. – Т. 17, № 4. – С. 452-464. – DOI 10.18500/1816-9791-2017-17-4-452-464. – EDN ZXJPON.
10. Худойбердиев, Х. А. Амсиласозии раванди шинохти нутқ дар заминаи нутқи забони тоҷикӣ / Х. А. Худойбердиев, Б. Х. Ашурзода // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2022. – No. 2(58). – P. 39-42. – EDN VNMJGH.
11. Солиев,О.М.Система автоматической проверки орфографии таджикского языка - TajSpell / О. М. Солиев,Х.А.Худойбердиев, Г. М. Довудов // Вестник Технологического университета Таджикистана. – 2021. – № 3(46). – С. 188-194. – EDN WZYMGP.

12. Худойбердиев, Х.А. Web-приложение «Автоматические системы обработки информации на таджикском языке» www.tajlingvo.tj. – Свидетельство о государственной регистрации информационного ресурса, Республика Таджикистан. №4202200496 от 28/04/2022.

REFERENCES:

1. Khudoyberdiev, Kh.A. Development of the Tajik speech corpora for solving some problems of computer linguistics / Kh.A. Khudoyberdiev, D.Z. Muzafarov, Sh.N.Ashurova // Bulletin of PITTU named after academician M.S. Oshimi. – 2023. – № 2(27). – C. 7-14. – EDN FMCKBZ.
2. Prodeus, A.N. Speech cases: creation and problems / A.N. Prodeus // Electrical and computer systems. – 2013. – № 9(85). – C. 118-126. – EDN QAGMTL.
3. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, The Trustees of the University of Pennsylvania. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1 [Date of application - 08.08.2024]
4. Solomennik, A.I. Automation of new voice creation procedure for a Russian tts system / A.I. Solomennik, P.G. Chistikov, S.V. Rybin, A.O. Talanov, N.A. Tomashenko] // News of higher educational institutions. Instrumentation. News of higher educational institutions – 2013. – T. 56, № 2. – C. 29-32. – EDN PYBEDH.
5. Khudoyberdiev, Kh.A. Modeling a system for automatic processing of text in the Tajik language / Kh.A. Khudoyberdiev // International Journal of Open Information Technologies. – 2023. – T. 11, № 3. – C. 27-33. – EDN KRBOBH.
6. Kermanshahi M. A., Akbari A., Nasersharif B. Transfer learning for end-to-end ASR to deal with the low-resource problem in Persian language //2021 26th International Computer Conference, Computer Society of Iran (CSICC). – IEEE, 2021. – C. 1-5.
7. Khudoyberdiev, Kh.A. Modeling the process of speech recognition in the context of Tajik language speech / Kh.A. Khudoyberdiev, B.Kh. Ashurzoda // Polytechnic Bulletin. Series: Intelligence. Innovations. Investments. – 2022. – No. 2(58). – P. 39-42. – EDN VNMJGH.
8. Kryvnova, O. F. Speech Corpora on New Technologic Level / O. F. Kryvnova // Speech Technology. – 2008. – № 2. –P. 13 – 23.
9. Savin, A.N. Development of speech recognition systems based on hidden Markov models of individual words / A.N. Savin, N.E. Timofeeva, A.S. Geraskin, Yu.A. Mavlutova // News of Saratov University. New series. Series: Mathematics. Mechanics. Computer science. – 2017. – T. 17, № 4. – C. 452-464. – DOI 10.18500/1816-9791-2017-17-4-452-464. – EDN ZXJPON.
10. Usmanov, Z.D. Algorithm for representing a specific Tajik word form as a phrase piece / Z.D.Usmanov, G.M.Dovudov // News of the Academy of Sciences of the Republic of Tajikistan. Department of Physical, Mathematical, Chemical, Geological and Technical Sciences. – 2013. – № 4(153). – C. 69-76. – EDN SADBON.
11. Soliev, O.M. Tajik language automatic spell-checking system - TajSpell / O.M. Soliev, Kh.A/ Khudoyberdiev, G.M. Dovudov // Bulletin of the Technological University of Tajikistan. – 2021. – № 3(46). – C. 188-194. – EDN WZYMGP.
12. Khudoyberdiev KH.A. Web-prilozhenie «Avtomaticheskie sistemy obrabotki informatsii na tadzhikskom yazyke» www.tajlingvo.tj. – Svidetel'stvo o gosudarstvennoi registratsii informatsionnogo resursa, Respublika Tadzhikistan. №4202200496, 28/04/2022.