ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ ҲУЗУРИДАГИ ИЛМИЙ ДАРАЖАЛАР БЕРУВЧИ DSc.13/30.12.2019.T.07.01 РАҚАМЛИ ИЛМИЙ КЕНГАШ

ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ

БАКАЕВ ИЛҲОМ ИЗАТОВИЧ

ЎЗБЕК ТИЛИ СЎЗ ШАКЛЛАРИНИ МОРФОЛОГИК ТАҲЛИЛ ҚИЛИШ МОДЕЛЛАРИ ВА АЛГОРИТМЛАРИ

05.01.04 – Ҳисоблаш машиналари, мажмуалари ва компьютер тармоқларининг математик ва дастурий таъминоти

ТЕХНИКА ФАНЛАР БЎЙИЧА ФАЛСАФА ДОКТОРИ (PhD) ДИССЕРТАЦИЯСИ АВТОРЕФРАТИ

Тошкент – 2021

УДК 004.9+81'322.2:811.512.133

**Техника фанлари бўйича фалсафа доктори (PhD)
диссертацияси автореферати мундарижаси**

**Оглавление автореферата диссертации доктора философии (PhD)
по техническим наукам**

**Contents of dissertation abstract of doctor of philosophy (PhD)
on technical sciences**

ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ ҲУЗУРИДАГИ ИЛМИЙ ДАРАЖАЛАР БЕРУВЧИ DSc.13/30.12.2019.T.07.01 РАҚАМЛИ ИЛМИЙ КЕНГАШ

ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ

БАКАЕВ ИЛҲОМ ИЗАТОВИЧ

ЎЗБЕК ТИЛИ СЎЗ ШАКЛЛАРИНИ МОРФОЛОГИК ТАҲЛИЛ ҚИЛИШ МОДЕЛЛАРИ ВА АЛГОРИТМЛАРИ

05.01.04 – Ҳисоблаш машиналари, мажмуалари ва компьютер тармоқларининг математик ва дастурий таъминоти

ТЕХНИКА ФАНЛАР БЎЙИЧА ФАЛСАФА ДОКТОРИ (PhD) ДИССЕРТАЦИЯСИ АВТОРЕФРАТИ

Тошкент – 2021

Техника фанлари бўйича фалсафа доктори (PhD) диссертацияси мавзуси Ўзбекистон Республикаси Вазирлар Маҳкамаси ҳузуридаги Олий аттестация комиссиясида 2021.3.PhD/T2374 рақам билан рўйхатга олинган.

Диссертация Тошкент ахборот технологиялари университетида бажарилган.

Диссертация автореферати уч тилда (ўзбек, рус, инглиз (резюме)) Илмий кенгаш веб-саҳифасида (www.tuit.uz) ва «Ziyonet» Ахборот таълим порталида (www.ziyonet.uz) жойлаштирилган.

Илмий раҳбар:	**Равшанов Нормахмад** техника фанлари доктори, профессор
Расмий оппонентлар:	**Мухамедиева Дилноз Тулкуновна** техника фанлари доктори, профессор
	Норов Абдисаит Мурадович техника фанлари бўйича фалсафа доктори
Етакчи ташкилот:	**Ўзбекистон Миллий университети**

Диссертация ҳимояси Тошкент ахборот технологиялари университети ҳузуридаги DSc.13/30.12.2019.T.07.01 Илмий кенгашнинг 2021 йил «28» декабрь соат 14⁰⁰ даги мажлисида бўлиб ўтади. (Манзил: 100200, Тошкент шаҳри, Амир Темур кўчаси, 108-уй. Тел.: (99871) 238-64-43, факс: (99871) 238-65-52, e-mail: tuit@tuit.uz).

Диссертация билан Тошкент ахборот технологиялари университети Ахборот-ресурс марказида танишиш мумкин (236 рақам билан рўйхатга олинган). (Манзил: 100200, Тошкент шаҳри, Амир Темур кўчаси, 108-ўй. Тел.: (99871) 238-65-44).

Диссертация автореферати 2021 йил «16» декабрь куни тарқатилди.
(2021 йил «6» декабрь даги 71 рақамли реестр баённомаси).

Р.Х. Хамдамов
Илмий даражалар берувчи
илмий кенгаш раиси,
техника фанлари доктори, профессор

Ф.М. Нуралиев
Илмий даражалар берувчи
илмий кенгаш илмий котиби,
техника фанлари доктори, доцент

М.А. Рахматуллаев
Илмий даражалар берувчи
илмий кенгаш қошидаги илмий семинар раиси,
техника фанлари доктори, профессор

КИРИШ (фалсафа доктори (PhD) диссертацияси аннотацияси)

Диссертация мавзусининг долзарблиги ва зарурати. Диссертация мавзусининг долзарблиги ва зарурати. Жаҳонда ҳужжатларни айлантириш тизимларидан тортиб, кўп тилли ахборот қидириш тизимлари, машинавий таржима тизимлари, савол-жавоб тизимлари ҳамда маҳаллий маълумотлар базалари каби табиий тилдаги матнларни қайта ишлаш тизимларида маълумотларни самарали қайта ишлаш учун лингвистик тузилмалар асосида моделлар, алгоритмлар ва дастурий воситалар ишлаб чиқишга катта эътибор қаратилмоқда. Ушбу тизимлар учун матнлардан токенларни таниб олиш, сўзнинг техник ёки морфологик асоси аниқлаш, сўзларни морфем тузилмасига кўра таҳлил қилиш, сўзнинг турли хил вариантларини шакллантириш ва грамматик хусусиятларини аниқлаш каби морфологик таҳлил қилиш масалаларини назарий ва амалий жиҳатдан тадқиқ қилиш муҳим аҳамият касб этади. Шу сабабли, табиий тиллардаги матнларни қайта ишлаш учун морфологик таҳлил қилиш дастурий воситаларини янада ривожлантириш билан боғлиқ муаммолар АҚШ, Европа мамлакатлари, Хитой, Япония, Ҳиндистон, Россия Федерацияси ҳамда Бирлашган араб амирлиги каби мамлакатлар учун доимо долзарб бўлган.

Дунёда табиий тилларни қайта ишлаш бўйича етакчи мактаблар томонидан сўз шаклларни морфологик таҳлил қилиш учун математик моделлар, алгоритмлар ва дастурий воситаларни ишлаб чиқиш бўйича илмий изланишлар олиб борилмоқда. Бугунги кунда ушбу изланишларнинг натижалари кўп тилли ахборот қидириш тизимларида берилган сўровларни токенизация қилиш, сўзнинг асоси ва турли хил вариантларини шакллантириш, машинавий таржима тизимлари ва савол-жавоб тизимларида гапнинг структурасини белгилашда ҳар бир сўзнинг туркуми ва грамматик хусусиятларини аниқлаш имконини беради.

Ўзбекистонда ҳам табиий тилларни қайта ишлаш жараёнини назарий ва амалий жиҳатдан ўрганиш, ўзбек компьютер лингвистикасини ривожлантириш ва тил муаммоларини ҳал қилиш устида илмий тадқиқотлар олиб борилмоқда. Давлат тилининг замонавий ахборот технологиялари ва коммуникацияларига фаол интеграциялашувини таъминлаш мақсадида республикамиз раҳбарияти томонидан «...ўзбек тилини Интернет жаҳон ахборот тармоғида оммалаштириш, унда муносиб ўрин эгаллашини таъминлаш, дастурий маҳсулотларнинг ўзбекча иловаларини яратиш, ўзбек тилини ўргатувчи компьютер дастурларини кенг миқёсда амалиётга татбиқ қилиш; ўзбек тилидаги матнлар таҳририга мўлжалланган компьютер дастурларини яратиш...»[1] вазифалари белгиланган.

Бу вазифаларни муваффақиятли амалга оширилиши ўзбек тили сўз шаклларининг морфологик таҳлил қилиш моделлари ва алгоритмларини,

[1] Ўзбекистон Республикаси Президентининг 2020 йил 20 октябрдаги ПФ-6084 сон «Мамлакатимизда ўзбек тилини янада ривожлантириш ва тил сиёсатини такомиллаштириш чора-тадбирлари тўғрисида» ги Фармони.// www.xabar.uz

лотин ва кирилл алифбосини эътиборга олган ҳолда ўзбек тилидаги матнларни самарали қайта ишлашга мўлжалланган дастурий воситаларни ишлаб чиқиш технологиялари ривожлантириш бугунги куннинг долзарб масалаларидан бири ҳисобланади.

Ўзбекистон Республикаси Президентининг 2019 йил 21 октябрдаги ПФ-5850-сон «Ўзбек тилининг давлат тили сифатидаги нуфузи ва мавқеини тубдан ошириш чора-тадбирлари тўғрисида» ги Фармони, 2020 йил 5 октябрдаги ПФ-6079-сон «Рақамли Ўзбекистон – 2030» стратегиясини тасдиқлаш ва уни самарали амалга ошириш чора-тадбирлари тўғрисида»ги Фармони, 2020 йил 20 октябрдаги ПФ-6084-сон «Мамлакатимизда ўзбек тилини янада ривожлантириш ва тил сиёсатини такомиллаштириш чора-тадбирлари тўғрисида» ги Фармони, 2021 йил 17 февралдаги ПҚ-4996-сон «Сунъий интеллект технологияларини жадал жорий этиш учун шарт-шароитлар яратиш чора-тадбирлари тўғрисида»ги Қарори ҳамда мазкур соҳага тегишли меъёрий-ҳуқуқий хужжатларда белгиланган вазифаларни амалга оширишда ушбу диссертация иши муайян даражада хизмат қилади.

Тадқиқотнинг республика фан ва технологиялари ривожланишининг устувор йўналишларга мослиги. Мазкур тадқиқот республика фан ва технологиялар ривожланишининг IV. «Ахборотлаштириш ва ахборот-коммуникация технологияларини ривожлантириш» устувор йўналиши доирасида бажарилган.

Муаммонинг ўрганилганлик даражаси. Математик лингвистика, сўз шаклларини морфологик таҳлил қилиш модел ва алгоритмлари асосида табиий тиллардаги матнларни қайта ишлаш учун компьютер тизимларини яратиш ҳамда ривожлантириш кўплаб хорижий олимлар жумладан, Ю.Н. Марчук, И.С. Николаев, О.В. Митренина, Т.М. Ландо, К.К. Боярский, Е.И. Большакова, И.М. Ножов, Е.П. Соснина, И.А. Большаков, Р. Митков, А.В. Луканин, Т.В. Батура, Дж.Ш. Сулейманов, А.Р. Гатиатуллин, Р.А Гильмуллин, А.Б. Альменова ва бошқаларнинг ишларида кўриб чиқилган.

Ўзбекистонда машинавий лингвистика муаммолари, шунингдек, табиий тиллардаги нутқ сигналлари ва матнларни қайта ишлаш, машинавий таржима, туркий тиллар учун электрон луғатларни шакллантириш, ўзбек тили фонетикаси ва морфологияси ни назарий жиҳатдан тадқиқ этиш Ш.А. Назиров, М.М. Арипов, М. М. Мусаев, М.А. Рахматуллаев, Б.Б. Мўминов, М.Х. Хакимовлар бошлигида бир қатор олимлар ва томонидан амалга оширилди.

Бу олимларнинг илмий ишлари бугунги кунда ўзбек компьютер лингвистикасини ўрганиш учун энг нуфусли манба ҳисобланади. Гарчи улар томонидан назарий ва амалий жиҳатдан салмоқли натижаларга эришилган бўлса-да, шунга қарамай ўзбек тилининг сўз шаклларини морфологик таҳлил қилиш модел ва алгоритмлари асосида матнларни автоматик қайта ишлашнинг дастурий тизимларини яратиш ҳозирги вақтгача етарли даражада ўрганилмаган.

Диссертация тадқиқотининг диссертация бажарилган илмий тадқиқот муассасасининг илмий-тадқиқот ишлари режалари билан

боғлиқлиги. Диссертация тадқиқоти Муҳаммад Ал-Хоразмий номидаги Тошкент ахборот технологиялари университети ҳузуридаги АКТ Илмий инновацион маркази илмий-тадқиқот ишлари режасининг И-ОТ-2019-12 «Исломшунослик илмига салмоқли ҳисса қўшган аждодларимиз асарларининг замонавий илмий-амалий талқини ва таҳлилига бағишланган электрон «Ақлли кутубхона» платформасини яратиш ва оммалаштириш» (2019-2021) мавзусидаги лойиҳа доирасида бажарилган.

Тадқиқотнинг мақсади ўзбек тилининг сўз шаклларини морфологик таҳлил қилиш моделлари, алгоритмлари ва вебга йўналтирилган дастурий воситаларини ишлаб чиқишдан иборат.

Тадқиқотнинг вазифалари:

ўзбек тилидаги сўз шаклларининг морфем моделлари ва ўзбек тили учун сўз ясалиши алгоритмларини тадқиқ этиш ҳамда ишлаб чиқиш;

ўзбек тилидаги сўзларнинг морфологик асосини аниқловчи ва сўз шаклларини морфем таҳлил қилувчи алгоритмларни ишлаб чиқиш;

ўзбек тилидаги сўз шаклларининг грамматик хусусиятларини ифодаловчи лексиконни шакллантириш;

матндан токенларни таниб олиш учун чекли автоматдан фойдаланиб продукцион қоидалар асосидаги билимлар базасини ишлаб чиқиш;

ўзбек тилидаги сўзларни морфологик таҳлил қилиш жараёнининг функционал моделини ишлаб чиқиш;

ўзбек тилидаги сўз шаклларини морфологик таҳлил қилишнинг вебга йўналтирилган дастурий воситасини лойиҳалаштириш учун коммуникацион моделлар ва маълумотларнинг ахборот моделларини ишлаб чиқиш.

Тадқиқотнинг объекти сифатида табиий тиллардаги матнларни компьютерда қайта ишлаш жараёни танланган.

Тадқиқотнинг предмети ўзбек тилидаги сўз шакллари морфологик таҳлил қилишнинг морфем модел ва алгоритмлари ва дастурий воситалари.

Тадқиқотнинг усуллари. Диссертация мавзусини ёритишда математик статистика, компьютер лингвистикаси тўпламлар назарияси методлари, шунингдек, тақсимланган веб-иловаларни ва маълумотлар базасини ишлаб чиқиш технологиялари ҳамда лойиҳалаш усулларидан фойдаланилган.

Тадқиқотнинг илмий янгилиги қуйидагилардан иборат:

ўзбек тили орфографияси, пунктуацияси ва лингвистик хусусиятларини ҳисобга олган ҳолда сўз ясалишининг алгоритмлари ҳамда сўз шаклларининг морфем моделлари ишлаб чиқилган;

ўзбек тилидаги аффиксация усули билан ясалган сўз шаклларини морфем таҳлил қилувчи, ўзак морфемаларини аниқловчи алгоритмлар ишлаб чиқилган;

ўзбек тилидаги сўз асослари лексикони ва грамматик теглардан фойдаланиш асосида лексемаларни таниб олиш алгоритми ишлаб чиқилган;

чекли автоматлар тўпламидан иборат продукцион қоидалар асосида ўзбек тилидаги матнлардан токенларни таниб олувчи билимлар базаси ишлаб чиқилган;

ўзбек тилидаги сўз шаклларини морфологик таҳлил қилиш дастурий тизимини лойиҳалаштириш модели, шу жумладан, функционал IDEF0 модели, информацион IDEF1x модели, объектга йўналтирилган синфлар модели, REST архитектураси асосидаги коммуникация модели ишлаб чиқилган.

Тадқиқотнинг амалий натижалари қуйидагилардан иборат:

ўзбек тилининг сўз шаклларини морфологик таҳлил қилишнинг ташқи қидирув тизимларидаги сўровларга хизмат кўрсатувчи веб-сервис сифатида интерактив режимда фаолият юритувчи вебга йўналтирилган дастурий воситаси ишлаб чиқилган.

Ишлаб чиқилган дастурий восита содда туб, содда ясама, қўшма, жуфт, такрорий сўзлар ҳамда қисқартма сўзларнинг имлоси ва тузилишини, шунингдек, аффиксация усули билан ясалган сўз шаклларининг ўзак морфемаларини аниқлаш ва сўз шаклларини морфем таҳлил қилиш сифатини ҳисобга олган ҳолда, ўзбек тилидаги матнларнинг токенизациясини такомиллаштириш орқали республиканинг бир қатор ахборот-кутубхона марказларининг қидирув тизимларида матнли сўровларни қайта ишлаш самарадорлигини ва натижаларини тақдим этиш мувофиқлигини ошириш имкониятини берди.

Тадқиқот натижаларининг ишончлилиги қўйилган муаммоларнинг математик жиҳатдан тўлиқ ва аниқ ифодаланганлиги, лингвистик тушунчалар ўртасидаги ўхшашлик ва фарқларни аниқлашда морфем моделларнинг тўғри қўлланганлиги, назарий маълумотлар ҳамда тадқиқотнинг амалий натижаларининг ўзаро мувофиқ келганлиги билан изоҳланади.

Тадқиқот натижаларининг илмий ва амалий аҳамияти. Тадқиқот натижаларининг илмий аҳамияти шундан иборатки, улар компьютер лингвистикаси методлари ривожига, ўзбек тилининг сўз шаклларини моделлаштириш ва морфологик алгоритмларни ишлаб чиқишга, шунингдек, ўзбек тилидаги матнларни автоматик қайта ишлаш тизимларини лойиҳалаш ҳамда ишлаб чиқиш усулларини яратишга алоҳида ҳисса қўшганлиги билан белгиланади.

Тадқиқот натижаларининг амалий аҳамияти ўзбек тилидаги қидирув сўровларини қайта ишлашни такомиллаштириш, республиканинг бир қатор ахборот-кутубхона ва архив муассасаларида тўлиқ матнли ҳужжатлар кутубхоналари ва библиографик ёзувлар электрон каталогларида қидирувда мувофиқлик натижаларини ошириш мақсадида ишлаб чиқилган «Morphoanalyzer» вебга йўналтирилган дастурий воситасини қўллаш билан изоҳланади.

Тадқиқот натижаларининг жорий қилиниши. Ўзбек тилидаги сўз шаклларини морфологик таҳлил қилиш моделлари, алгоритмлари ҳамда веб-дастурий воситадан фойдаланиш асосида:

Электрон кутубхона тизимида ахборот ресурсларини излашда ўзбек тилидаги матнли сўровларни морфологик таҳлил қилиш учун моделлар, алгоритмлар ва дастурий таъминот Абу Али ибн Сино номидаги Бухоро вилояти Ахборот-кутубхона марказида ҳамда Бухоро давлат университети

Ахборот-ресурс марказида жорий қилинган. (Ўзбекистон Республикаси Президенти Администрацияси ҳузуридаги Ахборот ва оммавий коммуникациялар агенлигининг 2021 йил 22 сентябрдаги 11-3885-сон маълумотномаси). Жорий қилиш натижасида имло хатоларини автоматик тузатиш ва сўзлар имлосининг тўғри вариантларини бериш орқали электрон каталог тизимида ўзбек тилидаги ҳужжатларнинг библиографик тавсифларини тузиш операцияларини бажаришда вақт ва меҳнат сарфини 9-11 % камайиш имконини берган.

Ўзбек тилидаги сўз шаклларини морфологик таҳлил қилиш моделлари, алгоритмлари ва дастурий таъминоти Ўзбекистон Миллий кутубхонасига жорий этилди (Ўзбекистон Республикаси Президенти Администрацияси ҳузуридаги Ахборот ва оммавий коммуникациялар агенлигининг 2021 йил 22 сентябрдаги 11-3885-сон маълумотномаси). Жорий қилиш натижасида морфем таҳлил аниқлиги ҳисобига библиографик ёзувлар электрон катологи ва тўлиқ матнли ҳужжатлар электрон кутубхонаси тизимларида ахборот ресурсларини қидириш натижаларининг мувофиқлигини ошишига хизмат қилган.

Тадқиқот натижаларининг апробацияси. Диссертациянинг назарий ва амалий натижалари 6 та халқаро ва 6 та республика илмий амалий конференцияларида муҳокамадан ўтказилган.

Тадқиқот натижаларининг эълон қилиниши. Диссертация бўйича жами 23 та илмий иш, жумладан Ўзбекистон Республикаси Олий аттестация комиссиясининг докторлик диссертациялари асосий илмий натижаларини чоп этиш тавсия этилган илмий нашрларда 9 та мақола, 6 та республика ва 3 та хорижий журналларда нашр этилган ҳамда 2 та ЭҲМ учун яратилган дастурий воситаларни қайд қилиш гувоҳномалари олинган.

Диссертациянинг ҳажми ва тузилмаси. Диссертация кириш, тўртта боб, хулоса, фойдаланилган адабиётлар рўйхати ва иловалардан иборат. Диссертация ҳажми 116 бетни ташкил этади.

ДИССЕРТАЦИЯНИНГ АСОСИЙ МАЗМУНИ

Кириш қисмида диссертация мавзусининг долзарблиги ва зарурати асосланган, тадқиқотнинг Ўзбекистон Республикаси фан ва технологиялари ривожланишининг устувор йўналишларига мослиги кўрсатилган, тадқиқот мақсади, вазифалари, объекти, предмети, илмий янгилиги, амалий натижалари тавсифланган, тадқиқот натижаларининг ишончлиги, илмий ва амалий аҳамияти асосланган, тадқиқот натижаларини жорий қилиш ҳолати, нашр этилган ишлар ва диссертация тузилиши бўйича маълумотлар келтирилган.

Диссертациянинг «**Табиий тилларда матнни қайта ишлаш соҳасининг ҳолати ва ривожланиш тенденциялари**» деб номланган биринчи боби бешта параграфдан иборатдир. Биринчи параграфда табиий тилларни қайта ишлаш ривожланиш босқичлари, табиий тилларни қайта ишлашнинг бешта асосий компонеталари ҳақида маълумотлар келтирилган. Иккинчи параграфда морфологик таҳлил қилишнинг ёндашувлари афзаллиги

ва камчилигига кўра таҳлил қилинган. Учинчи параграфда морфологик таҳлил қилиш дастурларининг таснифи ва имкониятларига кўра қиёсий таҳлили келтирилган. Тўртинчи параграфида чет эл давлатларида ва республикамизда агглютинатив тилларни морфологик таҳлил қилиш дастурларининг функционал имкониятларига кўра олиб борилаётган илмий тадқиқот натижалари таҳлил қилинган. Бобнинг бешинчи параграфида ушбу муаммо бўйича илмий манбаларни батафсил таҳлил қилиш асосида тадқиқот мақсади ва вазифалар белгиланган.

Диссертациянинг «**Ўзбек тили сўз шаклларини морфологик таҳлил қилиш модел ва алгоритмлари**» деб номланган иккинчи бобида токенизация алгоритмлари ва токенизация бўйича илмий тадқиқотлар натижалари, сўз шаклларини асосини аниқловчи мавжуд стемминг алгоритмларини ишлаб чиқишдаги ёндашувлар, ўзбек тилидаги матнлардан токенларни ажратиб олиш учун ифодалар ва алгоритм, сўз шаклларининг морфем моделлари асосида сўзларни морфологик асосини аниқлаш ва морфем таҳлил қилиш алгоритмлари, сўз туркуми ва грамматик хусусиятларини аниқлаш алгоритмлари ишлаб чиқишга бағишланган.

Ўзбек тилидаги токенларни шартли равишда икки гуруҳга ажратилди Биринчи гуруҳга қисқартма шаклдаги, иккинчи гуруҳга сўз ясалишига кўра токенлар (қўшма, жуфт ва такрорий сўзлар) киритилди. Аксарият ҳолларда мавжуд токенизаторлар иккинчи гуруҳга тегишли токенларни ажратиб олишда семантик хатоликка йўл қўяди. Қайд этилган муаммони инобатга олган ҳолда, иккинчи гуруҳга мансуб қўшма, жуфт, такрорий сўзларни ёзишнинг имло қоидаларини тавсифловчи математик ифодалар таклиф қилинди:

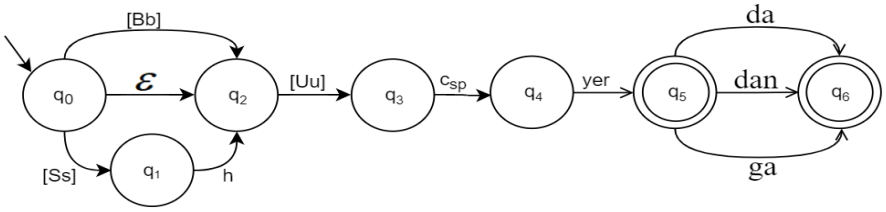
$$w_{qs} = w'_{qs} + c_{sp} + w''_{qs}, \quad (1)$$

$$w_j = w'_j + \rho + w''_j, \quad (2)$$

$$w_t = w'_t + \rho + w'_t, \quad (3)$$

Бу ерда, w_{qs} – қўшма сўз, w_j – жуфт сўз, w_t – такрорий сўз. w'_{qs}, w''_{qs}, w'_j, w''_j w'_t – мустақил маънога эга бўлган сўз, сўзнинг бу икки қисми янги луғавий маънони ифодалайди, ρ – дефис белгиси, c_{sp} – пробел белгиси.

Юқоридаги ифодалар орқали тасвирланган токенларни матндан таниб олиш учун чекли автомат(ЧА) ишлаб чиқилди. Масалан, w_{qs} турдаги 3 та қўшма сўзни (равиш сўз туркуми мисолида u yerga, bu yerda, shu yerdan) 1 та ЧА орқали қуйидагича тасвирлаш мумкин (1-расм).



1-расм. w_{qs} - турдаги 3та қўшма сўз учун чекли автомат

1-расмда тасвирланган ЧА қуйидаги регуляр ифода орқали ифодаланади: $((sh)|b)^{*}u) \setminus s\ (dan|da|ga)?$

Жуфт ва такрорий ҳамда қисқартма сўзларнинг математик ифодалари учун ҳам чекли автомат шу тартибда тузилади ва регуляр ифода кўринишига келтирилади. Ҳар бир математик ифода асосида тузилган регуляр ифодалар ёрдамида токенизация алгоритми тузилади. Матнларни токенизация қилиш алгоритми қуйидаги қадамлардан иборат:

1-қадам. Алгоритм бошланади.

2-қадам. Сатр туридаги «text» ўзгарувчисига матн киритилади (клавиатура, ахборот алмашинув буфери, ёки файл орқали).

3-қадам. Матндаги мавжуд қўшма сўз биграмм токенлар ва уларнинг маркеридан иборат «ng» луғат ҳосил қилинади. Қўшма сўз биграмм токенлар билимлар базаси орқали ажратиб олинади.

4-қадам. Қўшма сўз биграмм токенлар луғати бўш бўлмаса, 5-қадамга ўтилади, акс ҳолда 6-қадамга ўтилади.

5-қадам. «text» даги мавжуд қўшма сўз биграмм токенлар маркерланади.

6-қадам. «text» ўзгарувчисидаги матн пробел бўйича руйхатга айлантирилади. Агар «ng» луғат бўш бўлмаса, рўйхатдаги маркерлар қўшма сўзга алмаштирилади.

7-қадам. Рўйхат туридаги «x» ўзгарувчининг қиймати экранга чиқарилади.

8-қадам. Алгоритм ўз ишини тугатади.

Иккинчи бобнинг §2.2 параграфида сўз шаклларини морфологик асосини аниқлаш модел ва алгоритмлари ишлаб чиқилган. Ўзбек тилидаги сўз шаклларини морфологик асосини аниқлашда сўзнинг структураси муҳим аҳамият касб этади. Шу муносабат билан қуйидаги математик ифодалар орқали тасвирланган лингвистик қоидалар асосидаги аффиксация усули билан ясалган сўз шаклларининг морфемик моделлари таклиф этилди:

$$w_{sy} = K + \left\{ \sum_{l=1}^{m} sx_l \right\} + \left\{ \sum_{i=1}^{n} af_i \right\}. \qquad (5)$$

Бу ерда K- содда туб сўз (ўзак), sx- сўз ясовчи аффикс (суффикс), af - аффикслар, Σ- 0,1, …n кетма-кет конкатенация операцияси, l – суффикслар миқдори индекси, i – аффикслар миқдори индекси. w_{sy} - содда ясама сўз. Фикримизни далиллаш учун от сўз туркумидаги «gul-don-lar-imiz-ni» сўз шаклини таҳлил қиламиз. Бу мисолда K- «gul», sx- «-don», af - «-lar », «-imiz »,«-ni », «gul-don-lar-imiz-ni» - w_{sy} сўз шакли.

$$w_{psy} = \{\Pr\} + K + \left\{ \sum_{l=1}^{m} sx_l \right\} + \left\{ \sum_{i=1}^{n} af_i \right\}. \qquad (6)$$

Бу ерда, Pr – сўз ясовчи аффикс (префикс), w_{psy} - содда ясама сўз(префиксли). Масалан, «be-adab-garchilik-lar-ni» сўз шаклини таҳлил

қиламиз. Бунда « be-» - Pr , «adab»- K, «-garchilik»- sx ,«-lar»,«-ni»- af , «be-adab-garchilik-lar-ni»- w_{psy} сўз шакли.

$$w_{st} = K + \left\{ \sum_{i=1}^{n} af_i \right\}. \qquad (7)$$

Бу ерда, w_{st} - содда туб сўз. Масалан, «kitob-lar-imiz-ning» сўз шаклини таҳлил қиламиз. Бунда «kitob»- K, «-lar», «-imiz», «-ning»- af ,«kitob-lar-imiz-ning» - w_{st} сўз шакли.

(5), (6), (7) морфем моделлар учун алгоритм ишлаб чиқилди w_{st} содда туб сўзда аффикслар ўнг томондан, w_{psy} ва w_{sy} содда ясама сўзларда эса аффикслар чап ва ўнг томондан қўшилади. Бу каби сўзларнинг аффикслари чап ва ўнг томондан ўчириш орқали асоси топилади.

Алгоритм қуйидаги қадамлардан иборат:

1-қадам. Алгоритм бошланади.

2-қадам. «S» номли сатр ўзгарувчига сўз киритилади.

3-қадам. «S» ўзгарувчининг қиймати «SS» номли сатр ўзгарувчига ўзлаштирилади, яъни сўзнинг нусхаси яратилади.

4-қадам. «SS» қиймати dbStemShow(SS) функция орқали маълумотлар базасидан изланади. Агар сўз нусхаси маълумотлар базасида мавжуд бўлса, функция рост қийматни қайтаради, акс ҳолда ёлғон қийматни қайтаради. Кейин эса dbStemShow(SS) функциясининг қайтарадиган қиймати мантиқий турдаги «b» номли ўзгарувчига ўзлаштирилади.

5-қадам. «b» ўзгарувчининг қиймати рост бўлса, яъни «SS» қиймати маълумотлар базасида мавжуд бўлса, унда 16-қадамга ўтилади.

6-қадам. preffunc(SS) функцияси орқали киритилган сўзни префикси изланади. Агар сўзда префикс бор бўлса, preffunc(SS) функциянинг қайтарадиган қиймати «pref» номли сатрий ўзгарувчисига ўзлаштирилади. Акс ҳолда «pref» ўзгарувчисига «null» қиймат ўзлаштирилади.

7-қадам. Агар «pref» ўзгарувчисининг қиймати бўш бўлмаса, яъни префикс мавжуд бўлса, 13-қадамга, мавжуд бўлмаса, 8-қадам ўтилади.

8-қадам. Сўз узунлиги бўйича i:=(0,Length(S)-1,1) 1 та қадамли итерация бошланади.

9-қадам. Ҳар бир итерацияда сатрни тескари тартибда i-узунликдаги қисми ўчирилиб, dbStemShow(S.SubStr(0,Length(S)-i)) функцияси орқали маълумотлар базасидан изланади ва қайтарадиган қиймати «b» ўзгарувчига ўзлаштирилади.

10-қадам. Ҳар бир итерацияда dbStemShow(S.SubStr(0,(Length(S)-1)-i)) функциянинг мантиқий қиймати текширилади. Агар «b» ўзгарувчининг қиймати рост бўлса яъни сўз маълумотлар базасидан топилса, унда 11-қадам ўтилади, акс ҳолда кейинги итерация учун 8-қадамга ўтилади.

11-қадам. «SS» ўзгарувчисига S.SubStr(0,Length(S)-i) функциясининг қиймати ўзлаштирилади ва 16-қадамга ўтилади.

12-қадам. Агар итерация жараёнида асос топилмаса, «SS» ўзгарувчига «S» ўзгарувчининг қиймати ўзлаштирилади ва 16-қадамга ўтилади.

13-қадам. Сўзни префиксдан кейинги қисми кесиб олиниб, dbStemShow(S.SubStr(Length(pref),Length(S)-1) функцияси орқали маълумотлар базасидан изланади ва қайтарадиган қиймати «b» ўзгарувчига ўзлаштирилади.

14-қадам. «b» ўзгарувчининг қиймати рост бўлса, яъни сўз маълумотлар базасида мавжуд бўлса, 15-қадамга ўтилади. Сўз маълумотлар базасида мавжуд бўлмаса, 8-қадамга ўтилади.

15-қадам. Сўзнинг префиксдан кейинги қисми S.SubStr(Length(pref),Length(S)) орқали кесиб олиниб, «SS» ўзгарувчисига ўзлаштирилади.

16-қадам. «SS» ўзгарувчининг қиймати экранга чиқарилади.

17-қадам. Алгоритм тугайди.

Жуфт ва такрорий сўзларнинг асосини топиш модел ва алгоритмлари диссертациянинг §2.2 параграфида келтирилган.

Иккинчи бобнинг §2.3 параграфида сўз шаклларини морфемик таҳлил алгоритми ишлаб чиқилган. Сўзни морфем структурасига кўра таҳлил қилинганда, сўз морфемаларга ажратилади ва маъноли қисмлар аниқланади. Сўзларни морфемик таҳлил қилишда сўзнинг ясалиш структурасининг морфем моделидан фойдаланиб, морфемик таҳлил қилиш алгоритмини тузамиз. (5), (6), (7) - морфем моделлар орқали ифодаланган сўз шаклларини морфотактика қоидалари асосида ЧА фойдаланиб, морфемик таҳлил қилиш алгоритми тузилди. ЧА қуйидаги бешта элемент орқали аниқланади:

$$M = (\theta, Q, q_0, F, \delta),$$

М – ЧА, $\theta = \{PR, K, SX, AF\}$ – кирувчи алифбо, $Q = \{q_0, q_1, q_2, q_3, q_4\}$ – ҳолатлар тўплами, $\delta(Q,\theta): Q \times \theta \to Q'$ – ўтиш функцияси, q_0 , $(q_0 \in Q)$ – бошланғич ҳолат, F , $(F \subseteq Q)$ – натижавий ҳолатлар тўплами.

Чекли автомат орқали содда туб сўз шаклини аниқлаш учун ўтиш функцияси қуйидаги қийматларни қабул қилиши керак:

$$\delta(q_0, K) = q_2, \delta(q_2, \text{AF}) = q_4, \delta(q_4, AF) = q_4.$$

Содда ясама (суффиксли) сўз шаклини аниқлаш учун ўтиш функцияси қуйидаги қийматларни қабул қилиши керак:

$$\delta(q_0, K) = q_2, \delta(q_2, SX) = q_3, \delta(q_3, \text{SX}) = q_3, \delta(q_3, AF) = q_4.$$

Содда ясама (префиксли) сўз шаклини аниқлаш учун ўтиш функцияси қуйидаги қийматларни қабул қилиши керак:

$$\delta(q_0, PR) = q_1, \delta(q_1, K) = q_2, \delta(q_2, SX) = q_3, \delta(q_3, \text{AF}) = q_4,$$

$$\delta(q_3, \text{SX}) = q_3, \delta(q_3, AF) = q_4, \delta(q_4, AF) = q_4.$$

Иккинчи бобнинг §2.4 параграфида сўз туркумлари ва уларнинг грамматик хусусиятларни аниқлаш алгоритми ишлаб чиқилган. Сўзларнинг туркумини ва грамматик хусусиятларини аниқлаш алгоритмини тузишда қуйидаги босқичлар амалга оширилади: сўзни морфемаларга ажратиш; сўз туркумлари учун ифодаловчи «тэг»лар лексиконини ишлаб чиқиш; ҳар бир сўз

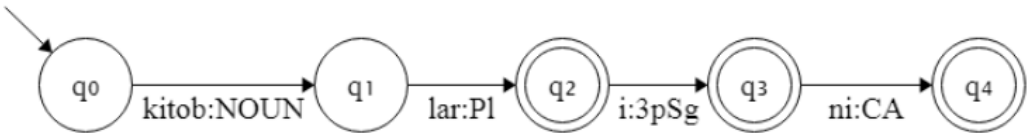
туркуми учун грамматик хусусиятларни ифодаловчи тэгларнинг лексиконини ишлаб чиқиш. Сўз шаклларининг грамматик хусусиятларини аниқлаш алгоритмини тузиш учун чекли шакл алмаштиргичдан фойдаланилди .

Чекли шакл алмаштиргич(ЧША) қуйидаги етти элементдан иборат:

$$M_{чша} = (Q, \theta_{in}, \theta_{out}, \delta, \omega, q_0, \mathrm{F}).$$

$M_{чша}$ – чекли шакл алмаштиргич, Q – ҳолатлар тўплами, $\theta_{in} = \{\mathrm{A}, \mathrm{l}_1, \mathrm{s}_1, \mathrm{s}_2\}$ – кирувчи алифбо, $\theta_{out} = \{N, "Pl", "3pSg", "CA"\}$ – чиқувчи алифбо, $\delta(Q, \theta_{in}) : Q \times \theta_{in} \to Q'$ – ўтиш функцияси, $\omega(Q, \theta_{in}) : Q \times \theta_{in} \to \theta_{out}$ – чиқиш функцияси, $q_0 \in Q$ – ЧША бошланғич ҳолати, $F \subseteq Q$ – натижавий ҳолатлар тўплами.

Кирувчи алифбода А-асос, l_1 –кўплик аффикси, s_1-эгалик аффикси, s_2-келишик аффикси элементлари қуйидагича аниқланади. Масалан, кирувчи алифбо сифатида от сўз туркумига тегишли «kitoblarini» сўз шаклини олсак, бу сўз шакли қуйидаги чекли шакл алмаштиргич кўринишида тасвирланади:



2-расм. От сўз туркуми учун ЧША

2-расмда келтирилган ЧША q_2, q_3, q_4 натижавий ҳолатларига эга ва бу орқали «kitoblar», «kitoblari», «kitoblarini» каби сўз шакллари учун грамматик тэгларни белгилаш мумкин. ЧША сўзни грамматик теглаш жараёни 1-жадвалда тасвирланган.

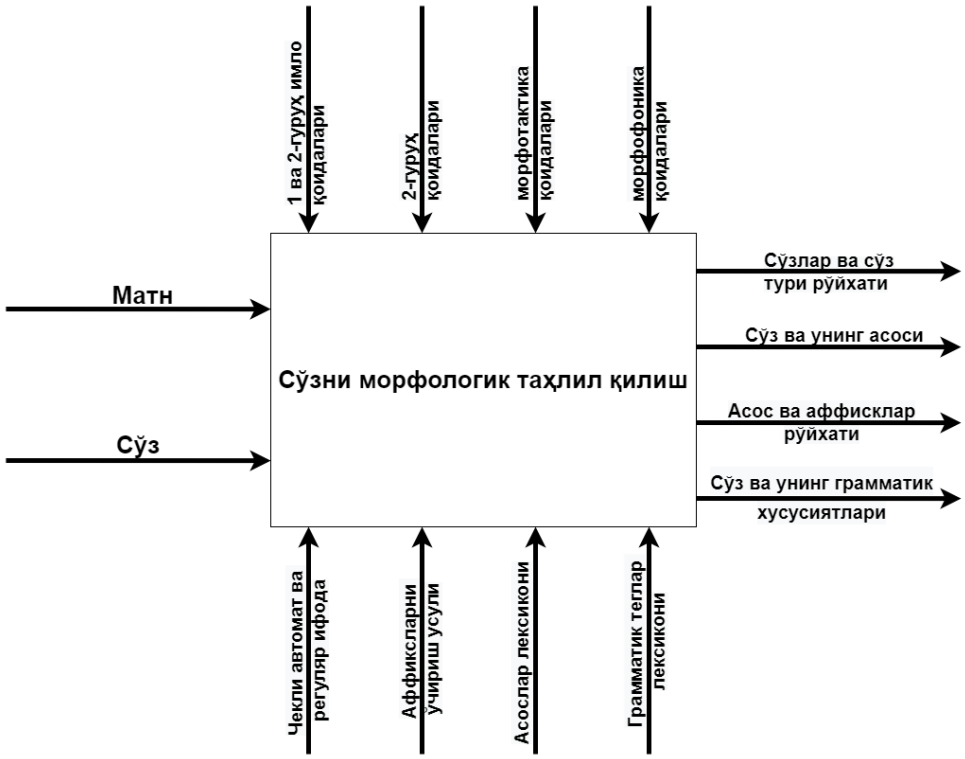
1-жадвал. ЧША учун ўтишлар жадвали

№	kitob	Lar	i	ni
q_0	(q_1, kitob)	-	-	-
q_1	-	(q_2,Pl)	-	-
q_2	-	-	(q_3,3pSg)	-
q_3	-	-	-	(q_4,ni)
q_4	-	-	-	-

Диссертациянинг **«Ўзбек тилидаги сўзларни морфологик таҳлил қилиш дастурий воситалари»** деб номланган учинчи бобида «Morphoanalyzer» дастурий воситасининг архитектураси ва IDEF0 функционал модели, объектга йўналтирилган синфлар модели, маълумотлар базасининг IDEF1x модели, продукцион қоидалар асосидаги билимлар базасини ишлаб чиқиш масалалари кўрилган.

«Morphoanalyzer» дастурий воситаси уч қатламли архитектура асосида қурилган бўлиб, мижоз қатлами, мантиқий қатлам ва маълумотлар қатламидан иборат. Мантиқий қатлам кирувчи маълумот сифатида сўз ва матн қабул

қилади ва чиқувчи маълумот сифатида сўзлар ва сўз тури рўйхати, сўз ва унинг асоси, асос ва аффикслар рўйхати, сўз ва унинг грамматик хусусиятларини қайтаради. «Morphoanalyzer» дастурий воситасининг мантиқий қатлами 3-расмда IDFE0 модели орқали тасвирланган.



3-расм. «Morphoanalyzer» дастурий воситасининг IDFE0 модели

Логика қатламининг IDFE0 модели тўртта функционал модуллардан ташкил топган**:** матнларни токенларга ажратиш модули; сўзларни нормаллаштириш модули; сўзни морфемик таҳлил қилиш модули; сўзни грамматик хусусиятларини топиш модули.

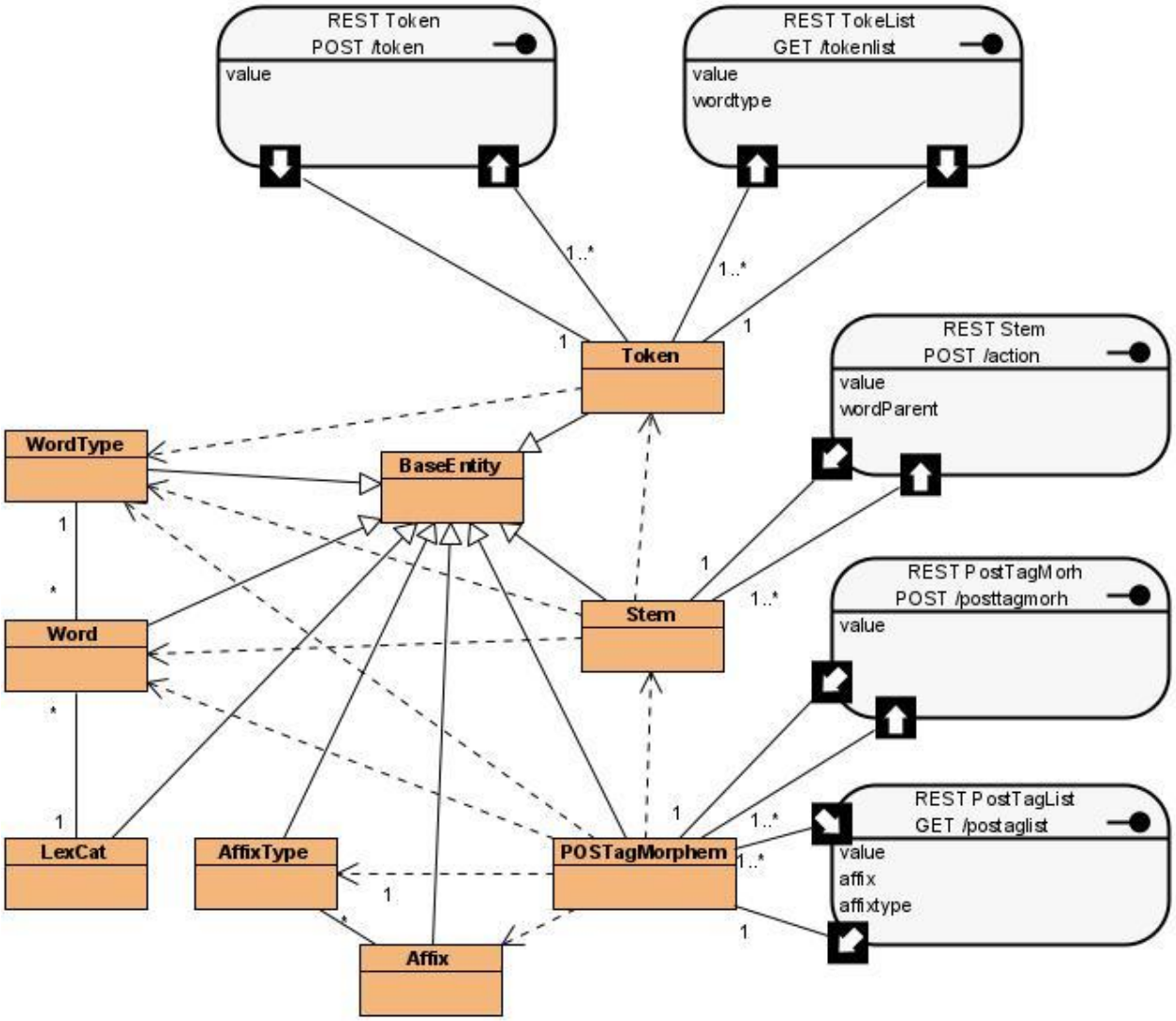
Ушбу модуллар вазифасини қуйидаги синфлар бажаради**:** масалан, матнларни токенларга ажратиш модули вазифасини “Token” синфи, сўзларни нормаллаштириш модули вазифасини “Stem” синфи, сўзни морфемик таҳлил қилиш ва грамматик хусусиятларини топиш модули вазифасини “PostTagMorphem” синфлари амалга оширади.

«Morphoanalyzer» дастурий воситасининг ушбу синфлари фойдаланувчига компьютер тармоғи орқали REST API архитектураси асосида ахборот алмашиниш имкониятини беради (4-расм).

«Morphoanalyzer» дастурий воситаси 2-жадвалда келтирилган бўлиб, URL манзиллар асосида ахборот алмашиниш имкониятини беради.

Масалан, «kitobxon» сўзини юборганимизда натижа JSON кўринишида қуйидагича қайтарилади:

{"value": "kitobxon", "affix": "xon", "affixtype": "so’z yasovchi suffuks"}.



4-расм. Синфлар модели билан REST API ўртасидаги коммуникацион модел

2-жадвал. REST API билан ахборот алмашиниш манзиллари

API учун номлари	URL манзиллар	Қўлланиладиган метод
Token	token/api_key/text	GET
TokenList	/tokenList/text	GET
Stem	/stem/value	GET
PostTag	/posttag/value	GET
PostTagList	/posttaglist/value	GET

«Morphoanalyzer» дастурий воситасининг ҳар бир функционал модули ўз вазифасини бажаришда продукцион қоидалар асосидаги билимлар базасидан фойдаланади. Масалан, токенизация алгоритми қўшма сўзларни ажратиб олиш қуйидагича амалга оширилади**:**

$$i =< Q; \mathrm{P}; \mathrm{A} \rightarrow \mathrm{B}; \mathrm{N} > .$$

Бу ерда, *i* – Токенизация(продукция номи), *Q* – Матндан ҳосил қилинган токенларнинг биграмм рўйхати(қўлланиладиган соҳа), P – билимлар базасида ҳар бир биграмм элементининг 0-индекслари комбинациясига номи мос келадиган қоидалар тўпламининг мавжудлигини текширишнинг бошланғич шарти, A – регуляр ифодага мос келувчи шарт(қоида), B– сўзларни маркировка ҳаракати, N – қўшма сўзни луғатга қўшиш ҳаракати.

Масалан, "har yerda"→"hy". Агар “hy” номли қоидалар тўплами бўлса, унда продукция фаоллаштириш шарти бажарилади, агар рост бўлса, белгилаш амалга оширилади ва луғатга “har yerda” сўзи киритилади.

Ушбу мисолда «hy» номли тўпламда 3 та қоида мавжуд бўлиб, киритилган сўзга 3 та қоида бирма-бир қўлланилади. Цикл киритилган сўзга барча қоидалар қўлланилганда ёки бирор-бир қоида рост қиймат қайтарганда тўхтайди (3-жадвал).

3-жадвал. Қўшма сўзлар учун билимлар базасининг ишлаш тамойили

Максимал итерация сони	Ишчи хотира (Киритилган сўз)	Тўплам	Тўпламнинг низоли қоидалари	Қўлланиладиган қоида
3	har yerda	hy	Q16, Q18, Q33	16

Диссертациянинг **«Morphoanalyzer» дастурий воситанинг тадбиқи»** деб номланган тўртинчи бобида морфологик таҳлил дастурий воситасининг аппарат-дастурий таъминотга бўлган талаблари, фойдаланувчининг функционал имкониятлари дастурий таъминотдан фойдаланиш йўриқномаси, эксперт томонидан берилган маълумотлар асосида олинган натижалар ҳамда дастурий таъминотни ишлаб чиқариш жараёнидаги тадбиқи ҳақида маълумотлар келтирилган.

«Morphoanalyzer» дастурий воситасининг икки турдаги фойдаланувчилари мавжуд: лингвист эксперт (администратор) ва фойдаланувчи. «Morphoanalyzer» дастурий воситаси фойдаланувчиларга бир неча ҳаракатларни бажариш имкониятини беради. Бу ҳаракатлар прецедент деб номланади. МТҚДВнинг ҳар бир прециденти ўз спецификациясига эга (4-жадвал).

4-жадвал. Матнни токенлаш прецеденти учун спецификация

Прецедент номи:Матнни токенлаш
Қисқача изоҳ: Матндан токенларни ажратади ва турини аниқлайди.
Вазифани бажарадиган синф: Token
Асосий фойдаланувчи: Фойдаланувчи
Бошланғич шартлар: Ихтиёрий браузер дастурининг мавжудлиги
Прецедентнинг вазифасини бажариш босқичлари: 1. Матндан токенлар ажратилади**.** 2. Токенларнинг тури аниқланади**.**

Охирги шартлар: Йўқ

«Morphoanalyzer» дастурий воситасида эртак, қонуний ҳужжатлар, ҳадислар, ўзбек ҳалқ мақоллари каби турли хил жанрдаги матнлардан фойдаланилди ва қуйидагича натижа берди (5,6-жадваллар).

5-жадвал. Токенизация алгоритми натижалари

№	Танланган корпуслар	Сўзлар сони	Сўз турлари		
			Қўшма	Жуфт такрорий	Қисқартма
1	Эртаклар	2100	147(7%)	84(4%)	0
2	Қонуний ҳужжатлар	2150	194(9%)	43(2%)	1769
3	Ҳадислар	2000	110(5.5%)	30(1.5%)	0

6-жадвал. Стемминг алгоритми натижалари

№	Танланган корпуслар	Сўзлар сони	Тўғри асослар
1	Эртаклар	1246	96%
2	Қонуний ҳужжатлар	1290	92.4%
3	Ҳадислар	1285	88.9%
4	Ўзбек халқ мақоллари	1200	94%

ХУЛОСА

«Ўзбек тили сўз шаклларини морфологик таҳлил қилиш моделлари ва алгоритмлари» мавзусидаги диссертация иши доирасида олинган асосий илмий натижалар қуйидагилардан иборат:

1. Табиий тиллардаги матнларни компьютерда қайта ишлаш жараёнининг асосий хусусиятлари тадқиқ қилинди ва морфологик таҳлил масалаларини ечиш учун моделлар, алгоритмлар ва дастурларни ишлаб чиқишдаги мавжуд ёндашувларнинг қиёсий таҳлили ўтказилди.
2. Агглютинатив тилларни морфологик таҳлил учун дастурларнинг функционал имкониятлари танқидий таҳлил қилинди.
3. Матнлардан қўшма, жуфт, такрорий сўзлар ва қисқартма сўзларни таниб олиш учун чекли автоматлар, шунингдек, ўзбек тилининг лингвистик хусусиятлари, имло қоидалари ва тиниш белгиларини ҳисобга олган ҳолда сўз ясалиш алгоритмлари ишлаб чиқилган.
4. Ўзбек тилида аффиксация усули билан ясалган сўз шаклларининг ўзак морфемаларини аниқлаш ва морфемик таҳлил қилиш алгоритмлари ишлаб чиқилди;
5. Ўзбек тилидаги сўзларнинг асослар лексикони ва грамматик теглардан фойдаланиш асосида лексемаларни таниб олиш алгоритми ишлаб чиқилди;

6. Ўзбек тилидаги матнлардан токенларни таниб олиш учун чекли автомат тўпламидан иборат продукцион қоидалар асосида билимлар базаси ишлаб чиқилди;

7. Ўзбек тилидаги сўз шаклларини морфологик таҳлил қилиш дастурий тизимини лойиҳалаш моделлари, жумладан, функционал IDEF0 ва IDEF1x ахборот моделлари, объектга йўналтирилган синфлар модели ва REST архитектураси асосидаги коммуникация модели ишлаб чиқилди;

8. Ўзбек тилининг сўз шаклларини морфологик таҳлил қилиш учун «Morphoanalyzer» веб-га йўналтирилган дастурий воситаси ишлаб чиқилган бўлиб, ушбу дастурий восита ҳам интерфаол тарзда, ҳам ташқи қидирув тизимларининг сўровларига хизмат кўрсатиш учун веб-сервис сифатида фаолият кўрсатади.

9. Ишлаб чиқилган «Morphoanalyzer» дастурий воситаси электрон каталог тизимларида ўзбек тилидаги ҳужжатларнинг библиографик тавсифларини тузишда операцияларни бажариш тезлигини 9-11% га ошириш имкониятини берди, шунингдек, ўзбек тилидаги сўз шаклларини морфемик таҳлил қилиш сифатини ошириш ва сўзларнинг имлоси ва тузилишини ҳисобга олган ҳолда ўзбек тилидаги матнларнинг токенизациясини такомиллаштириш хисобига республиканинг бир қатор ахборот-кутубхона марказларининг қидирув тизимларида матн сўровларини қайта ишлаш ва натижаларнинг мувофиқ тарзда тасвирлаш самарадорлигини ошириш имкониятини берди.

НАУЧНЫЙ СОВЕТ DSc.13/30.12.2019.T.07.01 ПО ПРИСУЖДЕНИЮ УЧЕНЫХ СТЕПЕНЕЙ ПРИ ТАШКЕНТСКОМ УНИВЕРСИТЕТЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

ТАШКЕНТСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

БАКАЕВ ИЛХОМ ИЗАТОВИЧ

МОДЕЛИ И АЛГОРИТМЫ МОРФОЛОГИЧЕСКОГО АНАЛИЗА СЛОВОФОРМ УЗБЕКСКОГО ЯЗЫКА

05.01.04 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ ДИССЕРТАЦИИ ДОКТОРА ФИЛОСОФИИ (PhD) ПО ТЕХНИЧЕСКИМ НАУКАМ

Ташкент – 2021

Тема диссертации доктора философии (PhD) по техническим наукам зарегистрирована в Высшей аттестационной комиссии при Кабинете Министров Республики Узбекистан за номером B2021.3.PhD/T2374.

Диссертация выполнена в Ташкентском университете информационных технологий.

Автореферат диссертации на трех языках (узбекский, русский, английский (резюме)) размещен на веб-странице научного совета (www.tuit.uz) и на Информационно-образовательном портале «ZiyoNet» (www.ziyonet.uz).

Научный руководитель:	**Равшанов Нормахмад** доктор технических наук, профессор
Официальные оппоненты:	**Мухамедиева Дилноз Тулкуновна** доктор технических наук, профессор
	Норов Абдисаит Мурадович доктор философии по техническим наукам
Ведущая организация:	**Национальный университет Узбекистана**

Защита диссертации состоится «28» декабря 2021 г. в 14⁰⁰ часов на заседании Научного совета DSc.13/30.12.2019.T.07.01 при Ташкентском университете информационных технологий. (Адрес: 100200, г. Ташкент, ул. Амира Темура, 108. Тел.: (99871) 238-64-43; факс: (99871) 238-65-52; e-mail: tuit@tuit.uz).

С диссертацией можно ознакомиться в Информационно-ресурсном центре Ташкентского университета информационных технологий (регистрационный номер № 236). (Адрес: 100200, г. Ташкент, ул. Амира Темура, 108. Тел.: (99871) 238-65-44).

Автореферат диссертации разослан « 16 » декабря 2021 года.
(протокол рассылки № 41 от « 6 » декабря 2021 г.).

Р.Х. Хамдамов
Председатель научного совета
по присуждению учёных степеней,
доктор технических наук, профессор

Ф.М. Нуралиев
Ученый секретарь научного совета
по присуждению учёных степеней,
доктор технических наук, доцент

М.А. Рахматуллаев
Председатель научного семинара
при научном совете
по присуждению ученых степеней,
доктор технических наук, профессор

ВВЕДЕНИЕ (аннотация диссертации доктора философии (PhD))

Актуальность и востребованность темы диссертации. В мире уделяется большое внимание разработке моделей, алгоритмов и программного обеспечения на основе лингвистических структур для эффективной обработки данных в системах обработки текстов на естественных языках – от систем управления документами до многоязычных систем поиска информации, а также систем машинного перевода, систем «вопрос-ответ» и локальных баз данных. Для разработки подобных систем важно исследование теоретических и практических аспектов морфологического анализа, таких как распознавание токенов в текстах, определение производящей основы слов, анализ слов в соответствии их с морфемной структурой, формирование различных вариантов слов и определение грамматических особенностей. Поэтому проблемы, связанные с дальнейшим развитием программных средств морфологического анализа для обработки текстов на естественных языках, всегда были актуальны для развитых стран, таких как США, страны Европы, Китай, Япония, Индия, Российская Федерация и Объединенные Арабские Эмираты.

В мире ведущие научные школы в области обработки текстов на естественных языках проводят исследования по разработке математических моделей, алгоритмов и программных средств для морфологического анализа словоформ. Результаты этих исследований, сегодня, позволяют осуществлять токенизацию текстовых запросов в многоязычных поисковых системах, выявлять основу и различные варианты слов, определять категорию и грамматические особенности каждого слова при определении структуры предложений в системах машинного перевода и системах «вопрос-ответ».

В Узбекистане также успешно проводятся научные исследования по теоретическому и практическому изучению процесса обработки текстов на естественном языке, по развитию узбекской компьютерной лингвистики и решению языковых проблем в целом.

В целях обеспечения активной интеграции государственного языка с современными информационными технологиями и коммуникациями, руководством нашей республики определены такие задачи как «...популяризация и обеспечение занятия достойного места узбекского языка во всемирной информационной сети Интернет; создание приложений программных продуктов на узбекском языке; широкое внедрение в практику компьютерных программ по обучению узбекскому языку; создание программ для редактирования текстов на узбекском языке...»[2].

Успешное выполнение указанных задач определяет востребованность дальнейшего развития математических моделей и алгоритмов морфологического анализа словоформ узбекского языка, а также технологий разработки программных продуктов, предназначенных для эффективной

[2] Указ Президента Республики Узбекистан №УП-6084 от 20 октября 2020 года «О мерах по дальнейшему развитию узбекского языка и совершенствованию языковой политики в стране».

обработки текстов на узбекском языке с учетом особенностей латинской и кириллической графики.

Настоящее диссертационное исследование в определенной степени служит реализации задач, определенных Указами Президента Республики Узбекистан № УП-5850 от 21 октября 2019 года «О мерах по кардинальному повышению роли и авторитета узбекского языка в качестве государственного языка», № УП-6079 от 5 октября 2020 года «Об утверждении Стратегии «Цифровой Узбекистан – 2030» и мерах по ее эффективной реализации», № УП-6084 от 20 октября 2020 года «О мерах по дальнейшему развитию узбекского языка и совершенствованию языковой политики в стране», Постановлением Президента Республики Узбекистан № ПП-4996 от 17 февраля 2021 года «О мерах по созданию условий для ускоренного внедрения технологий искусственного интеллекта», а также другими нормативно-правовыми документами, принятыми в данной сфере.

Соответствие исследования приоритетным направлениям развития науки и технологий республики. Данное исследование выполнено в соответствии с приоритетным направлением развития науки и технологий Республики Узбекистан IV. «Информатизация и развитие информационно-коммуникационных технологий».

Степень изученности проблемы. Проблемы математической лингвистики, создания и развития компьютерных систем обработки текстов на естественных языках на основе моделей и алгоритмов морфологического анализа словоформ рассматривались в работах многих зарубежных ученых, таких как Ю.Н. Марчук, И.С. Николаев, О.В. Митренина, Т.М. Ландо, К.К. Боярский, Е.И. Большакова, И.М. Ножов, Е.П. Соснина, И.А. Большаков, Р. Митков, А.В. Луканин, Т.В. Батура, Дж.Ш. Сулейманов, А.Р. Гатиатуллин, Р.А Гильмуллин, А.Б. Альменова и других.

В Узбекистане, теоретические исследования проблем машинной лингвистики, в том числе обработки текстов и речевых сигналов на естественных языках, машинного перевода, формирования электронных словарей тюркских языков, фонетики и морфологии узбекского языка проводились такими учеными как, Ш.А. Назиров, М.М. Арипов, М.М. Мусаев, М.А. Рахматуллаев, Б.Б. Мўминов, М.Х. Хакимов и другими.

Научные труды этих ученых сегодня являются наиболее авторитеными источниками для изучения узбекской компьютерной лингвистики. Хотя ими уже достигнуты значительные результаты теоретического и прикладного характера, тем не менее, вопросы разработки программных систем автоматической обработки текстов на основе моделей и алгоритмов морфологического анализа словоформ узбекского языка, до настоящего времени остаются недостаточно изученными.

Связь диссертационного исследования с планами научно-исследовательских работ высшего образовательного учреждения, где выполнена диссертация. Диссертационное исследование выполнено в соответствии с планами работ Научно-исследовательского института развития цифровых технологий и искусственного интеллекта в рамках проекта:

И-ОТ-2019-12 «Создание и популяризация электронной платформы «Умная библиотека», посвященной современной научно-практической интерпретации и анализу произведений наших предков, внесших значительный вклад в науку исламоведения» (2019-2021).

Целью исследования является разработка моделей, алгоритмов и веб-ориентированного программных средств морфологического анализа словоформ узбекского языка.

Задачи исследования:

Исследование и разработка морфемных моделей словоформ и алгоритмов словообразования для узбекского языка;

разработка алгоритмов определения морфологической основы слов и морфемного анализа словоформ узбекского языка;

формирование лексикона, отражающего грамматические особенности словоформ узбекского языка;

разработка база знаний на основе продукционных правил с использованием конечных автоматов для распознавания токенов в текстах;

разработка функциональной модели процесса морфологического разбора слов узбекского языка;

разработка информационной модели данных и коммуникационной модели для проектирования веб-ориентированного программного средства морфологического анализа словоформ узбекского языка.

Объектом исследования является процесс компьютерной обработки текстов на естественных языках.

Предметом исследования являются модели, алгоритмы и программные средства морфологического анализа словоформ узбекского языка.

Методы исследования. В ходе исследования применялись методы теории множеств, математической статистики, компьютерной лингвистики, а также методы проектирования и технологии разработки баз данных и распределенных веб-приложений.

Научная новизна исследования заключается в следующем:

разработаны морфемные модели словоформ и алгоритмы словообразования с учетом лингвистических особенностей, орфографии и пунктуации узбекского языка;

разработаны алгоритмы определения корневых морфем и морфемного анализа словоформ узбекского языка, образованных методом аффиксации;

разработан алгоритм распознавания лексем на основе использования лексикона основ слов узбекского языка и грамматических тегов;

разработана база знаний для распознавания токенов из текстов на узбекском языке на основе продукционных правил, состоящих из конечных автоматов;

разработаны модели проектирования программной системы морфологического анализа словоформ узбекского языка, включая функциональную IDEF0 и информационную IDEF1x модели, объектно-ориентированную модель классов и коммуникационную модель на базе архитектуры REST.

Практические результаты исследования заключаются в следующем:

разработано веб-ориентированное программное средство для морфологического анализа словоформ узбекского языка, функционирующее как в интерактивном режиме, так и в качестве веб-сервиса для обслуживания запросов сторонних поисковых систем;

разработанное программное средство позволило повысить эффективность обработки текстовых запросов и релевантность выдачи результатов в поисковых системах ряда информационно-библиотечных центров республики за счет улучшения токенизации текстов на узбекском языке с учетом правописания и структуры слов, включая простые непроизводные, простые производные, сложные, парные, повторяющиеся и сложносокращённые слова, а также за счет повышения точности определения корневых морфем и качества морфемного анализа словоформ узбекского языка, образованных методом аффиксации.

Достоверность результатов исследования. Достоверность результатов исследования обосновывается математической полнотой и точностью поставленных задач, правильным применением морфемных моделей при определении сходств и различий лингвистических понятий, согласованностью теоретических данных и практических результатов исследований.

Научная и практическая значимость результатов исследования. Научная значимость результатов исследования объясняется тем, что они вносят определенный вклад в развитие методов компьютерной лингвистики, моделирования словоформ узбекского языка и разработки алгоритмов морфологического, а также методов проектирования и разработки систем автоматической обработки текстов на узбекском языке.

Практическая значимость результатов исследования объясняется применением разработанного веб-ориентированного программного средства «Morphoanalyzer» для улучшения обработки поисковых запросов на узбекском языке и повышения релевантности результатов поиска в электронных каталогах библиографических записей и библиотеках полнотекстовых документов ряда информационно-библиотечных и архивных учреждений республики.

Внедрение результатов исследования. На основе применения моделей, алгоритмов и веб-ориентированного программного средства морфологического анализа словоформ узбекского языка:

модели, алгоритмы и программное обеспечение для морфологического анализа текстовых запросов на узбекском языке при поиске информационных ресурсов в системе электронной библиотеки внедрены в Бухарском областном информационно-библиотечном центре имени Абу Али ибн Сино и в центре информационных ресурсов Бухарского государственного университета (Справка Агентства информации и массовых коммуникаций при Администрации Президента Республики Узбекистан №11-3885 от 22 сентября 2021 года). В результате внедрения обеспечена возможность сокращения времени и трудозатрат на 9-11% при выполнении операций составления библиографических описаний документов на узбекском языке в системе

электронного каталога за счет автоматической коррекции орфографических ошибок и выдачи правильного варианта написания слов.

модели, алгоритмы и программное обеспечение для морфологического анализа словоформ узбекского языка внедрены в Национальной библиотеке Узбекистана (Справка Агентства информации и массовых коммуникаций при Администрации Президента Республики Узбекистан №11-3885 от 22 сентября 2021 года). В результате внедрения обеспечено повышение релевантности поисковой выдачи информационных ресурсов в системах электронного каталога библиографических записей и электронной библиотеки полнотекстовых документов за счет точности морфемного анализа.

Апробация результатов исследования. Результаты данного исследования были обсуждены на 6 международных и 6 республиканских научных конференциях.

Опубликованность результатов исследования. По теме диссертации опубликованы 23 научные работы, из которых 9 статей в научных изданиях, рекомендованных Высшей аттестационной комиссией Республики Узбекистан для публикации основных научных результатов диссертаций, в том числе 3 в зарубежных и 6 в республиканских журналах, а также получены 2 свидетельства об официальной регистрации программы для ЭВМ.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения, списка использованной литературы и приложений. Объем диссертации составляет 116 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во **введении** обоснована актуальность и востребованность темы диссертации, сформулированы цель и задачи исследования, указаны объект и предмет исследования, определено соответствие темы исследования приоритетным направлениям развития науки и технологий Республики Узбекистан, изложены научная и практическая значимость полученных результатов, приведены сведения об опубликованности результатов исследования и их внедрении в деятельности хозяйствующих субъектов, а также сведения о структуре диссертации.

Первая глава диссертации **«Состояние и тенденции развития проблем обработки текстов на естественных языках»** состоит из пяти параграфов. В первом параграфе рассмотрены теоретические аспекты компьютерной лингвистики и эволюция методов обработки текстов на естественных языках. Во втором параграфе приведен критический анализ существующих подходов к морфологическому анализу с точки зрения их достоинств и недостатков. В третьем параграфе представлен сравнительный анализ функциональных возможностей существующих программных средств морфологического анализа. В четвертом параграфе приведен анализ результаты научных исследований зарубежных и отечественных ученых по разработке математического и программного обеспечения систем морфологического анализа текстов на агглютинативных языках. В пятом параграфе на основе

анализа научных работ и выявленных пробелов в области компьютерной обработки текстов на узбекском языке, сформулированы цель и задачи данного диссертационного исследования.

Во второй главе **«Модели и алгоритмы морфологического анализа словоформ узбекского языка»** с опорой на современные научные достижения в области развития методов решения задач токенизации и стемминга, разработаны морфемные модели словоформ и алгоритмы словообразования для узбекского языка, алгоритмы токенизации и определения морфологической основы слов, а также морфемного анализа словоформ узбекского языка с учетом грамматических особенностей.

Токены в текстах на узбекском языке можно условно разделить на две группы. В первую группу включаются токены сложносокращенных слов, а во вторую – токены сложных, парных и парно повторяющихся слов в соответствии с правилами словообразования. В большинстве случаев существующие токенизаторы допускают ошибку при различении токенов, принадлежащих второй группе.

С учетом отмеченной проблемы были предложены математические выражения, описывающие орфографические правила написания сложных, парных, парно повторяющих слов, относящихся ко второй группе:

$$w_{qs} = w'_{qs} + c_{sp} + w''_{qs}; \quad (1)$$

$$w_j = w'_j + \rho + w''_j; \quad (2)$$

$$w_t = w'_t + \rho + w'_t. \quad (3)$$

Здесь w_{qs} – сложное слово; w_j – парное слово, w_t – парно повторяющееся слово, w'_{qs} , w''_{qs} , w'_j , w''_j , w'_t – слова с самостоятельным значением, эти две части слова представляют новое лексическое значение, ρ – дефис, c_{sp} – пробел.

Для распознавания токенов в текстах, согласно по математическим выражениям (1)-(3), был разработан конечный автомат (КА).

Принцип действия КА показан на рис. 1 на примере трех сложных слов типа w_{qs} , представляющих собой указательные местоименные наречия на узбекском языке: «u yerga» (туда), «bu yerda» (здесь), «shu yerdan» (отсюда).

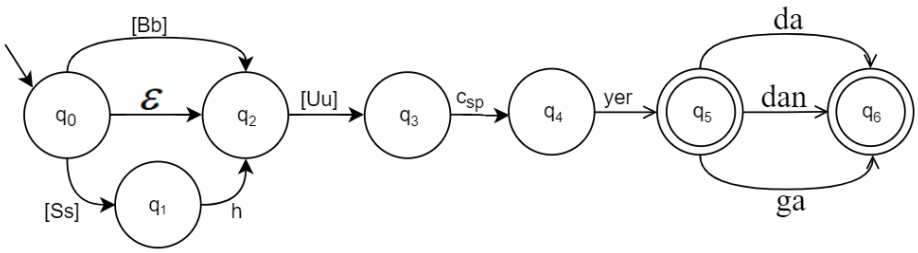


Рис. 1. w_{qs} **– конечный автомат для трех сложных слов типа** w_{qs}

КА, приведенный на рис. 1, представлен следующим регулярным выражением:

$$((\text{sh})\,|\,\text{b})^{*}u)\setminus s\ (\text{dan}\,|\,\text{da}\,|\,\text{ga})?\,.$$

Для математических выражений парных, парно повторяющихся и сложносокращенных слов конечные автоматы строятся в том же порядке и преобразуются в регулярные выражения. Таким образом, алгоритм токенизации строится с использованием регулярных выражений на основе каждый математических выражений. На примере сложных слов (рис. 1), алгоритм состоит из следующих шагов:

Шаг 1. Запуск работы алгоритма.

Шаг 2. Исходный текст (с клавиатуры, из буфера обмена или из файла) присваивается в качестве значения строковой переменной «text».

Шаг 3. Существующее сложное слово в тексте формируется из словаря «ng», состоящего из токенов биграмм и их маркеров. Токены биграмм сложного слова извлекаются из базы знаний.

Шаг 4. Если словарь токенов биграмм сложного слова не равен значению null, то переход к шагу 5, в противном случае – к шагу 6.

Шаг 5. Существующее составное слово в строковой переменной «text» помечается маркерами биграмм.

Шаг 6. Текст в переменной «text» преобразуется в список, используя символ пробела в качестве разделителя. Если словарь «ng» не пуст, то маркеры в списке заменяются сложным словом.

Шаг 7. На экране отображается значение переменной «x» в типе списка.

Шаг 8. Алгоритм завершает свою работу.

Во втором параграфе данной главы разработаны модели и алгоритмы определения морфологической основы слов. Следует отметить, что структура слов в узбекском языке играет важную роль в определении их производящей основы. В этой связи, предложены морфемные модели словоформ, образованных методом аффиксации на основе лингвистических правил, описываемых нижеследующими математическими выражениями.

Для простых производных слов

$$w_{sy} = K + \left\{\sum_{l=1}^{m} sx_l\right\} + \left\{\sum_{i=1}^{n} af_i\right\}. \qquad (5)$$

Здесь K – простое непроизводное слово (корень); sx – суффикс (слово-образовательный аффикс); af – аффиксы; Σ – операция последовательной конкатенации; l – индекс количества суффиксов; i – индекс количества аффиксов.

Чтобы доказать свою точку зрения, проанализируем словоформу «gul-don-lar-imiz-ni» в категории существительных. В этом примере K = gul, sx = don, $af = af_1 \cdot af_2 \cdot af_3 = lar \cdot imiz \cdot ni = larimizni$. В итоге словорма w_{sy} принимает значение «guldonlarimizni».

В случае простых производных слов с префиксами

$$w_{psy} = \{\text{Pr}\} + K + \left\{\sum_{l=1}^{m} sx_l\right\} + \left\{\sum_{i=1}^{n} af_i\right\}. \qquad (6)$$

Здесь Pr – префикс (словообразовательный аффикс), остальные обозначения те же.

Проанализируем словоформу «be-adab-garchilik-lar-ni». В этом примере Pr = be; K = adab; sx = garchilik; $af = af_1 \cdot af_2 = lar \cdot ni = larni$. В итоге словорма w_{psy} принимает значение «beadabgarchiliklarni».

Для простых непроизводных слов

$$w_{st} = K + \left\{ \sum_{i=1}^{n} af_i \right\}. \quad (7)$$

Проиллюстрируем (7) на примере словоформы «kitob-lar-imiz-ning». В данном примере K = kitob; $af = af_1 \cdot af_2 \cdot af_3 = lar \cdot imiz \cdot ning = larimizning$.

Как можно видеть из (5)-(7), в простых непроизводных словах w_{st} аффиксы добавляются справа, а в простых производных словах w_{psy} и w_{sy} аффиксы добавляются как слева, так и справа. Корень таких слов находится путем отсечения аффиксов слева и справа.

Непосредственно, алгоритм определения морфологической основы слов состоит из следующих шагов:

Шаг 1. Запуск работы алгоритма.

Шаг 2. Очередное слово из исходного текста присваивается в качестве значения строковой переменной «S».

Шаг 3. Значение переменной «S» присваивается строковой переменной с именем «SS», т.е. создается копия слова.

Шаг 4. Поиск значения «SS» осуществляется базе данных с помощью булевой функции dbStemShow(SS). Если копия слова существует в базе данных, то функция возвращает значение «true» (истина), в противном случае – «false» (ложь). Возвращаемое значение функции dbStemShow(SS) затем присваивается переменной логического типа «b».

Шаг 5. Если значение переменной «b» истинно, то переходим к шагу 16.

Шаг 6. Выполняется поиск префикса слова посредством функции preffunc(SS). Если слово имеет префикс, возвращаемое значение функции preffunc(SS) присваивается строковой переменной с именем «pref». В противном случае переменной pref присваивается значение «null».

Шаг 7. Если значение переменной «pref» не пустое, т.е. если префикс присутствует, переходим к шагу 13, в противном случае – к шагу 8.

Шаг 8. По длине слова i: = (0, Length(S)-1) начинаем одноступенчатую итерацию.

Шаг 9. На каждой итерации часть строки длиной i удаляется в обратном порядке, при этом выполняется поиск выделенной подстроки в базе данных с помощью функции dbStemShow(S.SubStr(0,Length(S)-i)), а возвращаемое значение присваивается переменной «b».

Шаг 10. На каждой итерации проверяется логическое значение функции dbStemShow(S.SubStr(0,(Length(S)-1)-i)). Если значение переменной «b» истинно, т.е. подстрока найдена в базе данных, то переходим к шагу 12, иначе – к шагу 8.

Шаг 11. Значение функции S.SubStr(0,Length(S)-i) присваивается переменной «SS» и выполняется переход к шагу 16.

Шаг 12. Если в ходе итерационного процесса не найдено никакой основы, то значение переменной «S» приравнивается переменной «SS» и выполняется переход к шагу 16.

Шаг 13. Часть слова после префикса отсекается и выполняется поиск в базе данных с помощью функции dbStemShow(S.SubStr(Length(pref), Length(S)-1)), а возвращаемое значение присваивается переменной «b».

Шаг 14. Если значение переменной «b» истинно, то есть искомая часть слова имеется в базе данных, переходим к шагу 15, в противном случае – к шагу 8.

Шаг 15. Часть слова после префикса обрезается S.SubStr(Length(pref), Length(S)) и присваивается переменной «SS».

Шаг 16. На экране отображается значение переменной «SS».

Шаг 17. Завершение работы алгоритма.

Аналогичным образом в данном параграфе второй главы рассмотрены детали разработки моделей и алгоритмов для определения морфологической основы парных, парно повторяющихся и сложносокращенных слов.

В третьем параграфе второй главы разработан алгоритм морфемного анализа словоформ. В процессе анализа отдельных слов на предмет их морфемной структуры, слова делятся на морфемы и определяются их значимые части. Поэтому, алгоритм морфемного анализа разработан с использованием морфемных моделей (5)-(7), описывающих структуру словообразования, а также с использованием КА на основе правил морфотактики.

В данном случае, КА определяется следующим образом:

$$M = (\theta, Q, q_0, F, \delta),$$

где $\theta = \{PR, K, SX, AF\}$ – входящий алфавит, $Q = \{q_0, q_1, q_2, q_3, q_4\}$ – множества обстоятельств, $\delta(Q,\theta): Q \times \theta \to Q'$ – функция перехода, q_0 , $(q_0 \in Q)$ – начальное состояние, F , $(F \subseteq Q)$ – множества результирующих обстоятельств.

Для определения простой непроизводной словоформы конечным автоматом, функция перехода должна принимать следующие значения:

$$\delta(q_0, K) = q_2, \delta(q_2, \text{AF}) = q_4, \delta(q_4, AF) = q_4.$$

Для определения простой производной (с суффиксом) словоформы конечным автоматом, функция перехода должна принимать следующие значения:

$$\delta(q_0, K) = q_2, \delta(q_2, SX) = q_3, \delta(q_3, \text{SX}) = q_3, \delta(q_3, AF) = q_4.$$

Для определения простой производной (с префиксом) словоформы конечным автоматом, функция перехода должна принимать следующие значения:

$$\delta(q_0, PR) = q_1, \delta(q_1, K) = q_2, \delta(q_2, SX) = q_3, \delta(q_3, \text{AF}) = q_4,$$
$$\delta(q_3, \text{SX}) = q_3, \delta(q_3, AF) = q_4, \delta(q_4, AF) = q_4.$$

В четвертом параграфе второй главы разработан алгоритм определения части речи слов и их грамматических особенностей. Данный алгоритм предусматривает выполнение следующие шагов: разбиение слова на морфемы; разработка лексикона тегов, представляющих части речи; разработка лексикона тегов, представляющих грамматические особенности для каждой части речи. Построение алгоритма определения грамматических свойств словоформ основано на использовании конечного преобразователя формы (КПФ).

Конечный преобразователь формы $M_{чша}$ задается в следующем виде:

$$M_{чша} = (Q, \theta_{in}, \theta_{out}, \delta, \omega, q_0, \mathrm{F}),$$

где Q – множества обстоятельств, $\theta_{in} = \{\mathrm{A}, \mathrm{l}_1, s_1, s_2\}$ – входящий алфавит, $\theta_{out} = \{N, "Pl", "3pSg", "CA"\}$ – выходной алфавит, $\delta(Q, \theta_{in}) : Q \times \theta_{in} \to Q'$ – функция перехода, $\omega(Q, \theta_{in}) : Q \times \theta_{in} \to \theta_{out}$ – функция вывода, $\mathrm{q}_0 \in \mathrm{Q}$ – начальное состояние КПФ, $\mathrm{F} \subseteq \mathrm{Q}$ – множества результирующих обстоятельств.

Во входящем алфавите приняты следующие обозначения элементов: A – основа; l_1 – аффикс множественного числа, s_1 – аффикс принадлежности, s_2 – падежный аффикс.

Например, если мы возьмем словоформу «kitob-lar-i-ni», относящуюся к категории существительных, в качестве входящего алфавита, то эта словоформа описывается в виде следующего конечного преобразователя формы:

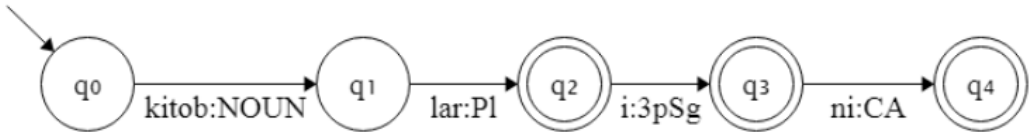


Рис 2. КПФ для категории существительных

КПФ, показанный на рис. 2, имеет результирующие варианты q_2, q_3, q_4, с помощью которых можно назначать грамматические теги для таких словоформ, как «kitoblar» (книги), «kitoblari» (его книги), «kitoblarini» (его книги). Процесс грамматической разметки слова КПФ описан в таб. 1.

Таблица 1. График перехода для КПФ

№	kitob	lar	i	ni
q_0	(q_1, kitob)	-	-	-
q_1	-	(q_2,Pl)	-	-
q_2	-	-	(q_3,3pSg)	-
q_3	-	-	-	(q_4,ni)
q_4	-	-	-	-

Третья глава диссертации **«Программное обеспечение для морфологического анализа слов узбекского языка»** посвящена описанию архитектуры программного обеспечения «Morphoanalyzer» и созданию функциональной модели IDEF0, объектно-ориентированной модели классов, модели базы данных IDEF1x и базы знаний на основе продукционных правил.

Программное обеспечение «Morphoanalyzer» основано на трехуровневой архитектуре, включающей клиентский уровень, уровень бизнес-логики и уровень данных.

Логический уровень принимает слова и тексты как входящую информацию и возвращает списки слов и типов слов, списки морфологических основ и аффиксов, списки слов и их грамматических свойств.

Логический уровень программного обеспечения «Morphoanalyzer» проиллюстрирован на рис. 3 в виде модели IDFE0.

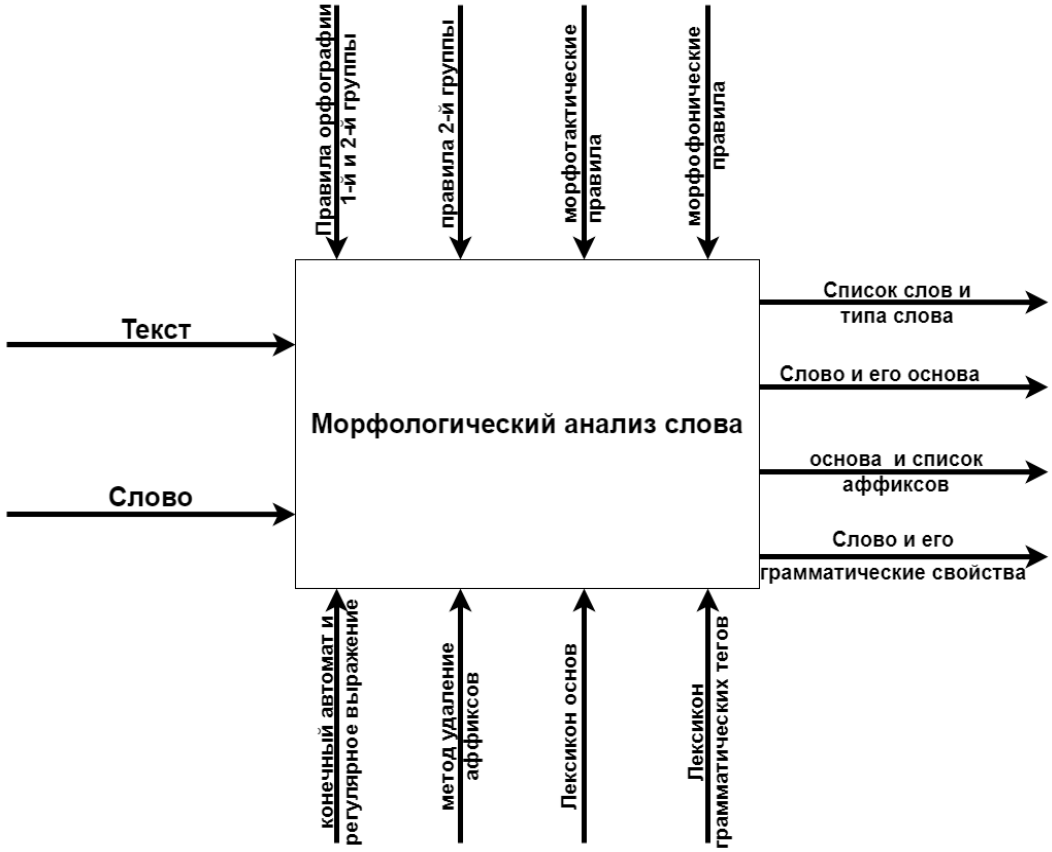


Рис. 3. IDFE0 модель программного обеспечения «Morphoanalyzer»

Модель логического уровня IDFE0 состоит из четырех функциональных модулей: модуль разбиения текста на токены; модуль нормализации слов; модуль морфемного анализа слов; модуль поиска грамматических особенностей слов;

Функции этих модулей выполняются методами, реализованных в следующих классах: «Token» – функционирует как модуль распределения токенов; «Stem» – действует как модуль нормализации слов;

«PostTagMorphem» – выполняет функции модуля морфемного анализа слов и поиска грамматических особенностей.

Программное обеспечение «Morphoanalyzer» позволяет пользователю обмениваться информацией в компьютерной сети, выступая в роли веб-сервиса, реализованного согласно архитектурному стилю REST API (рис. 4).

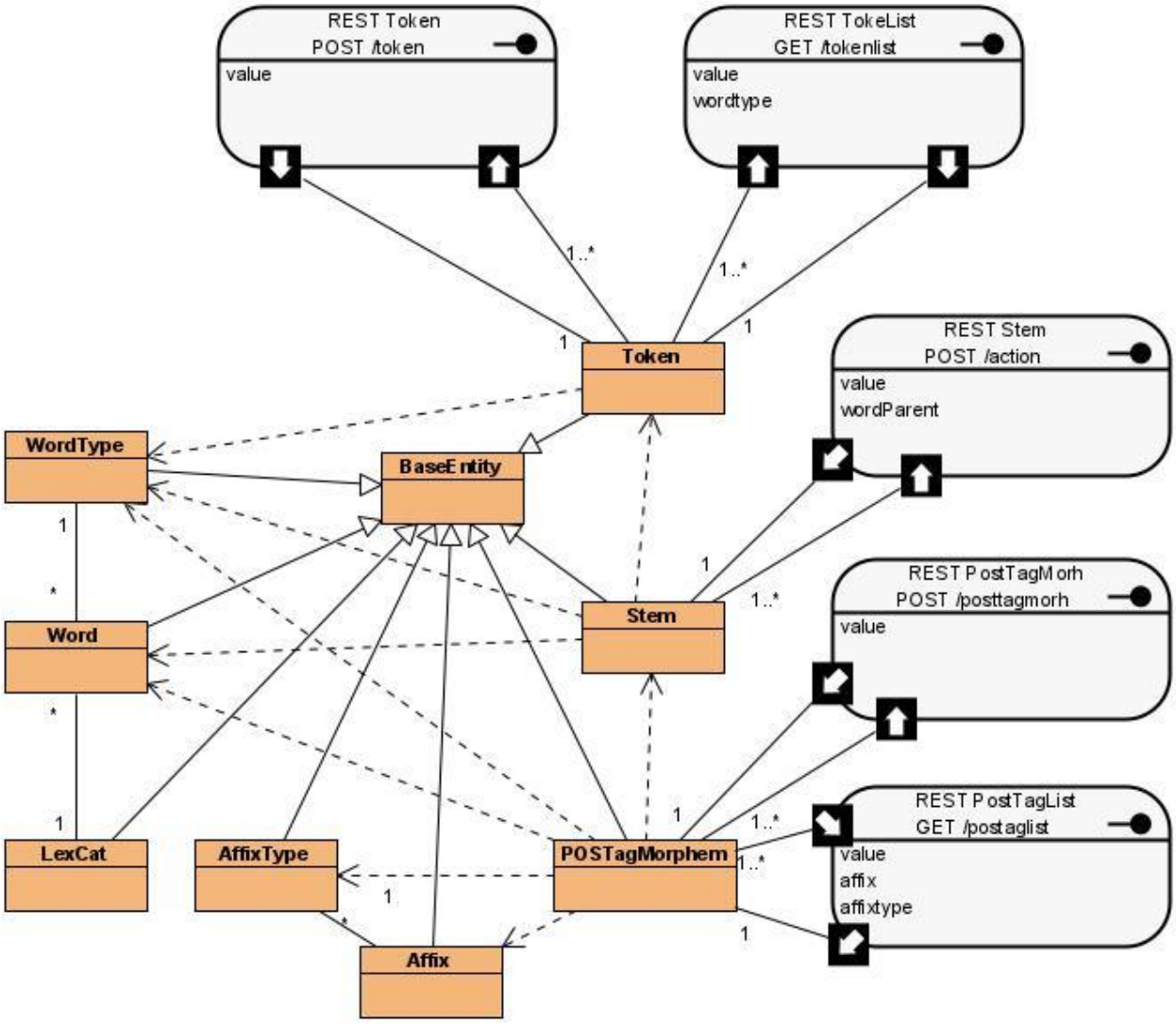


Рис. 4. Коммуникационная модель взаимодействия компонентов «Morphoanalyzer» в сетевой среде

Основные программные интерфейсы (API) «Morphoanalyzer» приведены в таб. 2.

Таблица 2. REST API для доступа к функциям «Morphoanalyzer»

Имя API	Адрес URL	Метод
TokenAPI	/token/text	GET
TokenList	/tokenlist/text	GET
Stem	/stem/value	GET
PostTag	/posttag/value	GET

PostTagList	/posttaglist/value	GET

Например, когда мы отправляем слово «kitobxon» в качестве запроса, результат возвращается в формате JSON следующим образом:

{"value": "kitobxon", "affix": "xon", "affixtype": "so'z yasovchi suffuks"}.

Каждый модуль программного обеспечения «Morphoanalyzer» использует базы знаний на основе продукционных правил для выполнения своих функций. Например, алгоритм токенизации разделяет сложные слова следующим образом:

$$i =< Q; \mathrm{P}; \mathrm{A} \rightarrow \mathrm{B}; \mathrm{N} > .$$

Здесь, i – токенизация (наименование продукции), Q – список биграмм токенов, сгенерированных из текста (область применения), P– первоначальное условие проверки наличия множества правил в базе знаний с именем, соответствующим комбинации 0-х индексов каждого элемента биграмы, A – условие соответствия регулярному выражению (правилу), B – действие маркировки слов, N – действие добавления сложного слова в словарь.

Например, «har yerda»→«hy». Если сушествует множество правил с именем «hy», тогда условие активации продукции выполняется, если условие истинно, то выполняется маркировка и добавление слова «har yerda» в словарь.

В приведенном примере есть три правила в наборе под названием «hy», и эти три правила применяются к введенному слову последовательно одно за другим. Цикл останавливается, когда все правила применяются к введенному слову или, когда какое-либо правило возвращает истинное значение (таб. 3).

Таблица 3. Принцип работы базы знаний по сложным словам

Максимальное количество итераций	Рабочая память (Вставленное слово)	множество	Конфликтные правила множества	Применимое правило
3	har yerda	hy	Q16, Q18, Q33	16

Четвертая глава диссертации **«Применение программного обеспечения Morphoanalyzer»** содержит сведения об аппаратных требованиях программного обеспечения для морфологического анализа, описание его функциональных возможностей, инструкции по использованию программного обеспечения, а также результаты экспериментов и их интерпретацию.

В рамках программного обеспечения «Morphoanalyzer» предусмотрены два типа ролей пользователей: лингвист-эксперт (администратор) и пользователь. Программное обеспечение «Morphoanalyzer», в том числе, позволяет пользователям выполнять ряд действий, называемых прецедентами.

Каждый прецедент для морфологического анализа имеет свою спецификацию (таб. 4).

Таблица 4. Спецификация прецедента текстового токена

Имя прецедента – токенизация текста
Краткое описание – отделяет токены от текста и определяет их тип
Класс – выполняет задачу «Token»
Главный пользователь – Пользователь
Предварительные требования – доступность дополнительного программного обеспечения для браузера
Этапы исполнения прецедента: 1. Отделяются токены из текста. 2. Определяется тип токенов.
Окончательные условия – нет

С помощью программного обеспечения «Morphoanalyzer» проведен ряд экспериментов по анализу текстов различных жанров, в том числе, сказки, юридические документы, хадисы и узбекские народные пословицы. Результаты работы разработанных алгоритмов приведены в таб. 5-7.

Таблица 5. Результаты алгоритма токенизации

№	Выбранные корпуса	Количество слов	Типы слов		
			Сложные	Парные и парно повторяющиеся	Сложно-сокращённые
1	Сказки	2100	147(7%)	84(4%)	0
2	Законные документы	2150	194(9%)	43(2%)	1769
3	Хадисы	2000	110(5.5%)	30(1.5%)	0

Таблица 6. Результаты алгоритма стемминга

№	Выбранные корпуса	Количество слов	Правильные основы
1	Сказки	1246	96%
2	Законные документы	1290	92.4%
3	Хадисы	1285	88.9%
4	Узбекские народные пословицы	1200	94%

ЗАКЛЮЧЕНИЕ

Основные результаты, полученные в рамках диссертационной работы «Модели и алгоритмы морфологического анализа словоформ узбекского языка», сводятся к следующему:

1. Исследованы основные характеристики процесса компьютерной обработки текстов на естественных языках и проведён сравнительный анализ существующих подходов к разработке моделей, алгоритмов и программ для решения задач морфологического анализа;
2. Проведен критический анализ функциональных возможностей программ для морфологического анализа агглютинативных языков;
3. Разработаны конечные автоматы для распознавания парных, парно повторяющихся, сложных и сложносокращённых слов, а также алгоритмы словообразования с учетом лингвистических особенностей, орфографии и пунктуации узбекского языка;
4. Разработаны алгоритмы определения корневых морфем и морфемного анализа словоформ узбекского языка, образованных методом аффиксации;
5. Разработан алгоритм распознавания лексем на основе использования лексикона основ слов узбекского языка и грамматических тегов;
6. Разработана база знаний на основе продукционных правил, состоящих из множества конечных автоматов для распознавания токенов в текстах на узбекском языке;
7. Разработаны модели проектирования программной системы морфологического анализа словоформ узбекского языка, включая функциональную IDEF0 и информационную IDEF1x модели, объектно-ориентированную модель классов и коммуникационную модель на базе архитектуры REST.
8. Разработано веб-ориентированное программное средство «Morphoanalyzer» для морфологического анализа словоформ узбекского языка, функционирующее как в интерактивном режиме, так и в качестве веб-сервиса для обслуживания запросов сторонних поисковых систем.
9. Разработанное программное средство «Morphoanalyzer» позволило повысить на 9-11% скорость выполнения операций составления библиографических описаний документов на узбекском языке в системах электронного каталога, а также повысить эффективность обработки текстовых запросов и релевантность выдачи результатов в поисковых системах ряда информационно-библиотечных центров республики за счет улучшения токенизации текстов на узбекском языке с учетом правописания и структуры слов и за счет повышения качества морфемного анализа словоформ узбекского языка.

SCIENTIFIC COUNCIL AWARDING SCIENTIFIC DEGREES DSc.13/30.12.2019.T.07.01 AT TASHKENT UNIVERSITY OF INFORMATION TECHNOLOGIES

TASHKENT UNIVERSITY OF INFORMATION TECHNOLOGIES

BAKAEV ILKHOM IZATOVICH

MODELS AND ALGORITHMS OF MORPHOLOGICAL ANALYSIS OF WORD FORMS OF THE UZBEK LANGUAGE

05.01.04 – Mathematical and software support of computers, complexes and computer networks

DISSERTATION ABSTRACT OF THE DOCTOR OF PHILOSOPHY (PhD) ON TECHNICAL SCIENCES

Tashkent – 2021

The theme of dissertation of doctor of philosophy (PhD) on technical sciences was registered at the Supreme Attestation Commission at the Cabinet of Ministers of the Republic of Uzbekistan under number B2021.3.PhD/T2374.

The dissertation has been prepared at Tashkent University of Information Technologies Named After Mukhammad al-Kharizmi.

The abstract of the dissertation is posted in three languages (Uzbek, Russian, English (resume)) on the website www.tuit.uz and on the website of «ZiyoNet» Information and Educational portal www.ziyonet.uz.

Scientific adviser:	**Ravshanov Normakhmad** Doctor of Technical Sciences, Professor
Official opponents:	**Mukhamedieva Dilnoz Tulkunovna** Doctor of Technical Sciences, Professor
	Norov Abdisait Muradovich Doctor of Philosophy on Technical Sciences
Leading organization:	**National University of Uzbekistan**

The defense of dissertation will take place on «28» december 2021 at 14^{00} at the meeting of Scientific Council No. DSc.13/30.12.2019.T.07.01 at Tashkent University of Information Technologies (Address: 100200, Tashkent, Amir Temur str., 108. Tel.: (+99871) 238-64-43, fax: (+99871) 238-65-52, e-mail: tuit@tuit.uz).

The dissertation can be reviewed at the Information Resourse Centre of Tashkent University of Information Technologies (is registered under No. 236). (Address: 100202, Tashkent, Amir Temur str., 108. Tel.: (+99871) 238-64-43, fax: (+99871) 238-65-52).

Abstract of the dissertation was sent out on «16» december 2021 y.
(mailing report № 41 of «6» december 2021 y.).

R.Kh. Khamdamov
Chairman of the Scientific Council
Awarding Scientific Degrees,
Doctor of Technical Sciences, Professor

F.M. Nuraliev
Scientific secretary of the Scientific Council
Awarding Scientific Degrees,
Doctor of Technical Sciences, Docent

M.A. Rakhmatullaev
Chairman of the Scientific Seminar Under
the Scientific Council Awarding Scientific Degrees,
Doctor of Technical Sciences, Professor

INTRODUCTION (abstract of the PhD thesis)

The aim of the research work is to develop models, algorithms and web-based software for morphological analysis of word forms of the Uzbek language.

The object of research work is the process of computer processing of texts in natural languages.

The scientific novelty of the research work is as follows:

-developed morphemic models of word forms and word formation algorithms, taking into account linguistic features, spelling and punctuation of the Uzbek language;

algorithms for determining root morphemes and morphemic analysis of word forms of the Uzbek language, formed by the affixation method, have been developed;

developed an algorithm for recognizing lexemes based on the use of the lexicon of the foundations of the words of the Uzbek language and grammatical tags;

developed a knowledge base based on production rules, consisting of a set of finite state machines for recognizing tokens in texts in the Uzbek language;

design models of a software system for morphological analysis of word forms of the Uzbek language have been developed, including functional IDEF0 and information IDEF1x models, an object-oriented class model and a communication model based on the REST architecture.

Implementation of the research results. Based on the use of models, algorithms and a web-based software tool for morphological analysis of word forms of the Uzbek language:

models, algorithms and software for morphological analysis of text queries in the Uzbek language when searching for information resources in the electronic library system have been introduced in the Bukhara Regional Information and Library Center named after Abu Ali ibn Sino and in the center of information resources of Bukhara State University (Information from the Agency of Information and Mass Communications under the Administration of the President of the Republic of Uzbekistan No. 11-3885 dated September 22, 2021). As a result of the implementation, it is possible to reduce time and labor costs by 9-11% when performing operations of compiling bibliographic descriptions of documents in the Uzbek language in the electronic catalog system by automatically correcting spelling errors and issuing the correct spelling of words.

models, algorithms and software for morphological analysis of word forms of the Uzbek language have been introduced in the National Library of Uzbekistan (Reference from the Agency of Information and Mass Communications under the Administration of the President of the Republic of Uzbekistan No. 11-3885 dated September 22, 2021). As a result of the implementation, an increase in the relevance of the search results of information resources in the systems of the electronic catalog of bibliographic records and the electronic library of full-text documents was ensured due to the accuracy of morphemic analysis.

Structure and volume of the dissertation. The dissertation consists of an introduction, four chapters, a conclusion, a bibliography and annexes. The volume of the thesis is 116 pages.

ЭЪЛОН ҚИЛИНГАН ИШЛАР РЎЙҲАТИ
СПИСОК ОПУБЛИКОВАННЫХ РАБОТ
LIST OF PUBLISHED WORKS

I бўлим (I часть; I part)

1. Равшанов Н., Бакаев И. Разработка объектно-ориентированной модели базы данных «умной библиотеки» // Проблемы вычислительной и прикладной математики. — 2019. — № 5(23). — С. 117–129.
2. Bakaev I.I., Shafiev T.R. Morphemic analysis of Uzbek nouns with Finite State Techniques// Journal of Physics: Conference Series. London, 2020. Vol. 1546. Issue 012076(№3, Scopus; IF=0.54).
3. Bakaev I.I., Bakaeva R.I. Creation of a morphological analyzer based on finite-state techniques for the Uzbek language// Journal of Physics: Conference Series. London, 2020. Vol. 1791. Issue 012068 (№3, Scopus; IF=0.54).
4. Bakaev I.I. Linguistic features tokenization of text corpora of the Uzbek language//// Bulletin of TUIT: Management and Communication Technologies:2021. №3. Vol.4. Issue 3. (Олий аттестация комиссияси Раёсати қарори №283/7.1-сон, 30.07.2020 й.).
5. Бакаев И.И., Шафиев Т.Р., Организация полнотекстового поиска на веб-ресурсах// Проблемы вычислительной и прикладной математики. — 2020. – №1(25). — С. 118–127. (05.00.00; №23)
6. Бакаев И.И., Шафиев Т.Р. Методы построения алгоритмов стемминга для естественных языков// Проблемы вычислительной и прикладной математики. — 2020. — №3(27). — С.146–153. (05.00.00; №23)
7. Равшанов Н., Бакаев И.И., Шафиев Т.Р. Особенности разработки морфологического анализатора узбекского языка// Проблемы вычислительной и прикладной математики. –2020.–№4(28).–С.121-131. (05.00.00; №23)
8. Бакаев И.И. Разработка алгоритма стемминга для слов узбекского языка// Кибернетика и программирование. Москва. 2021.–№1. – С.1-12. (05.00.00; №45)
9. Бакаев И.И. Stemming algorithm for Uzbek words// Муҳаммад ал-Хоразмий авлодлари. 2021. № 3(17). –С. 147–152. (05.00.00; №10).

II бўлим (II часть; II part)

10. Бакаев И.И. Технология машинного обучения для поддержки принятия решений//Ахборот-коммуникация технологиялари ва телекоммуникацияларнинг замонавий муаммолари ва ечимлари республика илмий-техник анжуман. Фарғона. 2019. С. 86–88.
11. Bakaev I.I. Development of a stemming algorithm based on a linguistic approach for words of the uzbek language// International Conference on Scientific, Educational & Humanitarian Advancements. Turkey July 15th, 2021. ICSEHA-2021, pp 195-202.

12. Бакаев И.И., Бакаева Р.И. Создание морфологического анализатора для узбекского языка // Актуальные вызовы современной науки: LXIV Международная научная конференция. – Переяслав, 2021. – С. 60–63.

13. Бакаев И.И. Подходы токенизации текстовых корпусов узбекского языка// Компьютер лингвистикаси: муаммо ва ечимлар мавзусидаги халқаро онлайн илмий-амалий конференция материаллари тўплами. Тошкент, 2021. – С. 92-98.

14. Бакаев И.И., Бакаева Р.И. Токенизация текстовых корпусов узбекского языка // Амалий математика ва ахборот технологияларининг замонавий муаммолари халқаро миқёсидаги илмий-амалий анжуман материаллари. Бухоро, 2021. – С. 388-393.

15. Бакаев И.И., Шарипова М.М., Бакаева Р. И. Матнларни компьютерли морфологик таҳлил қилиш усуллари// «Ишлаб чиқаришга инновацион технологияларни жорий этиш ва қайта тикланадиган энергия манбаларидан фойдаланиш муаммолари» мавзусидаги Республика миқиёсидаги илмий-техник анжуманининг материаллари тўплами. Жиззах. 2020. – С. 203-206.

16. Бакаев И.И., Бакаева Р.И. Морфологический анализ слов на основе конечного преобразователя для узбекского языка// Инновацион ёндашувлар илм-фан тараққиёти калити сифатида: ечимлар ва истиқболлар мавзусидаги Республика миқёсидаги илмий-техник анжумани материаллари тўплами. Жиззах. 2020. – С. 34-38.

17. Бакаев И.И., Шафиев Т.Р. Применение теория конечных автоматов для морфемного разбор имен существительных в узбекском языке// Ахборот коммуникация технологиялари ва дастурий таъминот яратиш мавзусида профессор-ўқитувчилар ва талабаларнинг XV-илмий-амалий конференцияси. Самарқанд.2020. – С. 106-109.

18. Равшанов Н., Бакаев И.И., Шафиев Т.Р. Создание алгоритма для морфотаксических правил узбекского языка// Инновацион замонавий ахборот технологияларини таълим, фан ва бошқарув соҳаларида қўллаш истиқболлари халқаро илмий-амалий конференция. Самарқанд.2020. – С. 394-398.

19. Бакаев И.И., Шарипова М.М., Бакаева Р. И. Матнларни компьютерли морфологик таҳлил қилиш усуллари// Математик моделлаштириш, ҳисоблаш математикаси ва дастурий таъминот инженериясининг долзарб муаммолари мавзусида Республика миқёсидаги илмий анжуман материаллари тўплами. Қарши.2020. – С. 286-290.

20. Бакаев И.И., Шафиев Т.Р. Использование алгоритмов автозаполнение для формирования запросов пользователей// Инновацион замонавий ахборот технологияларини таълим, фан ва бошқарув соҳаларида қўллаш истиқболлари халқаро илмий-амалий конференция. Самарқанд.2020. – С. 399-402.

21. Бакаев И.И., Бакаева Р.И. Морфологический анализ слов на основе конечного преобразователя для узбекского языка//Математик моделлаштириш, ҳисоблаш математикаси ва дастурий таъминот

инженериясининг долзарб муаммолари мавзусида Республика миқёсидаги илмий анжуман материаллари тўплами. Қарши.2020. – С. 110-112.

22. Равшанов Н., Бакаев И.И. Программа для ЭВМ «Search by word index» // Агентство по интеллектуальной собственности РУз. Свидетельство № DGU 07446. 28.11.2019.

23. Равшанов Н., Бакаев И.И., Шафиев Т.Р. Программа для ЭВМ «Morphemic parser» // Агентство по интеллектуальной собственности РУз. Свидетельство № DGU 09170. 19.09.2020.

Автореферат «Информатика ва энергетика муаммолари Ўзбекистон журнали» таҳририятида таҳрирдан ўтқазилди ва ўзбек, рус тилларидаги матнларини мослиги текширилди.