**БУХОРО ДАВЛАТ УНИВЕРСИТЕТИ
ҲУЗУРИДАГИ ИЛМИЙ ДАРАЖАЛАР БЕРУВЧИ
DSc.03/04.06.2021.Fil.72.09 РАҚАМЛИ ИЛМИЙ КЕНГАШ**

**БУХОРО ДАВЛАТ УНИВЕРСИТЕТИ**

**ТОИРОВА ГУЛИ ИБРАГИМОВНА**

# ЎЗБЕК ТИЛИ МИЛЛИЙ КОРПУСИНИ ЯРАТИШНИНГ НАЗАРИЙ ВА АМАЛИЙ МАСАЛАЛАРИ

**10.00.01 – Ўзбек тили**

**Филология фанлари доктори (DSc) диссертацияси
АВТОРЕФЕРАТИ**

**Бухоро – 2021**

**УДК: 81`1:004=512.133**
**81'322.2**

**Докторлик (DSc) диссертацияси автореферати мундарижаси**
**Оглавление автореферата докторской (DSc) диссертации**
**Contents of the abstract of doctoral (DSc) dissertation**

**БУХОРО ДАВЛАТ УНИВЕРСИТЕТИ
ҲУЗУРИДАГИ ИЛМИЙ ДАРАЖАЛАР БЕРУВЧИ
DSc.03/04.06.2021.Fil.72.09 РАҚАМЛИ ИЛМИЙ КЕНГАШ**

**БУХОРО ДАВЛАТ УНИВЕРСИТЕТИ**

**ТОИРОВА ГУЛИ ИБРАГИМОВНА**

# ЎЗБЕК ТИЛИ МИЛЛИЙ КОРПУСИНИ ЯРАТИШНИНГ НАЗАРИЙ ВА АМАЛИЙ МАСАЛАЛАРИ

**10.00.01 – Ўзбек тили**

**Филология фанлари доктори (DSc) диссертацияси
АВТОРЕФЕРАТИ**

**Бухоро – 2021**

**Фан доктори (DSc) диссертацияси мавзуси Ўзбекистон Республикаси Вазирлар Маҳкамаси ҳузуридаги Олий аттестация комиссиясида B2019.3.DSc/Fil180 рақам билан рўйхатга олинган.**

Диссертация Бухоро давлат университетида бажарилган.

Диссертация автореферати уч тилда (ўзбек, рус, инглиз (резюме)) Бухоро давлат университети веб-сайти (www.fdu.uz) ҳамда «ZiyoNet» Ахборот-таълим порталида (www.ziyonet.uz) жойлаштирилган.

| | |
|---|---|
| **Илмий раҳбар:** | **Менглиев Бахтиёр Ражабович**<br>филология фанлари доктори, профессор |
| **Расмий оппонентлар:** | **Муҳаммедова Саодат Худойбердиевна**<br>филология фанлари доктори, профессор |
| | **Каримов Суюн Амирович**<br>филология фанлари доктори, профессор |
| | **Ҳакимов Муҳаммадхон Хўжахонович**<br>филология фанлари доктори, профессор |
| **Етакчи ташкилот:** | **Термиз давлат университети** |

Диссертация ҳимояси Бухоро давлат университети ҳузуридаги DSc.03/04.06.2021.Fil.72.09 рақамли Илмий кенгашнинг 2021 йил «____»_____ соат _____даги мажлисида бўлиб ўтади. (Манзил: 200118, Бухоро шаҳри, М.Иқбол кўчаси, 11-уй. Тел.: +99865221-29-14; факс: +99865221-27-07, e-mail: buxdu_rektor@buxdu.uz).

Диссертация билан Бухоро давлат университети Ахборот-ресурс марказида танишиш мумкин (____ рақами билан рўйхатга олинган).(Манзил: 200118, Бухоро шаҳри, М.Иқбол кўчаси, 11-уй. Тел.: +99865221-29-14).

Диссертация автореферати 2021 йил «___»_____куни тарқатилди.

(2021 йил «_____»_____ даги _____ -рақамли реестр баённомаси).

**Ражабов Д.З.**
Илмий даражалар берувчи илмий кенгаш раиси, филология фанлари доктори (DSc), профессор

**Эшонқулов Ҳ.П.**
Илмий даражалар берувчи илмий кенгаш илмий котиби, филология фанлари доктори (DSc), доцент

**Муродов Ғ.Н.**
Илмий даражалар берувчи илмий кенгаш қошидаги илмий семинар раиси, филология фанлари доктори ( DSc), доцент

## КИРИШ (докторлик (DSc) диссертацияси аннотацияси)

**Диссертация мавзусининг долзарблиги ва зарурати.** Жаҳон тилшунослигида XXI асрнинг иккинчи ўн йиллигига келиб муайян тилни сақлаб қолишнинг, унинг ўрганилиш доирасини кенгайтиришнинг, тил имкониятларини намоён этишнинг асосий воситаси интернет тизимида тил корпусларини ўрганишга алоҳида эътибор қаратилмоқда. Жумладан, бугунги кунда XX асрнинг буюк кашфиёти бўлган компьютер технологиялари бошқа соҳалар каби тилшунослик фани учун ҳам кенг имкониятлар эшигини очиши, ҳам унинг зиммасига компьютер тили билан боғлиқ улкан вазифаларни юклаши, компьютер лингвистикаси йўналишининг пайдо бўлиши табиий тилларнинг тараққиёти учун муҳим аҳамият касб этади.

Дунё тилшунослигида тилни лингвистик моделлаштириш, сўзларни леммалаш ва теглар алгоритмини яратиш ҳамда миллий-маданий меросдан фойдаланиш имконини ошириш мақсадида муайян тилда яратилган оғзаки, ёзма ёдгорликлар, маънавий мерос намуналарини электронлаштириш ишлари қизғин тус олди. Компьютер технологиялари воситасида ахборотни қайта ишлаш, ахборот-ресурсларни жорий этиш учун зарур дастурий ҳамда методик таъминотни яратиш, интернет тизимида тил корпусини ва шу асосда миллий тил корпусининг илмий-назарий жиҳатларини ишлаб чиқишга алоҳида диққат қаратилмоқда.

Ўзбек тилшунослигида автоматик таржима қилиш, муаллифлик корпусининг лингвистик асосларини ишлаб чиқиш, лексикографик матнларга ишлов бериш ва лингвостатистик таҳлил этиш борасида бир қатор тадқиқот ишлари амалга оширилмоқда. Жумладан, “...давлат тилининг софлигини сақлаш, уни бойитиб бориш ва аҳолининг нутқ маданиятини ошириш; давлат тилининг замонавий ахборот технологиялари ва коммуникацияларига фаол интеграциялашувини таъминлаш”[1]га алоҳида эътибор қаратилди. Ўзбек тилининг халқаро миқёсдаги мақомини ошириш, уни жаҳон мулоқот тили даражасига кўтариш, ўзбек тилини чет элларда ўрганиш ва ўргатиш, миллий тилимизнинг имкониятларини кенгайтириш ва сайқаллаш ишлари ҳам бевосита миллий корпус орқали қулай амалга ошишини назарда тутсак, «ўзбек тили миллий корпусини яратишнинг назарий ва амалий масалалари»ни ҳал этиш долзарблик касб этади. Шу маънода, матн корпуси ҳамда миллий корпуснинг лингвистик асослари, унинг дастурий таъминотини яратиш технологияси юзасидан илмий тадқиқотларни янада чуқурлаштириш зарурати мавжуд.

Ўзбекистон Республикаси Президентининг 2016 йил 13 майдаги ПФ-4997-сон «Алишер Навоий номидаги Тошкент давлат ўзбек тили ва адабиёти университетини ташкил этиш тўғрисида», 2017 йил 7 февралдаги ПФ-4947-сон «Ўзбекистон Республикасини янада ривожлантириш бўйича Ҳаракатлар стратегияси тўғрисида», 2019 йил 21 октябрдаги ПФ-5850-сон «Ўзбек тилининг давлат тили сифатидаги нуфузи ва мавқеини тубдан ошириш чора-тадбирлари тўғрисида»ги фармонлари, 2017 йил 17 февралдаги ПҚ-2789-сон

[1]Ўзбекистон Республикаси Президентининг «Мамлакатимизда ўзбек тилини янада ривожлантириш ва тил сиёсатини такомиллаштириш чора-тадбирлари тўғрисида» ги Фармони. Манба:// https://lex.uz/docs/5058351

«Фанлар академияси фаолияти, илмий тадқиқот ишларини ташкил этиш, бошқариш ва молиялаштиришни янада такомиллаштириш чора-тадбирлари тўғрисида», Ўзбекистон Республикаси Вазирлар Маҳкамасининг 2019 йил 12 декабрдаги 984-сон «Давлат тилини ривожлантириш департаменти тўғрисидаги Низомни тасдиқлаш ҳақида», 2020 йил 29 январдаги 40-сон «Ўзбекистон Республикаси Вазирлар Маҳкамаси ҳузуридаги Атамалар комиссиясининг фаолиятини ташкил қилиш чора-тадбирлари тўғрисида»ги қарорлари ҳамда соҳага оид бошқа меъёрий-ҳуқуқий ҳужжатларда белгиланган вазифаларни амалга оширишда ушбу диссертация муайян даражада хизмат қилади.

**Тадқиқотнинг республика фан ва технологиялар тараққиётининг устувор йўналишларига мослиги.** Мазкур тадқиқот республика фан ва технологиялари ривожланишининг I. «Ахборотлашган жамият ва демократик давлатни ижтимоий, ҳуқуқий, иқтисодий, маданий, маънавий-маърифий ривожлантириш, инновацион иқтисодиётни ривожлантириш» устувор йўналишига мувофиқ бажарилган.

**Диссертация мавзуси бўйича хорижий илмий-тадқиқотлар шарҳи[2].** Табиий тилларнинг формализацияси ва уларнинг процессорларини математик аппаратлар асосида компьютер моделларини яратиш, корпус ва унинг турларини ўрганишга йўналтирилган илмий изланишлар жаҳоннинг етакчи илмий марказлари ва олий таълим муассасалари, жумладан, Принстон университети (АҚШ), Библиография институти (Германия), Оксфорд университети тил маркази (Англия), Монпелье университети (Франция), Уппсала университети (Швеция), Чарлз университети (Прага), Лингвистик тадқиқотлар институти (РАН), Компьютер лингвистикаси лабораториялари (РФ), Ломоносов номидаги Москва давлат университети (Россия), Ауезов номидаги Жанубий Қозоғистон давлат университети (Қозоғистон), шунингдек, Алишер Навоий номидаги Тошкент давлат ўзбек тили ва адабиёти университети, Бухоро давлат университети, Қарши давлат университети, Ўзбекистон Республикаси Фанлар академияси Ўзбек тили, адабиёти ва фольклори институти (Ўзбекистон)да олиб борилмоқда.

Жаҳон компьютер лингвистикасида корпусни тузиш, ўрганиш корпус лингвистикаси соҳаси ривожланишидан анча олдин бошланган. Чунончи, XVIII асрларда Библияга оид тадқиқотлар (мас., Cruden), луғатлар (Johnson, Oxford English Dictionary, Webster Dictionary), тилларни ўқитиш (частотный корпус Thorndike'a, 1921), Квирк корпуси (Survey of English Usage) шулар жумласидандир. Биринчи корпус АҚШдаги Браун Университетида 1960 йилларда яратилган Браун корпуси (Brown Corpus)дир. Браун тамойиллари бўйича яратилган Уппсала Корпуси (Уппсала Университети, Швеция) бўлса, Буюк Британияда Англия Банк лойиҳаси (Bank of English) ва Британия миллий корпуси (BNC) вужудга келди. Британия тамойиллари асосида кўплаб Европа тилларининг (испан, италян, хорват) миллий корпуслари

[2]Диссертация мавзуси бўйича хорижий илмий-тадқиқотлар шарҳи www.university ofcalifornia.edu, www.harvard.edu, www.indiana.edu, www.uni-bonn.de, www.krugosvet.ru.valentnost, www.scicenter.sintagmatika, www.philol.msu.ru/, https://iling.spb.ru/grammatikon, www.princeton.edu, https://bigenc.ru/linguistics/text/, www.navoiy-uni.uz ва бошқа манбалар асосида амалга оширилди.

яратилди. Табиий тилларнинг формализацияси ва уларнинг процессорларини математик аппаратлар асосида компьютер моделларини яратиш масалалари бўйича Наум Хомскийнинг тадқиқотлари амалга оширилган. Чарлотте Тайлор корпуснинг умумий ва махсус турлари мавжудлиги, махсус турдаги корпуслар жанр, услуб, даврларга кўра фарқланиши, ҳар икки турдаги корпус ўз навбатида диахрон ва синхрон шаклида бўлишини қайд этган.

Бугунги кунда жаҳон тилшунослигида миллий корпусларни яратиш бўйича қатор, жумладан, қуйидаги устувор йўналишларда тадқиқотлар олиб борилмоқда: Польша миллий корпуси, Алма-ата қозоқ тили корпуси, Шарқий арман корпуси, Тожик тили миллий корпуси ва ҳоказо. Чарлз Меер (Cambridge University Press , 2004) корпус лингвистикасида тадқиқот олиб бораётган муайян кичик доирадаги тадқиқотга доир методологик усулни ҳам корпус деб номлаш мумкинлиги таклиф этади. Доуглес Бибер (Oxford university, 2015) корпус аналитик технологиянинг миқдор (статистик) ва сифат хусусиятларини ўз ичига олиши кераклигини таъкидлайди. М.З.Курди (Great Britain, 2016) корпусларнинг тематик кўламини матнларнинг соҳавий таснифига кўра мувозанатлашган, имконият даражасидаги корпус турларига ажратади.

**Муаммонинг ўрганилганлик даражаси.** Жаҳон тилшунослигида миллий корпусни яратиш, унинг аналитик технологиясини, корпус лингвистикаси соҳаси ривожланиш каби масалалар таҳлилида А.Н.Хомский, Г.Н.Луч, Л.Блумфилд, Ч.К.Фрайс, Ҳ.Бонджерс, Н.Френсис, Ч.Ф.Меер, Ж.Синклер, М.З.Курди кабиларнинг илмий изланишлари алоҳида эътирофга лойиқ [3].

Рус тилшунослигида эса бу борада В.Г.Бритвин, В.П.Захаров, И.А.Мелчук, А.Б.Кутузов, Р.Г.Котов, Л.И.Беляева, Е.В.Недошивина, В.В.Риков, В.Плунгян каби олимларнинг катта массивли матнлар мажмуаси, корпус тузиш принциплари, лингвистик маълумотлар базаси, корпус соҳасидаги мақсадли тадқиқотларини келтириб ўтиш ўринли [4].

---

[3] Chomsky N., The logical basis for linguistic theory, Proc. IXth Int. Cong, of Linguists, 1962; Leech G. The State of Art in Corpus Linguistics // English Corpus Linguistics / Aimer K., Altenberg K.(eds.) – London, 1991. – P. 8-29.; Блумфилд Л. Язык. – М.: «Прогресс», 1968. – 608 с.; Fries Ch.C. The structure of English. An introduction to the construction of English sentences. – L.,1969.; Bongers H. The history and principles of Vocabulary control. – Woerden: WOCOPI, 1947; Френсис Н., Кучера Г. Вычислительный анализ современного американского варианта английского языка. – М., 1967.; Синклер Д. Предисловие к книге «Как использовать корпуса в преподавании иностранного языка»/ Д.Синклер [Электронный ресурс]. – Режим доступа: http://www/ruscorpora.ru/corpora-infro.html, свободный; Charlez Meyer English corpus linguistics: An introduction. Cambridge University Press, 2004. 168 p.; Mohamed Zakaria Kurdi. Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax, Great Britain, USA: Wiley-ISTE, 2016, 300 p.

[4] Бритвин В.Г. Прикладное моделирование синтагматической семантики научно-технического текста (на примере автоматического индексирования). КД.- М.: МГУ, 1983; Мельчук И.А. Порядок слов при автоматическом синтезе русского слова (предварительные сообщения) // Научно –техническая информация. 1985, №12. – С.12-36.; Захаров В.П. Корпусная лингвистика: учебник для студентов гуманитарных вузов. – Иркутск, 2011. – 161 с.; Кутузов А.Б. Корпусная лингвистика. – [Электрон ресурс]: Лицензия Creative commons Attribution Share-Alike 3.0 Unported [Электрон ресурс] – //lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf; Котов Р.Г. Лингвистические аспекты автоматизированных систем управления. – Москва: Наука, 1977.; Беляева Л.И., Чижаковский В.А. Тезаурус в системах автоматической переработки текста. – Кишинев, 1983.; Недошивина Е.В. Программы для работы с корпусами текстов: обзор основных корпусных менежеров. Учебно-методическое пособие. – Санкт-Петербург. − 2006. 26 с.; Рыков В.В. Курс лекций по корпусной лингвистике. URL: http://rykov-cl.narod.ru/c.html; Плунгян В. Зачем мы делаем

Ўзбек тилшунослигида матнни лингвостатистик таҳлил этиш, унга лексикографик ишлов бериш, автоматик таҳрир қилувчи дастурнинг лингвистик таъминотини, таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари, миллий корпусининг синоним сўзлар базаси, муаллифлик корпусининг лингвистик асослари борасида Х.Исхакова, С.Муҳаммедов, С.Ризаев, С.Муҳаммедова, Б.Менглиев, Д.Ўринбоева, А.Пўлатов, У.Дысимова, Г.Валиева, Г.Жуманазарова, Н.Абдураҳмонова, Ш.Ҳамроева, М.Абжалова, А.Эшмўминовларнинг ишлари диққатга сазовор[5].

Диссертацияни ёзиш жараёнида номлари кўрсатилган ва бошқа бир қатор ўзбек ҳамда жаҳон тилшуносларининг илмий изланишлари эътиборга олинди. Тадқиқотимизда мазкур йўналишда бажарилган ишлардан фарқли равишда, ўзбек тили миллий корпусини яратишнинг назарий ва амалий масалалари монографик тарзда текширилган.

**Тадқиқотнинг диссертация бажарилган олий таълим ёки илмий-тадқиқот муассасасининг илмий-тадқиқот ишлари режалари билан боғлиқлиги.** Диссертация Бухоро давлат университети илмий-тадқиқот ишлари режасига мувофиқ «Ўзбек тилшунослигида тил, шахс ва жамият муносабатлари тадқиқи муаммолари» мавзуси доирасида бажарилган (2017-2021 йй.).

**Тадқиқотнинг мақсади** ўзбек тили миллий корпуси лингвистик базасини шакллантиришнинг назарий ва амалий асосларини ишлаб чиқишдан иборат.

**Тадқиқотнинг вазифалари:**

яратилган дунё корпуслари, жумладан рус тили миллий корпуси ҳамда туркий тилларда яратилган (қозоқ, татар, турк тиллари) тажрибаларини ўрганиб, ўзбек тили миллий корпуси тузишнинг умумий тамойилларини аниқлаш;

---

Национальный корпус русского языка? [Электрон ресурс] «Отечественные записки» 2005, –№2. http://magazines. russ.ru/oz/ 2005/2/2005_2_20-pr.html

[5] Исхакова Х.Ф. Исследования в области формальной морфологии тюркских языков (на материале татарского литературного языка в сопоставлении с турецким и узбекским). Канд.дис…филол.наук.–М., 1972.; Муҳаммедов С.А. Статистический анализ лексико-морфологической структуры узбекских газетных текстов: Автореф. дисс… канд. фил. наук. – Ташкент, 1980.; Ризаев С. Ўзбек тилининг лингвостатистик тадқиқи: Фил.фан.док.дис…автореф. – Тошкент, 2008.; Муҳаммедова С. Ҳаракат феъллари асосида компьютер дастурлари учун лингвистик таъминот яратиш. Методик қўлланма. –Тошкент, 2006.; Ўринбоева Д.Б. Ўзбек фольклори матнларининг лингвостатистик тадқиқи. – Тошкент: Фан, 2010.; Пўлатов А. Компьютер лингвистикаси. – Тошкент: Akademnashr, 2011.-500 б; Дысимова У. Матндаги феълларни автоматик таҳрир қилувчи дастурнинг лингвистик таъмини (расмий-идоравий услубдаги матнлар асосида). Магистрлик дисс.–Тошкент, 2002, 56 б.; Валиева Г. Расмий-идоравий услубнинг лисоний бирликларини моделлаштириш. Магистрлик диссер.. –Тошкент, 2003, 60 б.; Норов А. Компьютер лингвистикаси асослари. – Қарши, 2017. – 136 б.; Жуманазарова Г.У. Фозил Йўлдош ўғли достонлари тилининг лингвопоэтикаси: Фил. фан. док. дис…автореф. –Тошкент, 2017.; Б.Менглиев Ўзбек тили миллий корпуси. 2018 йил, 26 апрель, http://marifat.uz/marifat/ruknlar/fan/1241.htm ;Абдурахмонова Н.З. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (Содда гаплар мисолида). Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф. – Тошкент, 2018.; Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф. – Қарши, 2018.; Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (Расмий ва илмий услубдаги матнлар таҳрири дастури учун)**.**Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф. – Фарғона, 2019.; Эшмўминов А.Ўзбек тили миллий корпусининг синоним сўзлар базаси. Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф. – Қарши, 2019.

корпус учун матнларни тайёрлаш жараёнини тавсифлаш, теглашнинг мавжуд стандартлари ва технологияларини моделлаштириш йўлларини асослаш;

миллий корпус ҳамда электрон манбаларининг компьютер форматида ва интернет тармоғи уланганлик, зарур маълумотларнинг нусхасини дискка ёки қоғозга кўчириб олиш каби ўхшаш ҳамда сўзларни автоматик таҳлил қилиш, уни тўлдириш, тузатиш, таҳрир қилиш, қидирув тизимининг мавжудлиги каби фарқли томонларини ёритиш;

миллий корпуснинг маълумотлар базасини шакллантириш ва тизим архитектураси сўз шакли ёки лемма бўйича гибрид излаш каби турини аниқлаш;

миллий-маданий меросдан фойдаланиш имконини ошириш мақсадида ўзбек тилда яратилган оғзаки матн, ёзма ёдгорликлар, шу тилда яратилган барча илмий, назарий ва амалий ҳамда маънавий мерос намуналарини электронлаштириш бўйича тамойилларини ишлаб чиқиш;

тилни сақлаб қолиш ва унинг ўрганилиш доирасини кенгайтириш, тил имкониятларини намоён этишнинг асосий воситаси интернет тизимида тил корпусларини яратиш.

**Тадқиқотнинг объекти** сифатида «Ўзбек тилининг изоҳли луғати» ва ўзбек тилидаги матнлар танланди.

**Тадқиқотнинг предмети**ни миллий корпус яратишда лингвистик база, корпус интерфейси, корпус бирликлари, корпус разметкасини моделлаштириш, лемма ва теглар ташкил этади.

**Тадқиқотнинг усуллари.** Тадқиқот жараёнида рационал-типологик, қиёсий-чоғиштирма, субстанциал, дискурсив таҳлил усулларидан фойдаланилди.

**Тадқиқотнинг илмий янгилиги** қуйидагилардан иборат:

ўзбек тилидаги сўзшаклларни, лексик бирликларни грамматик тавсифлаш асосида ўзбек тили миллий корпусини яратиш технологияси, миллий корпусда маълумотларни тақдим этиш тамойиллари назарий жиҳатдан асосланган;

миллий корпус разметкасининг SGML/XML тилига асосланган мавжуд стандартлари моделлаштириш йўлларини тартибга солувчи оддий матн, сарлавҳа, шеърий парчага оид қолиплар назарий хулоса асосида далилланган;

ўзбек тили миллий корпусининг лингвистик базасини тўлдирувчи матнларни разметкалаш формати, лексик маълумотни кодлаш талаблари ҳамда сўз бирикмаси, қўшма сўз ясалишида лингвистик моделлардан фойдаланиш имконияти аниқланган;

ўзбек тили миллий корпусининг лингвистик базасини тўлдирувчи матнларни тайёрлаш жараёнида: матнни олдиндан қайта ишлаш, уни белгилаш, корпусга киришни таъминлаш, уларни тез равишда кўп параметрли қидириш ва статистик ишлов бериш каби босқичлар орқали корпус-менежер тизими ҳамда корпус интерфейсини шакллантириш назарий жиҳатдан асосланган;

миллий корпусининг яратилиши муносабати билан таълим-тарбия жараёнининг юқори самарали виртуал муҳитини яратиш, компьютер

технологиялари воситасида ахборотни қайта ишлаш таҳлиллари асосида универсал имкониятларга эга кичик ҳажмли online лойиҳа яратилган.

**Тадқиқотнинг амалий натижалари** қуйидагилардан иборат:

миллий корпус интерфейсини лойиҳалаштиришнинг назарий жиҳати ишлаб чиқилган;

ўзбек тили («Ўзбек тилининг изоҳли луғати»)даги сўзларни морфологик, семантик, синтактик разметкалаш ҳамда лингвистик моделлаштириш усулларининг назарий ва амалий масалалари очиб берилган;

миллий корпус менежери, маълумотлар базаси ва тизим архитектурасининг лойиҳаси яратилган;

синтактик белгилаш, синтактик алоқаларни ўрнатиш тадбирлари ва сўз ёки ибораларга маълум бир синтактик белгиларни тақаш амали ишлаб чиқилган;

корпус тилшуносларнинг зарурий иш қуроли, оғзаки ва ёзма ёдгорликлар, миллий-маданий меросни акс эттирувчи ахборот манбаи, миллий тилни йўқолиш хавфидан сақловчи, миллатнинг мавжудлигини жаҳонга танитувчи восита сифатидаги масалаларни ўрганишда амалий аҳамиятга эга эканлиги асосланган.

**Тадқиқот натижаларининг ишончлилиги** ўзбек тилида яратилган оғзаки, ёзма ёдгорликлар, шу тилда яратилган барча илмий, назарий ва амалий ҳамда маънавий мерос намуналарини электрон кўринишдаги ўзбек тили миллий корпусини яратиш каби муаммонинг аниқ қўйилганлиги, ишнинг ўрганилиш доираси аниқ белгиланиши, назарий маълумотлар ва фактик материаллар ишончли илмий манбалардан олинганлиги, қиёсий-тарихий, тавсифлаш, лингвокультурологик, компонент таҳлил усуллари воситасида асосланганлиги, назарий фикр ва хулосаларнинг амалиётга жорий этилганлиги, олинган натижаларнинг ваколатли ташкилотлар томонидан тасдиқланганлиги билан изоҳланади.

**Тадқиқот натижаларининг илмий ва амалий аҳамияти.** Тадқиқот натижаларининг илмий аҳамияти ўзбек тили миллий корпусини яратиш, компьютер лингвистикаси йўналишидаги тадқиқотларнинг юзага келиши, ўзбек тилининг интернет тизимидаги электрон-маълумотлар базаси сифатидаги тараққиётини тадқиқ этишга оид назарий хулосалардан тилшунослик йўналишларидаги ишларда манба сифатида фойдаланиш мумкинлиги билан белгиланади.

Тадқиқот натижаларининг амалий аҳамияти ишдаги назарий умумлашма ва таҳлиллардан «Компьютер лингвистикаси» лабораториясини яратиш, «Компьютер лингвистикаси», «Машина таржимаси», «Корпусга асосланган таржимашунослик» каби махсус курсларни ўқитиш, ўзбек тилининг халқаро миқёсдаги мақомини ошириш, уни жаҳон мулоқот тили даражасига кўтариш, ўзбек тилини чет элларда ўрганиш ва ўргатиш, миллий тилимизнинг имкониятларини кенгайтириш ҳамда сайқаллаш ишлари орқали миллий корпус тузишда фойдаланиш мумкинлиги билан изоҳланади.

**Тадқиқот натижаларининг жорий қилиниши**. Ўзбек тили миллий корпусининг назарий ва амалий жиҳатларини аниқлаш жараёнида эришилган илмий натижалар асосида:

миллий корпус разметкасининг SGML/XML тилига асосланган мавжуд стандартлари моделлаштириш йўлларини тартибга солувчи оддий матн, сарлавҳа, шеърий парчага оид қолиплар ҳамда таълим-тарбия жараёнининг юқори самарали виртуал муҳитини яратиш асосида масофали электрон таълимни жорий этиш каби хулосалардан А5-037-рақамли «АКТ соҳасидаги касб-ҳунар коллежлари учун масофавий таълим электрон тизимини ишлаб чиқиш» номли фундаментал лойиҳада (2015-2017) фойдаланилган (Ўзбекистон Республикаси Олий ва ўрта махсус таълим вазирлигининг 2021 йил 9 февралдаги 89-03-763-сон маълумотномаси). Натижада сифатли таълим хизматлари имкониятларини ошириш ҳамда ахборот-ресурсларни жорий этиш учун зарур дастурий ҳамда методик таъминотни яратиш бўйича хулосалар чиқаришга эришилган;

ўзбек тили миллий корпусининг лингвистик базасини тўлдирувчи матнларни тайёрлаш жараёнида: матнни олдиндан қайта ишлаш, уни белгилаш, корпусга киришни таъминлаш, уларни тез равишда кўп параметрли қидириш ва статистик ишлов бериш ҳақидаги назарий хулосалардан ОТ-Ф1-002 рақамли «Ёшларда миллий ғоя ва мафкуравий иммунитетни шакллантиришнинг психологик механизмлари» номли фундаментал лойиҳада (2017-2020 йй.) фойдаланилган (Ўзбекистон Республикаси Олий ва ўрта махсус таълим вазирлигининг 2021 йил 9 февралдаги 89-03-763-сон маълумотномаси). Натижада миллий-маданий меросдан фойдаланиш имконини ошириш мақсадида муайян тилда яратилган оғзаки, ёзма ёдгорликлар, маънавий мерос намуналарини электронлаштириш хусусиятларини очиб беришга хизмат қилган;

ўзбек тили миллий корпусининг лингвистик базасини тўлдирувчи матнларни разметкалаш формати, лексик маълумотни кодлаш талаблари ҳамда сўз бирикмаси, қўшма сўз ясалишида лингвистик моделлардан фойдаланиш каби хулосалардан Ф1-ФА-0-13229-рақамли «Ҳозирги қорақалпоқ тилида функционал сўз ясалиши» мавзусидаги фундаментал лойиҳада (2012-2016 йй.) фойдаланилган (ЎзР ФА Қорақалпоғистон бўлимининг 2021 йил 21 январдаги 292/1-сон маълумотномаси). Натижада лойиҳа янги илмий-назарий маълумотлар билан бойитилган;

ўзбек тили миллий корпусини яратиш технологиясининг назарий ва амалий жиҳатдан асосланган «Ўзбек тили миллий корпусини яратишнинг назарий ва амалий масалалари» деб номланган монографиясидан 5А120102 – Лингвистика (ўзбек тилшунослиги) мутахассислиги учун «Корпус лингвистикасига кириш» фани бўйича маъруза ва амалий машғулотларда фойдаланилган (Ўзбекистон Республикаси Олий ва ўрта махсус таълим вазирлигининг 2021 йил 9 февралдаги 89-03-763-сон маълумотномаси). Натижада корпус учун матнларни тайёрлаш жараёнини тавсифлаш, теглашнинг мавжуд стандартлари ва технологияларини моделлаштириш йўлларини асослашга хизмат қилган;

корпус лингвистикаси ва унга турдош соҳалар атамалари асосида олий таълим тизими 5А120102 – Лингвистика (ўзбек тилшунослиги) мутахассислиги ва соҳа мутахассисларига мўлжалланган «Корпус

лингвистикасининг атамалар луғати» (ISBN 978-620-0-61316-5) нашр қилинган (Ўзбекистон Республикаси Олий ва ўрта махсус таълим вазирлигининг 2021 йил 9 февралдаги 89-03-763-сон маълумотномаси). Натижада йиғилган материаллар асосида ўзбек тили миллий корпусини яратишда ва «Корпус лингвистикаси» курсини ўқитишда мавжуд 257 та сўзни қамраб олувчи ҳамда ўзбек лексикографияси фондини тўлдиришга хизмат қиладиган луғат яратилган;

ўзбек тилидаги сўзшаклларни, лексик бирликларни грамматик тавсифлаш асосида ўзбек тили миллий корпусини яратиш технологияси, миллий корпусда маълумотларни тақдим этиш тамойиллари каби хулосалардан 5120100 – Филология ва тилларни ўқитиш (рус тили) таълим йўналиши талабаларига мўлжалланган «O`zbek tili» номли дарслигида фойдаланилган (Ўзбекистон Республикаси Олий ва ўрта махсус таълим вазирлигининг 2021 йил 9 февралдаги 89-03-763-сон маълумотномаси). Натижада ўзбек тили миллий корпуси яратишнинг назарий асослари ишлаб чиқилган;

ўзбек тилида яратилган оғзаки, ёзма ёдгорликлар, маънавий мерос намуналарини электронлаштириш бўйича тамойилларини ишлаб чиқиш, компьютер технологиялари воситасида ахборотни қайта ишлаш, машина таржимаси, электрон луғатшуносликни ривожлантириш, тезауруслар тузиш, интернет тизимида тил корпусини яратиш каби илмий қарашлардан «Бухоро» телерадиоканалининг «Менга сўз беринг», «Адабий муҳит», «Долзарб мавзу» дастурлари сценарийларини тайёрлашда фойдаланилган (Ўзбекистон Миллий телерадиокомпаниясининг 2020 йил 11 декабрдаги 1-450-сон маълумотномаси). Натижада тилда яратилган оғзаки, ёзма ёдгорликлар, маънавий мерос намуналарини электронлаштириш, миллий тил корпусини яратиш, ўзбек тилини интернет «тушунадиган» тилга айлантириш ҳақидаги маълумотлар радиоканал ижодий жамоасининг мавзу доирасида чуқурроқ мулоҳаза юритиши учун замин яратган.

**Тадқиқот натижаларининг апробацияси**. Тадқиқот натижалари 20 та, жумладан, 8 та халқаро ва 12 та республика илмий-амалий анжуманларида қилинган маърузаларда жамоатчилик муҳокамасидан ўтказилган.

**Тадқиқот натижаларининг эълон қилинганлиги.** Диссертация мавзуси бўйича 44 та илмий иш, жумладан, Ўзбекистон Республикаси Олий аттестация комиссиясининг докторлик диссертацияларининг асосий натижаларини чоп этиш тавсия этилган илмий нашрларда 14 та мақола (8 та республика, 6 та хориж), шулардан 2 та илмий мақола халқаро журнал ва Scopus базасидаги журналларида 4 та мақола чоп эттирилган. Натижалар 1 та дарслик ва 1 та ўқув қўлланма, 1 та луғат, 1 та монография ҳамда 4 та усулий қўлланмаларда ўрин олган.

**Диссертациянинг тузилиши ва ҳажми.** Диссертация кириш, тўрт асосий боб, хулоса ва фойдаланилган адабиётлар рўйхати ҳамда иловалардан иборат бўлиб, ишнинг умумий ҳажми 251 саҳифадан иборат.

## ДИССЕРТАЦИЯНИНГ АСОСИЙ МАЗМУНИ

**Кириш** қисмида диссертация мавзусининг долзарблиги ва зарурати асосланган, тадқиқотнинг мақсад ва вазифалари, объекти ва предмети тавсифланган, республика фан ва технологиялари ривожланишининг устувор йўналишларига мослиги кўрсатилган, тадқиқотнинг илмий янгилиги ва амалий натижалари баён қилинган, олинган натижаларнинг илмий ва амалий аҳамияти очиб берилган, тадқиқот натижаларини амалиётга жорий қилиш, нашр этилган ишлар ва диссертация тузилиши бўйича маълумотлар келтирилган.

Диссертациянинг **«Миллий корпус – ўзбек тилининг электрон лингвистик манбаси сифатида»** деб номланган биринчи бобида ўзбек тилининг интернет ва электрон тилга айланиши, миллий тилининг электрон ресурсларини (ўзбек тили корпуси, электрон луғатлари, интернет саҳифаларидаги матнлар) ва технологияларини такомиллаштириб, уларни яратиш зарурати ва унинг аҳамияти, сунъий интеллект ва электрон манбалар, корпус тушунчаси ва унинг электрон кутубхонадан фарқи, миллий корпус дунёда яратилган корпусларининг умумий ва фарқли жиҳатлари таҳлили каби назарий масалалар ёритилган. Ушбу бобнинг *«Сунъий интеллект ва электрон манбалар»* номли биринчи фаслида сунъий интеллект орқали тилнинг имкониятларидан фойдаланиш борасида замонавий ахборот технологиялари бениҳоя кенг қулайликлар эшигини очганлиги, у инсон онги бажариши мумкин бўлган бир қанча вазифаларни бажара олаётганлиги, сунъий интеллект маҳсули бўлмиш электрон манбалар инсонга зарар етказмаслик ва инсонларнинг оғирини енгиллаштириш мақсадида яратилаётганлиги ҳақида фикр билдирилган. Компьютер технологиясининг ривожланиши электрон луғатлар, таржима порталлари, терминологик маълумотлар банки, виртуал (электрон) кутубхона, электрон матнли корпуслар, электрон ҳукумат, электрон нашр, электрон дарслик ва қўлланмалар каби электрон манбаларнинг яратилишига асос бўлди. Сунъий интеллект турли амалларни бажаришга мўлжалланган алгоритм ҳамда дастурий тизимлардан иборат ва у инсон онги бажариши мумкин бўлган бир қанча вазифаларнинг уддасидан чиқа олади.

Инсониятнинг тафаккури ва сунъий интеллект тизимини солиштирадиган бўлсак, қуйидаги хулосага келиш мумкин: инсон тафаккури ижод қилувчи, мослашувчан, ҳиссий идрокдан фойдалана оладиган, ҳар томонлама, кенг қамровли билимдан фойдаланувчи устунликларга эга. Унда қуйидаги камчиликлар мавжуд: мураккаб ўтказувчан (ифодаловчи), тез ҳужжатлаштира олмайди, инсон тафаккури барқарор эмас. Ахборот (маълумот) қидирув тизими эса доимийлик, осон ифодаланувчанлик ва бир хиллик хусусиятига эгалик, осон ҳужжатлаштира олиш каби устунликларга эга. Шунингдек, унда бир қатор камчиликлар ҳам мавжуд, чунончи, сунъий, тор йўналишли, олдиндан дастурлаштирилади, айтиб туриш керак, белгили идрокдан ва махсус билимдан фойдаланади. Ишда ҳар иккала тизимнинг афзалликлари ва камчиликларини таҳлил қилиб, инсон тафаккурининг асосий афзалликлари, жумладан у кўп соҳада, масалан, ижодкорликда,

топқирликда, маълумот узатишда ва умуман мазмунан сунъий интеллектдан устун эканлиги далилланади.

Бобнинг *«Корпус лингвистикасида корпус терминига оид маълумотлар таҳлили»* деб номланган иккинчи фаслида корпус – электрон шаклда мавжуд бўлган матнларни, ҳужжатларни маълумотларни қайта ишлаш, уларни автоматик таҳлил, яъни морфологик, синтактик ва семантик таҳлил қилиш, морфологик анализ ва синтез қилиш билан бирга, ярим автоматик режимда маънони бузмасдан ишончли нутқ материалини мослаштириш даражасини текширишувчи мослама эканлиги таъкидланади. Лингвистик корпус – мавжуд маълумотларни матн ҳолида тақдим этибгина қолмай, матнни автоматик анализ, синтез қилиш таҳлилий хусусияти билан электрон кутубхонадан афзалроқ эканлиги ҳақида фикр юритилади.

Электрон кутубхона ва тил корпусининг фарқли томонлари: электрон кутубхонанинг қидирув бирлиги бу – асарнинг яхлит матнидир. Ундан маълум бир асарни қидириш мумкин. Интернетдан маълумотни автоматик қидиришни таъминлайдиган манба электрон ёки виртуал кутубхона ҳисобланади. Ҳозирда бундай кутубхоналарни турли-туман ном билан атайдилар. Масалан, “электрон кутубхона”, “виртуал кутубхона”, “е-кутубхона”, “e-library”, “digital library” каби. Бундай кутубхонада китоблар, газета ва журналлар китоб шкафларида эмас, балки компьютер хотирасидан ўрин олган. Босма, аудио(овозли), видео(кўринишли) ва мультимедиа(ҳаракатли) маълумотлар компьютернинг қурилмасида рақамли форматда сақланадиган маълумотлар тўплами кўринишида бўлади. Маълумотлар ҳажмининг катта кичиклигига қараб, тизимли сервер битта ёки бир неча тармоқларга уланган компьютерлар орқали ишлайди. Бу кутубхона маълумотлари электрон кўринишда бўлиб, компьютерлардан жой олади. Масалан, дунёнинг ихтиёрий нуқтасидан туриб, албатта, интернет мавжуд бўлган жойдан электрон кутубхона маълумотларидан фойдалана олиш, зарур маълумотларнинг нусхасини дискка ёки қоғозга кўчириб олиш учун компьютер ва у билан боғлиқ мосламалар бўлса кифоя. Бир неча дақиқада маълумот компьютер экранида намоён бўлади. Бунга фақат компьютер ва махсус интернет тармоқ орқали эришиш мумкин. Бир неча йил аввал бир мақолани ёки унга тегишли бирор фикрни топишга бир неча ой вақт сарфлаш зарур эди. Бугун бунинг учун қайсидир кутубхонага ёки бошқа шаҳарга бориш ва вақт сарфлаш машаққати йўқолди. Чунки қулайлик сифатида электрон кутубхоналар яратилди. Электрон кутубхонанинг одатдаги кутубхоналардан бир қанча афзалликлари бор. Масалан, жойнинг тежалиши, яъни электрон кутубхонанинг турли маълумотлари интернет саҳифасида жамлангани боис китобларни сақлаш учун махсус жойга зарурат йўқ. Бу саҳифани кутубхоналардаги махсус марказ мутахассислари маълумотларни мунтазам равишда компьютерга киритиб, йиғиб ва тўлдириб боради. Корпуснинг қидирув бирлиги эса тил бирлиги ва нутқ бирлиги шаклида бўлиши мумкин. Бундай матндан фақат ўқиш учунгина фойдаланиш эмас, балки бу матнларнинг турли грамматик изоҳлари мавжудлиги сабабли улар устида лингвистик амаллар бажара олиш мумкин. Унинг тезаурусдан фарқи шундаки, тезаурусда тушунча асосида қидирув амалга оширилса, корпусда

эса сўзшакл ва унинг қўлланиш матни қидирилади. Корпус турли хил луғатларни (частота, топонимлар, грамматик сўзлар, иборалар ва ҳоказо) ўз ичига олган лексикографик бирлик сифатида устунлик қилади. Корпус замонавий луғатчиликда катта аҳамиятга эга. Шунинг учун у катта ҳажмли луғатларни тузиш учун манба вазифасини ўтайди. Вақт ўтиши билан корпус турли лингвистик йўналишлар учун зарур бўлиши билан катта (кенг) ахборот ресурсига айланиб боради. Корпус асосида луғатлар аввалгига нисбатан тезлик билан тузилади ва қайта ишланади. Корпус иш тизимида мавжуд матнлар саралаш хусусиятига эга. Тадқиқотчи ўзи учун керакли бўлган мисолни барча матнлардан эмас, балки тадқиқот учун аҳамиятли, зарур бўлганини ажратиш имконига эга бўлади. Электрон кутубхона эса санаб ўтилган хусусиятларга эга эмас.

Бобнинг *«Амалдаги дунё корпусларининг умумий ва фарқли жиҳатлари»* деб номланган учинчи фаслида интернет тизимида жойлашган 19 та тил корпусларининг тавсифи таҳлилга тортилган.

Дунёда яратилган корпусларни тузиш мезонлари қуйидагилар: матннинг яратилиши ва таркиби, синхронлаштириш, турли жанрларнинг тақдим этилиши, матнларнинг сонлари нисбати ва махсус эҳтимоллик амаллари асосида алоҳида матнларни саралаш, компьютер таҳлили учун матнларнинг қулайлиги (матнлараро ўзига хослигини етказиш учун махсус белгилар қўйиб чиқиш) эътиборга олинган.

Ҳозирда фаолиятда бўлган корпуслар тилдан фойдаланишда унинг статистик анализи, табиий тилни қайта ишлаш (NLP) дастурий таъминоти, лексик ресурларни яратиш, тил ўқитишда ёки ўрганиш каби мақсадларда қўлланилади. Шу ўринда таъкидлаш жоиз, М.Абжалова томонидан табиий тилни қайта ишлашда таҳрир ва таҳлил дастурларининг лингвистик модуллари яратилган, матнларни графематик, морфологик ва синтактик таҳлил қилиш жараёнлари тадқиқ этилган[6]. Корпус таркибига тақдим этилган матнлар тилнинг динамик ҳолатини тадқиқ қилишда ёки лингвистиканинг турли соҳа предметига кўра анализ қилишда муҳим ҳисобланади.

Дунё корпуслари ва яратилган корпусларнинг йиллар бўйича тақсимоти[7], матн корпуслари яратишнинг асосий даврлари, инглиз ва рус тили корпуслари, уларнинг турли таснифи[8] корпус лингвистикаси бўйича олиб борилган тадқиқотларда ўз аксини топган. Ўзбек тилшунослигида ва компьютер лингвистикаси соҳасида амалга оширилган тадқиқотларда яратилган дунё корпусларининг айримлари ҳақида маълумот берилган, аммо тадқиқотлардаги таснифлар тўлиқ ёритилмаган. Шу боис ушбу бобда мавзу моҳиятидан келиб чиққан ҳолда, яратилган тил корпусларининг, жумладан, 19 та дунё корпусларининг интернет тизимидаги жойлашуви, ишлатиш тили, яратилган йили, қўлланилган сўз миқдори, статистик анализи, табиий тилни қайта ишлаш (NLP) дастурий таъминоти, лексик ресурларни яратиш, 19 та

[6] Абжалова М. Таҳрир ва таҳлил дастурларининг лингвистик модуллари: Монография – Т., 2020. – 176 б.

[7] Эшмуминов А. Ўзбек тили миллий корпусининг синоним сўзлар базаси. Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф.. – Тошкент 2019.

[8] Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф. – Тошкент, 2018.

дунё корпусларининг тил ўқитишда ёки ўрганишдаги муштарак ва фарқли жиҳатлари таҳлил этилган.

Бобнинг *«Ўзбек тилшунослигида миллий корпуснинг лингвистик мақоми хусусида»* деб номланган тўртинчи фаслида миллий тилни йўқолиш хавфидан сақловчи, миллатнинг мавжудлигини жаҳонга танитувчи восита сифатида корпус – лингвистик хазина эканлиги илмий асослаб берилган.

Корпус бу сўзлар, сўз бирикмалар, грамматик шакллар маъносини маълум бир излаш тизими орқали топишнинг электрон шаклдаги матнлар тўпламидир. Корпусларнинг ҳар хил турлари мавжуд. Масалан, бир муаллиф корпуси[9], бир китоб корпуси (жумладан дастлабки корпуслар “Библия” учун қилингандир). Маълум бир тилнинг миллий корпуси шу тилнинг барча қирраларини, жанрларини, усулларини, худудий ва ижтимоий вариантларини ўзида ифода этади.

Лингводидактика сифатида тил корпуси она тили ва хорижий тилни ўрганишда бирдек аҳамиятли. У таълим олиш самарадорлигини ошириш учун янги имкониятлар эшигини очиб беради. Корпус орқали жуда осонлик билан кам ишлатиладиган сўз, ибора ҳамда сўз бирикмасини топиш ёки унинг қўлланиши ва имлоси (орфографияси) билан боғлиқ муаммо жуда қисқа муддат ичида ҳал этилади. Шуни таъкидлаб ўтиш жоизки, тил корпусидаги маълумотлар грамматика ёки дарсликда тавсифланганидек эмас, балки жамиятда қандай яшаса, шундай борича акс этади. Бу эса умумхалқ тили ва адабий тилни ўрганишда энг сермаҳсул восита сифатида хизмат қилади. Бугун грамматика билан шуғулланувчи олимдан кўра оддий тадқиқиотчи маълум бир сўз, ибора ёки конструкциянинг қўлланилиш ҳолати, даражаси, ким, қачон илк марта бу тузилмани қўллаганлиги, қандай услуб учун хосланишини билишга кўпроқ эҳтиёж сезади. Корпус эса мана шу каби муаммоларни ҳал қилишга йўналтирилгандир.

Миллий корпус мавжуд тилнинг лексикаси ва грамматикасини ўрганиш учун зарур. Корпуснинг бошқа вазифаси эса кўрсатиб ўтилган тилшуносликнинг сатҳ ва соҳалари бўйича (лексикология, акцентология, тил тарихи) тегишли маълумотларни етказиб беришдир. Тилнинг электрон корпуси – нафақат тилшунослар, балки ўзбек тилидан фойдаланувчи барча кишилар: турли соҳа мутахассислари, олимлар, сиёсатчилар, луғатшунослар, тадқиқотчилар учун фойдали. У ҳар хил мақсадларда ишлатилиши мумкин бўлган кенг қамровли универсал ахборот-қидирув тизимидир.

Миллий корпусни яратиш – статистик тадқиқ методи, компьютер таржимаси, нутқни синтезлаш ва уни таниш, орфографик текширув каби лингвистик амалларни бажариши корпус лингвистикасининг кейинги тараққиёт босқичини амалга оширишга кўмаклашади.

Ишнинг **«Миллий корпус тузишнинг умумий тамойиллари ва маълумотларни тайёрлаш технологияси»** деб номланган иккинчи боби уч фаслдан иборат. Унда миллий корпус тузишнинг умумий қоидалари, корпусда маълумотларни тақдим этиш тамойиллари, корпус учун матнларни тайёрлаш технологиясининг илмий жиҳатлари асослаб берилган.

[9] Қаранг: Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филол.фан.бўйича фалсафа доктори (PhD)…диссер. автореф. – Тошкент, – 2018

Бобнинг «*Миллий корпус тузишнинг умумий қоидалари»* деб номланган биринчи фаслида тўпланган материалларни, машина сақловчиларига (компьютерларга) жойлаштириш; электрон қидирувга (морфологик, синтактик сатҳда) имкон берадиган ўзига хос белгилар ва репрезентативлиги корпуснинг (тилдаги кўплаб жанрларнинг асл ҳолида тўлиқ акс этиш) муҳим омил эканлиги назарий жиҳатдан ўрганилган. Корпус структураси, дастурнинг интерфейси, дастур ишлашининг алгоритми, натижаларни келтириб чиқариш технологияси қандай амалга ошириш мумкинлиги хусусидаги таклифлар берилган.

Корпус яратишнинг технологик жараёни тўғрисида В.В.Рыков[10], Ю.Н.Марчук[11], И.А.Мельчук[12], Ш.Ҳамроева[13]ларнинг технологик жараён босқичларини қўллаб-қувватлаган ҳолда қуйида ўзбек тили миллий корпусининг технологик жараён босқичларини таклиф этамиз:

*1.Матнни олдиндан қайта ишлаш босқичи*. Ушбу босқичда турли манбалардан олинган барча матнлар имловий тузатилади ва таҳрир қилинади. Матн библиографик ва экстралингвистик тавсифга тайёрланади.

а) конверсия ва график таҳлил қилиш босқичи. Аксарият матнлар дастлабки қайта ишлаш жараёнида кўриб чиқилади. Хусусан, тилининг компьютер формати учун кодлаш ва автоматик таҳлил учун зарур бўлмаган элементлар (расмлар, жадваллар) ҳамда матндаги таг ости чизиқлар олиб ташланади.

б) *автоматик маркировка босқичи*. Бунда автоматик маркировка натижаларини тўғирлаш, яъни хатоларни тузатиш ва ажратиш амалга оширилади (қўлда ёки ярим автоматик).

*2. Матнни белгилаш босқичи*. Ушбу босқичда корпуснинг зарурий маълумотлари (метадата) киритилади. Корпус матнларининг мета-тавсифи: библиографик маълумотлар, матннинг жанри ва услуб хусусиятларини тавсифловчи белгилар, муаллиф ҳақидаги маълумотлар ва бошқаларни ўз ичига олади. Ушбу маълумотлар одатда қўл меҳнати орқали киритилади. Матннинг таркибий қисмлари (параграфлар, жумлалар, сўзларни танлаш) ва соф лингвистик белгилари кўпинча автоматик равишда амалга оширилади.

3. *Корпусга киришни таъминлаш босқичи*. Корпус дисплейи қуйидаги кўринишда: CD-ROMда тарқатилиши ва глобал тармоқ режимида мавжуд бўлиши мумкин. Фойдаланувчиларнинг турли тоифаларига турли хил ҳуқуқларга ва хилма хил имкониятларга эга бўлади.

*4. Якуний босқич* – тегли матнларни тезкор равишда кўп параметрли қидириш ва статистик ишлов беришни таъминлайдиган ихтисослаштирилган лингвистик маълумот олиш тизимининг таркибига ўзгартириш (корпус менежери) киритиш босқичи.

---

[10] Рыков В.В. Курс лекций по корпусной лингвистике. URL: http://rykov-cl.narod.ru/c.html.

[11]Марчук Ю.Н. Основы компьютерной лингвистики. - М.: Изд-во МПУ, 2000.

[12] Мельчук И.А. Порядок слов при автоматическом синтезе русского слова (предварительные сообщении) / Научно-техническая информация. 1985, № 12. – С.12-36

[13] Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филол.фан.бўйича фалсафа доктори (PhD)…диссер.–Қарши, 2018. -Б.45.

Албатта, ҳар бир ҳолатдаги босқичларнинг таркиби ва сони юқорида санаб ўтилганлардан фарқ қилиши ва реал технология анча мураккаб бўлиши ҳам мумкин.

Бобнинг «*Корпусда маълумотларни бериш тамойиллари*» деб номланган иккинчи фаслида Р.Г.Пиотровский, Д.Н.Лавров ва унинг шогирдлари томонидан матн танлаш, корпусдан маълумотларни тақдим этиш тамойилларини ишлаб чиқилганлиги[14], уларнинг тажрибаларига таянган ҳолда ўзбек тили миллий корпуси учун маълумотларни кодлаш форматига қўйиладиган талаблар тавсифлаб берилган. Ушбу фаслда ўзбек тили миллий корпусининг қидирув тизимига бўлган асосий талаблар қуйидагича ифодаланган:

1) сўз ва сўз бирикмасини уларнинг белгилари (грамматик, семантик ва б.қ)га қараб қидириш;
2) матн (нутқ ёки асарнинг тугал фикр англатувчи парчаси) ва сўзлар орасидаги масофани ҳисобга олиш;
3) метаматнли маълумотни қидириш;
4) тараққий этган тил талаблари, ўз ичига мантиқий боғламалар, қавс ва матн операторларни қамраб олиш;
5) индексациялаш самарадорлиги;
6) энг мураккаб саволга юқори тезликда жавоб топиш;
7) кенг кўламлилик, энг катта ҳажмгача сўзларни ишлатиш (юз миллионлаб сўзларни ишлатиш).

Корпус маълумотларини кодлаш энг нуфузли стандартлар асосида ишланади. Чунончи, TEI (Text Encoding Initiative), XCES (XML Corpus Encoding Standard), EAGLES (European Advisory Group on Language Engineering Standards. Миллий корпуснинг маълумотларини тақдим этишда лингвистик ахборотни ташувчи матн разметкаси SGML/XML тили негизида амалга оширилади.

Корпусда асосан икки турдаги матн маълумоти тақдим этилади:

*А. Катта массивли матн маълумоти.* Матнни тўлиқ ифодалайдиган белгиларни қамраб олган: муаллиф номи, унинг жинси, туғилган санаси, матн сарлавҳаси, матн яратилиш вақти, сўзлар ҳажми, тематикаси, матн тури, услуби, қўлланилиш соҳаси ва ҳ.к.

*В. Лексик маълумот.* Лексик маълумот қуйидаги белгиларни ўз ичига олган: алоҳида сўзларни ифодалайди, яъни матнлар корпусида аниқ бир жойда сўз шаклини ишлата олади. Бунга қуйидагилар киради:

*В.1. Морфологик белгилар:*

- лексема (сўз шакли);
- лексеманинг грамматик белгилари (сўз туркуми, жонли нарсалар, ўткинчи ҳодисалар);
- сўз шаклнинг грамматик белгилари (сон, келишик, майл, вақт, шахс).

[14] Шаров С.А. Представительный корпус русского языка в контексте мирового опыта. (Электрон ресурс. https//lamb.viniti.ru) Лавров Д.Н., Харламова М.А., Костюшина Е.А. Модель представления экстралингвистической и тематической разметки в корпусе народной речи // У1-я Междунар. науч. конф. «Математическое и компьютерное моделирование», посвящ. памяти проф. Б.А. Рогозина. 23 ноября 2018. — С. 115-118.; http://ruscorpora.ru/new/sbornik2005/11polyakov.pdf

*В.2. Семантик белгилари:*

семантик разряд, таксономик синф, мереология, баҳо, каузация, сўз ясовчи алоқалар ва б.[15].

Корпусда матн абзацлар кетма-кетлигидан иборат бўлса, абзацлар гаплардан, гаплар эса сўзлардан иборат. Бунда таҳлилнинг асосий бирлиги сўз деб олинса, матн бирлиги эса гап деб қабул қилинади. Корпусда қидирув тизими орқали аниқ бир белгига доир сўз ва сўз бирикмаларни фақат мазкур гапга оид топа олиш имкони мавжуд. Қидирув натижаси гаплар рўйхати ҳисобланиб, унда топилган сўзлар ажратилган шрифт орқали ифодаланади. Керак бўлган пайтда қидирув матни абзац чегарасигача кенгайтирилиши мумкин, лекин ундан ортиқ эмас.

Шундай қилиб, корпусда асосий структурали бирликларни ажратиш мумкин: сўз, гап, абзац, матн. Бунда матннинг структурали бўлиниши (қисмлар, боблар, бўлимлар)ни ифодалайдиган, абзацдан ташқарида бўлган бирликлар ва гапнинг синтактик структураси (клауз, гуруҳлар)ни ифодалайдиган бирликлар ишлатилмайди.

Бобнинг учинчи фасли «*Корпус учун матнларни тайёрлаш технологияси*» деб номланган. Унда миллий корпус учун танланган матннинг илк разметкаси, корпус таркибига кирадиган матнларнинг турлари, лингвистик разметка кўринишлари илмий жиҳатдан асослаб берилган.

«Ўзбек компьютер лингвистикаси ўзбек тилининг инглиз тилидан тамомила фарқ қиладиган хусусиятлари асосида шакллантирилади. Бу эса ўзбек компьютер лингвистикасини яратишдан олдин ўзбек тилини мукаммал даражада системалаштириш, формаллаштириш вазифаларини амалга ошириш зарурияти мавжудлигини кўрсатади. Ўзбек тили каби бой, кенг ва чукур ривожланган тил масалаларини компьютерда ечиш даражасига олиб чиқиш инглиз тилига қараганда катта ҳажмда иш бажаришни талаб қилади», - дея таъкидлайди А. Пўлатов[16]. Олимнинг фикрига қўшилган ҳолда, ўзбек компьютер лингвистикасини яратишда инглиз компьютер лингвистикасидан тўғридан-тўғри фойдаланиб бўлмаса-да, унинг асосий ғояларига таяниш мумкин. Ўзбек тилининг тил корпусини тузишга мўлжалланган лингвистик база ва миллий матнлар банки тайёрланишда Рус тили миллий корпуси бўйича олиб борилган тадқиқот ишларига мурожаат қилинди. Ишда корпус таркибига кирадиган матн турларни ажратишда В.П.Захаров[17], А.Е.Поляков[18] кузатишларига таяниб, корпус учун матнларни тайёрлаш жараёни қуйидагиларга ажратилган:

1) HTML минимал форматда матннинг илк разметкаси;

---

[15] Аброскин А. А. Поиск по корпусу: проблемы и методы их решения // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. –277–282 с.; ПоляковА.Е. Технология подготовки информации в национальном корпусе русского языка. http://www.ruscorpora.ru/new/corpora-biblio.html; Кустова Г. И., Ляшевская О. Н., Падучева Е. В., Рахилина Е. В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. –М., 2005.– С.155–174.

[16] Пўлатов А. Қ. Компьютер лингвистикаси / А.Қ.Пўлатов; масъул муҳаррир: А.А.Абдуазизов, М.М.Орипов. - Т.: Akademnashr, 2011. - 520 б. (-Б. 7.)

[17] Захаров В.П. Корпусная лингвистика. Учебно-методическое пособие. – Санкт-Петербург, 2005. – 48 с.

[18] Поляков А. Е. Технология подготовки информации в Национальном корпусе русского языка Текст. / А.Е. Поляков // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. – М., 2005. –С. 192.

2) морфологик разметка ва омонимия (корпус қисмида)нинг аниқланиши;

3) метаматнли разметка;

4) Яндекс-сервер учун чиқиш форматига ўзгартириш.

Электрон корпусдаги лексик маълумотни кодлаш HTML/XML қоидаларига мослаштирилади. Бу эса матннинг турли хил турдаги дастурлар, қидирув индексатори, морфологик парсер, конверторлар, таҳрирлаш босқичларида тезкор қайта ишланиши ва корпусда разметкани автоматлаштириш учун кенг имкониятларини очиб беради. Миллий корпус учун матнлар турли хил манбалардан олиб киритилади ва ҳар хил форматларда ифодаланади.Чунончи, оддий матн, HTML, RTF, PDF каби.

Матнни тайёрлаш жараёнида матндан муаллифга тегишли бўлмаган ёки тил ўрганиш учун аҳамиятли бўлмаган қуйидаги элементлар олиб ташланади: саҳифа рақамлари, устун сарлавҳалари, титул саҳифалар, мундарижа, чиқиш маълумотлар, тизимли ёзув, аннотациялар, муҳаррир изоҳлари (муаллиф томонидан ёзилган изоҳлар сақланади), расмлар, схемалар, формулалар (лекин улар остида имзолар сақланади);

Лингвистик ва экстралингвистик разметкалар маълумотлар ифодасининг ягона формати бўлиб, корпус бўйича маълумот алмашишга шароит яратади.

Миллий корпуснинг технологик жараёни қуйидагилардан иборат: танланган матнлар асосида лексема ва сўз шаклларининг такрорланиш луғатини яратиш; олинган такрорланиш луғатининг ҳар қандай бирлиги учун матнни кўриб чиқиш; графикли сўзни бўғинга ажратиш ва бўғинларнинг такрорланиш луғатини тузиш; сўз захираларини саралаш; бир вақтнинг ўзида чекланмаган файлларни қайта ишлаш; ташқи белгиларга эга бўлган матнлар корпусларини яратиш; яратиладиган матн корпус ҳамда корпусга кирувчи алоҳида матнлар учун статистик маълумотларни ҳисоблаб чиқиш кабилардир.

Диссертациянинг **«Ўзбек тили миллий корпусининг лингвистик базасини шакллантириш»** деб номланган учинчи боби беш фаслдан иборат бўлиб, унда электрон корпуснинг лингвистик базаси таъминоти, материали, моделлаштиришнинг аҳамияти ва сўз бирикмаси таҳлилида моделлардан фойдаланиш каби назарий масалалар кўрилган.

Бобнинг «*Матн корпусининг лингвистик база таъминоти*» деб номланган биринчи фаслида маълумотлар базаси(МБ), унинг таркибий қисмлари, ахборот тизимининг хусусиятлари, миллий корпус яратишда МБнинг асосий элементлари: жадваллар, сўровлар, маълумотлар схемалари, формалар, ҳисоботлар, макрослар ва модуллар каби муаммолар ечими ўз ифодасини топган. Шунингдек, ўзбек тили миллий корпусини яратиш учун МБ шакллантириш режаси ҳамда маълумотлар базасини моделлаштиришнинг босқичлари ёритилган.

Маълумотлар базаси – маълумотларни сақлаш, янгилаш, қидириш ва етказиб беришни таъминловчи ахборот, дастурий таъминот[19]. У аппарат ва ходимларнинг комбинациясини ифодаловчи автоматлаштирилган тизимдир.

[19] Дейт К. Дж. Введение в системы баз данных. 8-е издание.: Пер. с англ. – М.: Издательскийдом "Вильяме", 2005. –1328с.

Тилшуносликда ушбу технологияни ишлаб чиқиш ва шунга ўхшаш манбаларни яратиш қуйидаги муаммоларни ҳал қилади:

1) эмпирик материални тузилиши ва бирламчи таҳлил қилиш муаммоси, тил даражалари бирликларини (грамматикалар, луғатлар, фонетик маълумотлар базалари) вазифасидан бошлаб, тўлиқ матнларни ўрнатишга имкон беради. Бир томондан, тил тизимининг таркибий моделини тўлдириш ва аниқлаштириш, бошқа томондан, дискурсив ҳудудларнинг миллий моделлари ва умумий тил тизимининг моделини яратиш;

2) тил маълумотларини ўрнатиш ва сақлашнинг янги усулларини топиш, шунингдек, ушбу материалларга киришни ташкил этиш вазифаси;

3)тадқиқотни оптималлаштириш ва янги натижаларни олиш учун материални қайта ишлашнинг янги усулларини топиш вазифаси;

4) катта материалга мурожаат қилиб, ўрганиш натижаларини текшириш вазифаси ҳал қилади.

Лингвистик таъминот – муайян соҳасида тилнинг етарли даражада ишлашини таъминлайдиган тил воситаларининг умумийлигидир. Ўзбек тилининг электрон корпусини вужудга келтиришда лингвистик таъминот мукаммал даражада бўлсагина, унинг дастурий таъминоти яратилади.

Бобнинг иккинчи фасли «*Киберлексикография миллий корпуснинг лингвистик материали сифатида*» деб номланган. Ушбу фаслда лексикография соҳасида эришилган муайян ютуқлар, давр талабларидан келиб чиққан ҳолда, луғатчиликнинг янги назарий муаммоларига ҳам эътибор қаратилган. Шунингдек, киберлексикография[20] ва кибер-луғатчилик муаммоси ҳақида тўхталиб ўтилган.

Бугунги кунда у “виртуал луғатлар, улар билан ишлаш ҳамда яратиш технологиялари” мазмунида оммалашмоқда. Бунда соҳани интернет тизими доирасида идрок этиш лозим бўлади. Киберлексикография[21] атамаси шу маънода интернет электрон луғатларни – умумий ва махсус турдаги академик, энциклопедик ва лингвистик луғатларни яратишнинг назарий асосларини ўзида ифодалайди. Киберлуғатларнинг ёрқин намунаси бўлган корпусларнинг яратилиши лексикография соҳасини янги чўққиларга олиб чиқди. Маълумки, корпус ёзма матнларнинг электрон шаклдаги тўплами бўлиб, махсус ишлаб чиқилган компьютер дастурлари орқали тўпланган луғатлар жамланмаси киритилиши асосида яратилади[22]. Корпус компьютер дастури шаклида тилнинг барча жиҳатларини қамраб олади. Шунинг учун ҳам, корпус лингвистикаси тилнинг хусусиятларини эътиборга олган ҳолда уни қайта баҳолашга олиб келади.

---

[20] Карпова О. М., Менагаришвили О. В. Электронные словари и кибернетическая лексикография *:* метод. рекомендации к спецкурсу. –Иваново: Иван. гос. ун-т, 2002. – 45 с.; T.Valiyev. Kibernetik leksikografiya va til korpusi muammolariga doir. SamDU. Ilmiy axborotnoma filologiya 2016-yil, 2-son. -B.67-70

[21] http://://studfile.net/preview/1619320/page:3/ Кибернетическая лексикография Захаров В. 1 А -10 ФИЯ ЧГУ имени И. Н. Ульянова.[Электрон ресурс]. http://www.myshared.ru/slide/10492/

[22] Сивакова Н.А.Лексикографическое описание английских и русских фитонимов в электронном глоссарии. дисс. .д-ра филол. наук в форме науч. докл. Тюмень., 2004. - 72 с..; Саженин, И. И. Словарный корпус: проблемы определения и структурной организации / И. И. Саженин; отв. ред. И.П. Матханова. // Проблемы интерпретационной лингвистики: типы восприятия и их языковое воплощение: межвузовский сборник научных трудов. – Новосибирск: Изд-во НГПУ, 2013. – С. 294 – 298

Ахборот-қидирув тизимига солинган луғатлар тўплами бўлган корпус киберлексикографик корпуснинг асосий манбаси ҳисобланади. Киберлексикографик корпуслар махсус ишлаб чиқилган компьютер дастури орқали амалга оширилган бўлиб, зарурий равишда сўзларни акс эттиришни ўз зиммасига олади. Энг асосийси, корпус дастури берилган мақсадли сўзни тадқиқ этади, корпусдаги мисоллар сонини аниқлайди ва боғлиқлик частотасини ҳисоблайди, мақсадли қисмга хос мисолларни намойиш этади, бундан фойдаланувчи кейинги тадқиқотларни давом эттира олиш имконига эга бўлади. Ўзбек тилининг миллий киберлуғатларини яратиш долзарб масала бўлиб, бу ўзбек лексикографиясининг, унинг маҳсули сифатидаги миллий Интернет луғатларининг жаҳон виртуал олами билан боғланиш ва улардан озиқланиш имконини беради. Киберлексикография соҳасини ривожлантириш киберлуғатларнинг ўзбек тилидаги тўлиқ ва тўлақонли базасини яратиш, автоматик таҳрирлаш, ўзбек тилидан бошқа тилга ёки аксинча таржима қилувчи мукаммал дастурларни яратишни тақозо қилади.

Бобнинг «*Лингвистик базани тузишда лингвистик моделлаштиришнинг аҳамияти*» деб номланган учинчи фаслида лингвистик базани тузишда моделлаштириш методининг аҳамияти ва компьютер иши жараёнида бошқариладиган кўрсатмаларни ишлаб чиқаришнинг асосий алгоритмини тузиш, ҳар бир сўз туркумини разметкалаш учун махсус модел шаклларини ишлаб чиқилиши хусусида назарий маълумот ҳамда таҳлиллар берилган.

Компьютер лингвистикасида *лингвистик модуль* атамаси муҳим аҳамият касб этади. Чунончи, табиий тилнинг компьютер тилига ўтказилиши, яъни компьютер тизими орқали матнга ишлов бериш йўллари яратилади. Бунинг учун ўзга тилларда яратилган дастурларнинг илғор таржималардан фойдаланилади. Лингвистик модуль ана шундай дастурларнинг мустақил таркибий қисмлари ҳисобланади. Масалан, лексик модулда луғат қатлами (сўзлар) қуршаб олинса, графематик модулда рамзлар, тиниш белгилар, ҳарфий ва бошқа белгилар таҳрир қилинади, орфографик модулда имло қоидалари, морфологик модулда сўзшакллар анализи (сўзшаклдан лексемага қадар таҳлил) ва синтези (лексеманинг грамматик шаклланиши таҳлили жараёни), синтактик модулда суперсинтактик бирлик – гап ёки сўзларнинг ўзаро боғланиш ҳодисаси таҳлил қилинади. Ўзбек тили миллий корпусини яратишда тилнинг хос хусусиятидан келиб чиққан ҳолда, унинг алгоритми тузилади.

Ўзбек тили миллий корпуси ўзбек тилида мавжуд бўлган лексик бирликларнинг, чунончи, синоним, антоним, омоним, ўзлашма сўзлар, сўзларнинг даражаланиши, сўзнинг морфологик таркиби, сўзнинг ясалиши, сўзларнинг маъноси, унинг морфологик хусусиятларини автоматик таҳлил қилиб бера олиши керак. Яъни корпусни тузиш, леммалаш, разметкалаш жараёнида корпус таркибига кирган шундай сўзларни бирма-бир қидирув асосида матнлар ичидан топиш ва уларни махсус изоҳлаш керак бўлади. Бунинг учун эса юқорида айтилган алгоритм, лингвистик моделлаштириш ишлари амалга оширилиши лозим. Бунда М.Абжалованинг “Ўзбек тилидаги

матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари”[23] деб номланган тадқиқотидан, лексик бирликлар билан боғлиқ муаммоларда А.Эшмўминовнинг “Ўзбек тили миллий корпусининг синоним сўзлар базаси”[24] тадқиқотидан, сўзларнинг морфологик хусусиятларини автоматик таҳлилида Ш.Ҳамроеванинг **“**Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари”[25] мавзусидаги тадқиқотининг айрим ўринларидан, ўзбек тилидаги лексик бирликларни таржима қилиш билан боғлиқ масалаларда Н.Абдурахмонованинг “Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти”[26] тадқиқотидан ўринли фойдаланиш зарур.

Лексик бирликларни разметкалашда ўзбек тилшунослигида мавжуд бўлган “Ўзбек тили синонимлар луғати”, “Ўзбек тилининг ўзлашма сўзлар изоҳли луғати”, “Ўзбек тилининг эскирган сўзлар ўқув луғати”, “Ўзбек тилининг маънодош сўзлар ўқув луғати”, “Ўзбек тилининг шаклдош сўзлар ўқув луғати”, “Ўзбек тилининг зид маъноли сўзлар ўқув луғати”, “Ўзбек тилининг сўзлар даражаланиши ўқув луғати”, “Ўзбек тилининг ўқув этимологик луғати”, “Ўзбек тилининг ўқув топонимик луғати” лингвистик таъминот вазифасини бажара олади. Фақат бундай луғатлар қайта ишловдан ўтиши, сўзларни леммалаш, сўзларнинг хусусиятидан келиб чиққан ҳолда уларнинг қаторини чегаралаш ва леммаланган қатор аъзоларини бир-бири билан боғлаш амалга оширилиши лозим. Шундагина қайта ишланган луғат дастурчи учун дастурий таъминот асосини ташкил эта олади.

Разметкаларни лингвистик моделлаштириш мақсадга мувофиқ, чунки лингвистик моделда морфологик тег шартли қисқартма шаклини олади. Ҳар бир сўз туркумини разметкалаш учун махсус лингвистик модел шакллари ишлаб чиқилади. Лингвистик базани морфологик разметкалаш алгоритмини ишлаб чиқиш зарур. Лингвистик базани семантик разметкалашлар билан таъминлаш йўлларини аниқлаш керак. Миллий корпус яратишда ва унинг лингвистик базасини тузишда лингвистик разметкалашнинг аҳамияти жуда катта.

Бобнинг «*Сўз бирикмаси таҳлилида моделлардан фойдаланиш*» деб номланган тўртинчи фаслида грамматик қоидалар асосида тузилган ва турли сўз туркумлари билан ифодаланган мослашув ҳамда битишув муносабатли сўз бирикмаларининг моделларини тушунарлилигига эришиш, моделнинг қайси бирлик учун тегишли эканлигини, таркибидаги қисмларнинг морфологик шаклини ҳамда грамматик кўрсаткичларини фарқлай олиш учун сўз бирикмасининг модели келтириб ўтилган.

---

[23] Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (Расмий ва илмий услубдаги матнлар таҳрири дастури учун)**.**Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф. –Фарғона, 2019.

[24] Эшмўминов А.Ўзбек тили миллий корпусининг синоним сўзлар базаси. Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф. – Қарши, 2019.

[25] Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф. –Қарши, 2018.

[26] Абдурахмонова Н.З. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (Содда гаплар мисолида). Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф. – Тошкент, 2018.

Сўз бирикмаси компонентларининг қолипларини яратишда уларнинг биринчи ҳарфини асос сифатида қабул қиламиз. Яъни тобе сўз учун катта [Т], ҳоким сўз учун эса катта [ Ҳ ] ҳарфлари сўз бирикмасининг умумий модели сифатида танланади. Қисмларнинг ўзаро бир-бирига тобелик алоқасида боғланишини англатиш учун тобеланишнинг модели сифатида йўналишни кўрсатувчи [ $\Longrightarrow$ ] белгиси қабул қилинади. Ушбу моделнинг тобе сўздан ҳоким сўзга қараб йўналганлиги, ҳоким компонентнинг етакчилик, тобе компонентнинг эса бўйсунувчилик хусусияти билан боғлиқ. Айтилган фикрларнинг янада тушунарлироқ бўлиши учун сўз бирикмасининг моделларини жадвалларда бериб ўтамиз.

<table>
<tr><th>№</th><th>Сўз бирикмаси қисмлари</th><th>Моделлар</th><th>Жойлашиш ўрнига кўра сўз бирикмаси</th><th>Жойлашиш ўрнига кўра моделлар</th></tr>
<tr><td>1</td><td>Тобе сўз</td><td>Т</td><td rowspan="3">Тобе ⇨ ҳоким</td><td rowspan="3">Т $\Longrightarrow$ Н</td></tr>
<tr><td colspan="2">Тобеланиш</td><td>⇨</td></tr>
<tr><td>2</td><td>Ҳоким сўз</td><td>Н</td></tr>
</table>

Сўз бирикмаси таҳлилида моделлаштириш методини татбиқ қилиш бирикманинг синтактик турларини қолиплаш учун муҳимдир. Бу жараёнда анъанавий усулга таянган ҳолда модел сифатида ўша бирликларнинг биринчи ҳарфлари олинади. Масалан, мослашув муносабатли сўз бирикмаси [Мб]; битишув муносабатли сўз бирикмаси [Бб]; бошқарув муносабатли сўз бирикмаси эса грамматик воситаларнинг турига кўра гуруҳларга ажралади. Яъни келишик қўшимчалари билан боғланган бирикмалар; келишикли бошқарув муносабатли сўз бирикмаси [ККБб]; кўмакчилар билан боғланган сўз бирикмаси; кўмакчили бошқув муносабатли сўз бирикмаси эса [КўБб] белгилари билан қолипланади. Сўз бирикмаси қолиплари ҳам грамматик қоидалар ва грамматик кўрсаткичларнинг турига мувофиқ бир-биридан фарқли кўринишларга эга бўлади.

Бобнинг «**Қўшма сўз ясалишида лингвистик моделлардан фойдаланиш**» деб номланган бешинчи фаслида ўзбек тилшунослигида сўз ясалиши (деривация) масаласи, аффиксация усулининг 3 хил кўринишдаги сўз ясаш модели, композиция усулининг сўз туркумига боғлиқ бир неча кўринишли моделлари таклиф этилиб, мисоллар орқали далилланган.

“Ўзбек тили миллий корпуси”нинг лингвистик базасини яратишда ясама сўзларнинг моделларини тузиш аҳамиятлидир. Бу ўринда М.Абжалова таклиф этган лингвистик модуллардан фойдаланиб, аффиксация усулида 3 хил кўринишдаги қуйидаги сўз ясаш моделини таклиф этиш мумкин:

Бундан: А=асос, N=ҳосила сўз

1. N= асос + ли; N= асос + ла; N= асос + сиз; N= асос + лик; N= асос + чи; N= асос + хон; N= асос + дон.....

Аффиксация усули билан сўз ясашда аффикслар, одатда, асосдан кейин қўшилади. Шунга кўра бу усул билан ҳосил бўлган ясама сўзлар «асос + қўшимча (суффикс)»тарзида бўлади (*маза-ли, маза-сиз, мазлум-лик, сабзавот, сабзавот-чи, сабзавотчи-лик, супур-ги* каби).

2. N= бе+асос ; N= но+асос; N= ҳам+асос; N= бад+асос.....

Ясама сўзлар «олд қўшимча(префикс) + асос» шаклида ҳам бўлади. Бу ҳодиса асосан тожик тилидан ўзлашган аффикслар орқали сўз ясашда учрайди: *бе-лаззат, бе-ҳикмат, бе-ҳисоб, бе-ҳол,бе-ҳузур, бе-ҳужжат, сер-барака, сер-бар, сер-иштаҳа, сер-мазмун, но-инсоф, но-умид, но-қулай, но-муносиб* каби.

3. N= бе+асос +лик; N= но+асос +лик; N= ҳам+асос +лик; N= бад+асос +лик.....

Ясама сўз таркибида сўнг қўшимча (суффикс) ҳам, олд қўшимча (префикс) ҳам келиши мумкин: *бе-сабр-лик, бе-парво-лик, бе-саранжом-лик, бе-саришта-лик, ҳам-жиҳат-лик, ҳам-нафас-лик, но-инсоф-лик, но-мард-лик, но-маъқул-чилик, но-маҳрам-лик, но-аниқ-лик, бад-бахт-лик* сингари.

Айтиш мумкинки, моделлаштириш методидан тилшунослик соҳасида, хусусан, грамматик таҳлил жараёнида фойдаланиш бир қараганда кўпчиликнинг тасаввурида соддаликдан мураккабликка ўтишдек кўринса-да, ёритилаётган мавзунинг тушунарлилик ва аниқлик даражасини оширишга хизмат қилади.

Диссертациянинг **«Ўзбек тили миллий корпус менежерини яратиш технологияси»** деб номланган тўртинчи боби тўрт фаслдан иборат бўлиб, унда корпус-менежер тизими: корпус маълумотлари архитектураси ва модели, парсинг, корпус фрагменти интерфейсини шакллантириш алгоритми, ўзбек тили миллий корпусининг онлайн варианти фрагментининг тузилиши кабилар ҳақида фикр юритилган.

Менежер корпус ( корпус браузер ёки корпус сўров тизими ) кўп тилли корпус таҳлили учун восита бўлиб, одатда менежер корпус тил шакллари ва кетма-кетликлар учун қидирувда қўлланиладиган мураккаб тизим ҳисобланади. У матн хусусида маълумот бериши ёки қидирувчи томонидан маълумотлар хусусияти мавқеи жиҳатидан (лемма ва тег сингарилар тўғрисида) маълумот бериши мумкин. У «конкорданс» деб аталади. Бошқа амаллар коллокациялар қидирувини, частоталар статистикасини ва матнда қайта ишланадиган метамаълумотларни ўзида мужассамлаштиради. Менежер корпуснинг нисбатан қисқача мазмуни серверга ёки корпуснинг сўров двигателига оид бўлади. Бунда мижоз томонига оид бўлган жиҳатлари фойдаланувчи интерфрейслари дейилади. Менежер корпус шахсий компьютерда дастур сифатида таъминланиши ёки веб – сервис сифатида берилиши мумкин[27].

«Матнлар корпуси» тушунчасининг ажралмас қисми бу матнли ёки лингвистик маълумотларни бошқариш тизимидир. Сўнгги пайтларда уни кўпрок корпус менежер деб аташади. Корпусли менежер – бу ихтисослашган қидирув тизими. У корпусдаги маълумотларни қидириш дастурий воситалари, статистик маълумотларни олиш ва натижаларни фойдаланувчиларга қулай тарзда етказишни ўзида мужассамлаштиради.

[27] Suleymanov D., Nevzorova O., Gatiatullin A., Gilmullin R., Khakimov B. National corpus of the Tatar language “Tugan Tel”: grammatical annotation and implementation. Procedia-Social and Behavioral Sciences, 2013, vol. 95, pp. 68–74. DOI: 10.1016/j. sbspro.2013.10.623.

Масалан: Рус тили миллий корпуси бўйича қидирувда ишлар Yandex.Server Professional асосида амалга оширилади. Yandex.Server эса грамматик ва метатекстлардаги яширин хусусиятлар ва алоҳида лавҳалар маълумотларини топишга жалб этилган. Қидиришдаги берилиш Yandex.Server орқали шаклланади. У корпоратив тармокдаги веб-сервердаги рус тили морфологик хусусиятларини инобатга олиб маълумотларни тўлиқ матн асосида қидиришни таъминлайди. Қидирув рус, инглиз, украин тиллари морфологиясини ҳисобга олган ҳолда амалга оширилади. Шу билан бирга интернетда Яндекс бўйича иш юритади. Агар «идти» сўзи бўйича қидирув берилса, унда «идти», «идёт», «шел», «шла» сўзлари учрайдиган ҳужжатларда намоён бўлади. Релевантлиги бўйича тартиблаштирилган ҳужжатлар қидирув натижаси бўлади. Улар нафақат ҳужжатлар сонини, балки сўзларнинг асиллигини, уларнинг қўлланилиш частотаси ва сўзлар орасидаги масофаларни ҳам ҳисобга олади.

Сўровлар ўзининг предметли ва формал мазмуни бўйича таҳлил қилинади ва корпус билан ишлайдиган илмий атамалар луғатида изоҳланади. Қидирув корпуснинг алоҳида элементларини тартиб билан қиёслашдан ва уларнинг ўзаро мослигини аниқлашдан иборатдир. Бундай ҳолда корпус матнлари релевант ҳисобланади ва беришга тавсия этилади[28].

Ўзбек тили миллий корпусининг сўровлар тили модели умумий жиҳатдан қуйидаги элементларни ўзида мужассамлаштиради:

1) бевосита қидирув элементлари (атамалар ва информацион сўровлар);

2) матн сўров элементларини морфологик меъёрлаштириш воситалари;

3) операторлар (конъюнкция: кўпайтириш; дизъюнкция: қўшиш; инкор этиш);

4) линияли грамматика воситалари (масофа ва мавқе операторлари);

5) қидирувнинг қўшимча шароитлари:

-корпуснинг белгиланган майдонларидаги қидирув(масалан теглар ичида);

-қидирув соҳасини чеклаш (айрим муаллифлар асарлари, айрим ҳужжатлар ва улар турлари бўйича);

6) бериладиган натижалар бўйича саралаш (ранжировка) талаблари;

7)натижаларни бериш шакли ва турига нисбатан талаблар.

Ишлаб чиқишнинг биринчи босқичида излаш тизими учун зарур бўлган маълумотлар захирасини ва маълумотлар базасини бошқариш тизимини(МББТ) танлаш муҳим. МББТни ишлатиш имконлари ва маълумотлар захираси реал вақт режимида катта ҳажмга киришни тез ва ишончли қилади ҳамда қуйидаги мезонларга жавоб бериши лозим бўлади:

- самарадорлик (МБнинг ишлаш тезлиги 100 млн қаторли жадвални қўшиб ҳисоблаганда секундига 1 та сўров);

- масштаблилиги (бир неча машинага тақсимланган жараёнларга мувофиқ тизимнинг функционаллигига мос талаблар амали);

- тан нархи (таҳлил бепул коммерцияни ва маълумотлар захирасини ўзида мужассамлаштиради);

[28] Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Suchomel V. The Sketch Engine: ten years on. Lexicography, 2014, no. 1, pp. 7–36. DOI: 10.1093/ijl/ecw029.

- ПО билан биргаликдаги мувофиқлик (PHP ва Unix каби тизимлар билан ишлаш имкониятини қўллаб-қувватлаш );

- ҳужжатларнинг имконлиги (рус, инглиз ва татар тилидаги ҳужжатларнинг тўлик имконлиги);

- тараққиёт истиқболи (лойиҳани ишлаб чиқиш динамикаси, фойдаланувчилар жамоати, ишлаб чиқувчилар режалари);

МБ ва тизим архитектураси қуйидаги турдаги саволларга жавоб бериш учун ишлаб чиқилган:

- сўз шакли ёки лемма бўйича бевосита қидирув учун;

- ва, ёки, ё каби конъюнкция, дизъюнкция, инкор шакллари сифатида келтирилган кўринишлар морфологик хусусиятларини қайтар тарзда ўрганиш;

- сўз шакли ва леммада морфологик хусусиятларни гибрид излаш тури учун.

Корпус – менежери учун яратилган архитектурадан фойдаланиш кўплаб муаммоларни ҳал этишга имкон беради. Келажакда бу архитектурани лингвистик маълумотлар таҳлили, жумладан морфологик анализатор, кўп маъноли морфологик модуль ечими ва бошқа турли сервислар интеграциясида бемалол қўллаш мумкин.

Лингвистик корпуслар учун ишлаш тизимлари масаласини ечишдаги бу ёндошув ушбу махсус ишлаб чиқиладиган тизимни нафақат татар тили электрон корпуси иши учун, ўзбек тили корпуслари ўзгаришлардаги ечимларда ҳам қўллаш мумкин[29].

Бобнинг *«Корпусда матнларнинг синтактик анализи(парсинг)*» деб номланган иккинчи фаслида парсинг, унинг вазифалар, морфологик анализатор кабилар назарий жиҳатдан таҳлил қилинган.

Парсинг **(**Parsing**)** – бу синтактик таҳлилнинг компьютер фанидан таърифи. Бунинг учун дастурлаш тилларидан бири билан тавсифланган токенларни расмий грамматика билан таққослаш учун математик модел яратилади. Масалан, PHP, Perl, Ruby, Python. Бирор киши ўқиётганда, филология фани нуқтаи назаридан, қоғозда кўрган сўзларни (токенларни) ўзининг луғатидаги (расмий грамматикага) таққослаб, синтактик таҳлил қилади. Компьютерга *«*ўқиш» - таклиф қилинган сўзларни бутунжаҳон интернетдаги сўзлар билан солиштиришга имкон берадиган дастур (скрипт) синтактик деб номланади. Бундай дастурларнинг кўлами жуда кенг, аммо уларнинг барчаси деярли бир хил алгоритм асосида ишлайди. Парсинг билан ишлаш алгоритми қуйидагича: қайси расмий дастурлаш тилида ёзилган бўлишидан қатъий назар, унинг ишлаш алгоритми бир хил бўлиб қолади. Интернетга кириш, веб-ресурс кодига ўтиш (кириш) ва уни юклаб олиш; маълумотларни ўқиш, олиш ва қайта ишлаш; олинган маълумотларни қайта ишланадиган шаклда тақдим этиш - .txt, .sql, .xml, .html ва бошқа форматдаги файллар.

Парсинг қуйидаги вазифаларни ҳал этади:

1. Лексик таҳлил – матнни гаплар ва сўзларга ажратиш.

[29] Хакимов Б.Э., Гильмуллин Р.А., Гатауллин Р.Р. Разрешение грамматической многозначности в корпусе татарского языка // Ученые записки Казанского ун-та: Гуманитарные науки. 2014. Т. 156. № 5. С. 236–244.

2. Сўзларнинг морфологик таҳлили (токенизация ва лемматизация) – сўзлар матнини (маъно устуворлигини) инобатга олган ҳолда нутқ қисмини, келишигини, турини (родини) ва бошқа грамматик белгиларини аниқлаш.

3. Синтактик таҳлил (dependency parsing) – гаплардаги сўзлар боғланишини аниқлаш, эга ҳамда кесимни излаш, гапларнинг эга, тўлдирувчи ва ҳол бўйича гуруҳларга ажратилиши.

5. Соддалаштирилган синтактик таҳлил (chunking) – мураккаб матн таркибини ички абзацларга ажратиш.

Юқорида санаб ўтилганларнинг барчаси учун ва шу жумладан луғатда бўлмаган сўзлар учун ҳам улар бажарилади. Бундан ташқари морфологик анализатор орфографик хатоларни тузатиш режимини улаш ҳам мумкин бўлади.

Парсинг жуда катта ҳажмли матнларни (ўнлаб килобайт ёки юзлаб мегабайт) тезкор равишда таҳлил этиш учун лойиҳалаштирилган. Парсингнинг юқори самарадорлигига эришиш учун у ишга туширилаётганда сўз захирасининг ҳаммаси унинг оператив хотирасига киритади. Парсинг таҳлил этилаётган матн ҳажми ёки ундан фойдаланиш вақтига чеклов қўймайди.

Токенлар морфологик анализатор объектлари бўлиб, уларнинг вазифаси стемминг, лематизация ва морфологик таҳлилни амалга оширишдан иборат. Ўзбек тили ўз таркиби жиҳатидан агглютинатив тилларга оид бўлиб, у морфологик анализатор иш алгоритмига таъсир қилади. Таҳлил эса луғат ёндошуви асосида амалга оширилган. Унда ҳар бир сўз семаси аниқланган бўлиб, у ёки бу сўзнинг қандай парадигмага оидлиги аниқланади. Грамматик луғат ягона лексикографик академик тадқиқот ҳисобланган «Ўзбек тилининг изоҳли луғати»га асосланади.

Демак, синтактик алоқаларни ўрнатиш тадбирлари ва сўз ёки ибораларга маълум бир синтактик белгиларни тақаш амалини парсинг бажаради.

Бобнинг учинчи фасли *«Корпус интерфейсини шакллантириш босқичлари»* деб номланган. Унда интерфейс ва унинг турлари, дизайни таҳлил қилинган.

«Интерфейс» сўзи инглиз тилидан олинган бўлиб, «ташқи кўриниш» деган маъноларда ишлатилади. Ушбу сўз кўпинча компьютер технологиясида қўлланилади. Компьютер – инсон ва машина ўртасида турли хил ахборот алмашинувини таъминлайдиган ягона алоқа тизимидир. Интерфейс - бу битта тизимнинг иккита элементи ва ушбу тизим ёрдамида ишлайдиган боғловчи бўғин. Интерфейс турли хил тугунлар ва мураккаб ускуналар блоклари, шунингдек, технология ва фойдаланувчи ўртасидаги алоқа тизими ҳисобланади. У мантиқий (ахборот вакиллик тизими) ва формал (ахборот узатиш хоссалари) шаклида ифодаланади. Унинг ёрдамида муаян вазифалари учун буйруқлар берилади. Бундай интерфейс фойдаланувчи интерфейси деб номланади. Ҳар қандай қурилманинг интерфейси бажарадиган вазифаларига қараб ташқи ва ички кўринишларга бўлинади. Ички интерфейсга фойдаланувчи тўғридан-тўғри кириш ҳуқуқига эга бўлмайди, у хусусий имкониятга эга. Ташқи интерфейс билан фойдаланувчи бевосита алоқа қила олади ва унинг ёрдамида қурилмани

бошқариши мумкин. Ушбу икки турдаги интерфейс ҳар доим битта қурилмага киради ва унинг ишлашини таъминлайди, улар алоҳида мавжуд бўлолмайди. Фойдаланувчи интерфейсини 2 қисмга бўлиш мумкин. Чунончи, бу қурилмага маълумот киритиш учун жавобгар бўлган ва фойдаланувчига унинг чиқиши учун жавобгар бўлган қисм. Агар биз оддий иш компьютери ҳақида гапирадиган бўлсак, унда биринчи тоифада биз компьютерда ишлайдиган барча нарсалар мавжуд. Шунга кўра, ҳамма нарса иккинчи тоифага тегишли бўлиб, унинг ёрдамида компьютер фойдаланувчига маълумотни бир хил клавиатура, сичқонча ва бошқа кириш мосламалари, яъни мониторлар, карнайлар, наушниклар, принтерлар, плутерлар ва бошқалар орқали берилган буйруқларга жавоб бериб, узатади[30]. Компьютер технологиясида ишлатиладиган интерфейс қуйидаги турларда бўлади:

*Визуал.* Мониторда намойиш этиладиган визуал тасвирлар ёрдамида маълумотларни узатадиган стандарт компьютер интерфейси.

*Имо-ишора.* Қоида тариқасида, у телефонлар ёки планшетлар учун интерфейс бўлиб хизмат қилади. Кўпгина ҳолларда, бу тизимни бошқарадиган одамнинг бармоқларининг ҳаракатларига жавоб берадиган ва ҳар бир аниқ ҳаракатга маълум даражада жавоб берадиган сенсорли панел. Оддий визуал интерфейснинг соддалаштирилган версияси деб аташ мумкин.

*Овоз.* Ушбу турдаги интерфейс нисбатан яқинда пайдо бўлди. Овозли буйруқлар ёрдамида тизимни бошқариш имкониятини беради. Тизим, ўз навбатида, фойдаланувчи билан мулоқот орқали ҳам жавоб беради. Энг қизиғи шундаки, замонавий технологиялар бизга нафақат телефонлар ёки компьютерларнинг овозини, балки маиший техника ва ҳатто бортли компьютерларнинг овозини бошқаришга имкон беради.

Ушбу соҳадаги энг янги йўналишлардан бири бу сенсорли интерфейс. Унинг ишлаш принципи маълум объектлар орқали амалга ошириладиган фойдаланувчи ва машинанинг жисмоний ўзаро таъсирига асосланади.

Миллий корпуснинг интерфейси турли дизайн, тузилишга эга бўлиб, унинг мукаммаллиги корпус яратувчи муаллифнинг зиммасига юкланади. Чунки интерфейс корпус ҳақида илк таассурот қолдирувчи, ўзига жалб этувчи умумий кўринишдир. Интерфейс яратишда миллий коллоритни акс эттирувчи безаклар ҳамда мумтоз ёки замонавийликни акс эттирувчи белгилар эътиборга олиниши лозим.

Демак, миллий корпусдан фойдаланишда унинг қулай ва энг самарали имкониятларининг намоён бўлишида интерфейснинг мукаммал ҳамда тизимли равишда ишлаб чиқилганлиги муҳим саналади. Шу боис интерфейсни замонавий дастурий дизайн талабларига жавоб берадиган, фойдаланувга тушунарли, ишлаш учун қулай шаклда яратиш лозим.

Бобнинг тўртинчи фасли «*Ўзбек тили миллий корпусининг онлайн варианти фрагментининг тузилиши*» деб номланади. Ушбу фаслда онлайн варианти фрагменти намуна сифатида берилган. Корпуснинг қидирув

[30] Raxilina ye. V., Marushkina A. S. Corpus studies of speech features of non-standard speakers ("xeritajnыy russkiy") // Acta Linguistica Petropolitana. Trud instituta lingvisticheskix issledovaniy. 2015. –T. XI. –№ 1. – S. 621-639.

ойнаси(интерфейси)нинг «Корпусдан қидириш» – деб номланган тугмасида сўз ёки гапни қидириш имкони мавжуд. Унга «**осмон**» ёки «**ер**» ёки «**замон**» сўзини ёзиб қидирув сўрови берилса, унинг натижаси янги ойнада қуйидагича кўринишга эга бўлади:

«**осмон**» – 1. Ер устида гумбаз шаклида кўриниб турган мовий фазо; 2. Ер атрофини ўраб олган олам фазоси (астрономик макон);

Ёки:

«**ер**» – 1.Қуёшдан кейинги учинчи планета. 2. Шу планетанинг қуруқлик қисми (сув билан қопланган қисмига қарама-қарши қўйилганда). 3.Планетамиз қобиғининг сиртқи қатлами. 4. Бирор нарса банд қилиб турган, эгаллаган ўрин, жой, макон;

Ёки:

«**замон**» – 1. Д*авр, вақт*. Материя (объект) ҳолатларининг ҳамда ҳодиса (жараён)ларнинг изчил алмашиниш шакли; давомлилик, такрорланмаслик, қайтарилмаслик каби умумий хоссаларга эга бўлган вақт. 2. Сўз бораётган пайт; умуман вақт, пайт, давр, маҳал.

Шунингдек, лингвистик база таркибида киритилган матнлар ичида шу сўз мавжуд бўлса, у бошқа ойнада (алоҳида кўринишда) қидирув натижаси сифатида берилади.

Корпуснинг «Лексик-грамматик қидирув» тугмасидан сўзнинг маъноси, грамматик ҳолати, синонимлиги, омонимлиги, паронимлиги, антонимлиги, сўзнинг варианти, сўзнинг қўлланилиш даври, сўзнинг қўлланилиш услуби каби хусусиятларини билиб олиш мумкин. «Лексик-грамматик қидирув» тугмасига **«ингичка»** сўзини ёзиб қидирувга берилса, унинг натижаси янги ойнада қуйидагича кўринишга эга бўлади:

- **сўзнинг маъноси**: 1. Кўндаланг кесими меъёрдан кичик; 2. Чийилдоқ, ўткир (товуш ҳақида);
- **грамматик ҳолати**: [сифат];
- **синонимлиги:** майин;
- **омонимлиги**: йўқ;
- **паронимлиги**: йўқ;
- **антонимлиги:** йўғон;
- **варианти**:йўқ;
- **қўлланилиш даври:** замондош;
- **қўлланилиш услуби:** бетараф;

«Синтактик қидирув» деб номланган ойнада гапнинг мақсадига кўра турлари кўриниб туради: дарак, сўроқ, буйруқ Шунингдек, гапнинг тузилишига кўра тури шу қидирув тугмасида кўриниб туради: содда ва қўшма, уюшган гаплар. Масалан, «Синтактик қидирув» тугмасига «**дарак гап**» деб ёзиб қидирувга берилса, унинг натижаси янги ойнада қуйидагича кўринишга эга бўлади:

Бир тўда ўртоқларим, ҳар кундаги каби йиғилишиб, Шайхонтохурга кетишди (Ойбек);

Ой нурига терс ўтиргани учун унинг юзида кўзида қандоқ, ифода касб этаётганини билмасдим (С.Аҳмад);

Афзалхон акам анчагача чўккалаб ўтирдида-да, секин қаддини ростлади (Ў.Ҳошимов);

Кийимидан йигит кўпроқ студентга ўхшар эди (А.Қаҳҳор).

«Корпус нима?» тугмасида корпус тушунчаси ва ҳозирга қадар яратилган корпуслар ҳақидаги маълумотлар жойлаштирилган.

«Таҳлиллар» тугмасида сўзнинг ясалиши (200га яқин сўз),ўзлашма сўзлар (200дан ортиқ сўз), сўзнинг даражаланиши (150дан ортиқ сўз), иборалар (100дан ортиқ сўз) ҳақида маълумотлар берилган.

Хуллас, ўзбек тили миллий корпусининг умумий кўриниши бир неча ойналарга ҳамда ўнг ва чап устунчаларга ажратилган. Унда қуйидаги ойналар мавжуд бўлади: «Лексик қидирув», «Морфологик қидирув», «Синтактик қидирув. Қидирув тугмачалари сўз ва гапларни бир неча сония вақт ичида автоматик таҳлил қилиб беради.

## ХУЛОСА

1. Корпус лингвистикаси - тилшуносликнинг энг ривожланган соҳаси, корпус эса тилшуносларнинг зарурий иш қуроли; оғзаки, ёзма ёдгорликлар, миллий-маданий меросни акс эттирувчи ахборот манбаидир. Корпус-қидирув дастурига бўйсундирилган матнлар йиғиндиси, мукаммал разметкага эга корпус лингвистик тадқиқотлар самарадорлигини таъминлашда барқарор лингвистик база вазифасини бажаради. Лингвистик корпус сунъий интеллект маҳсули сифатида электрон луғат, таржима портали, терминологик маълумотлар банки, виртуал (электрон) кутубхона, электрон ҳукумат, электрон нашр, электрон дарслик ва қўлланмалар қаторида туради. Сунъий интеллект маҳсули бўлган лингвистик электрон манбалар муайян тил корпусини яратиш учун хом ашё саналади.

2. Ўзбек тилининг халқаро миқёсдаги мақомини ошириш, уни жаҳон мулоқот тили даражасига кўтариш, ўзбек тилини чет элларда ўрганиш ва ўргатиш, миллий тилимизнинг имкониятларини кенгайтириш ва сайқаллаш ишларини бевосита миллий корпус орқали амалга ошириш даркор. Миллий тилни йўқолиш хавфидан сақловчи, миллатнинг мавжудлигини жаҳонга танитувчи восита сифатида тил корпуслари бебаҳо хазинадир.

3. Электрон шаклда мавжуд бўлган матнларни, ҳужжатларни маълумотларни қайта ишлаш, уларни автоматик таҳлил, яъни морфологик синтактик ва семантик таҳлил қилиш, морфологик анализ ва синтез қилиш билан бирга ярим автоматик режимда маънони бузмасдан ишончли нутқ материалини мослаштириш даражасини текширишувчи мослама тил корпуси ҳисобланади. У матнни автоматик анализ, синтез шаклида таҳлил қила олиш хусусияти билан электрон кутубхонадан афзалроқдир.

4. Миллий корпус яратиш икки босқичда амалга оширилади: манбалар рўйхатини аниқлаш ва матнларни рақамлаштириш (компьютер шаклига ўтказиш). Унинг технологик жараёни қуйидагилардан иборат: танланган матнлар асосида лексема ва сўз шаклларининг такрорланиш луғатини яратиш; олинган такрорланиш луғатининг ҳар қандай бирлиги учун матнни кўриб чиқиш; графикли сўзни бўғинга ажратиш ва бўғинларнинг такрорланиш луғатини тузиш; сўз захираларини саралаш; бир вақтнинг ўзида

чекланмаган файлларни қайта ишлаш; ташқи белгиларга эга бўлган матнлар корпусларини яратиш; яратиладиган матнлар корпуслари ҳамда корпусга кирувчи алоҳида матнлар учун статистик маълумотларни ҳисоблаб чиқиш; дастлабки матнлар билан txt, doc и rtf форматда ишлаш, кодлаштиришни автоматик тарзда белгилаш кабилар.

5. Корпус маълумотларини кодлаш энг самарали стандартлар танланади. Унинг маълумотларини тақдим этиш матн разметкаси SGML/XML тили негизида бажарилади. Лексик маълумотни кодлашда HTML/XML қоидаларига мослаштирилади. Миллий корпус учун танланган матнлар турли хил манбалардан олинади ва ҳар хил форматларда ифодаланади: оддий матн, HTML, RTF, PDF.

6. Дунё миқёсида амалда бўлган тил корпуслари каби ўзбек миллий корпусида ҳам лингвистик ва экстралингвистик разметкалар маълумотлар ифодасининг ягона форматда яратилади. Морфологик ва синтактик разметканинг назарий асосларини академ грамматикага асосланиб, қайта кўриб чиқиш, семантик разметка теглари тизимини қисқартириш билан боғлиқ амалий аҳамиятга молик ишлар амалга оширилади. Корпусда разметканинг аҳамияти беқиёс, чунки корпусдан фойдаланиш имконининг кенг ёки торлиги корпуснинг разметкасига боғлиқ. Мукаммал разметка кенг имкониятли, универсал корпус гаровидир.

7. Маълумотларни сақлаш, янгилаш, қидириш ва етказиб беришни таъминловчи ахборот, дастурий таъминот вазифасини бажарувчи, автоматлаштирилган тизим бу - маълумотлар базасидир. У табиий тилдаги энг яхши ишлов берилган, тезкорлик ва аниқлик учун хизмат қиладиган асл матнлар она тилимизнинг сунъий интеллектини бойитадиган лингвистик база ҳисобланади. “Ўзбек тилининг изоҳли луғати”, шунингдек ўзбек тилида яратилган турли услубдаги асарлар ўзбек тили миллий корпусининг лингвистик таъминоти вазифасини бажара олади.

8. Ўзбек тили миллий корпуснинг лингвистик материали сифатида замонавий лексикографиянинг илғор соҳаси киберлексикография бўла олади.

9. Корпусли менежер – бу ихтисослашган қидирув тизими бўлиб, корпусдаги маълумотларни қидириш дастурий воситалари, статистик маълумотларни олиш ва натижаларни фойдаланувчиларга қулай тарзда етказишни ўзида мужассамлаштиради. Миллий коруснинг маълумотлар базаси ва тизим архитектурасида бевосита қидирув ва гибрид излаш тури ишлаб чиқилади.

10. Замонавий дастурий дизайн талабларига жавоб берадиган, фойдаланувга тушунарли, ишлаш учун қулай шаклдаги алоқа тизими интерфейс ҳисобланади. У визуал, имо-ишоравий, овозли каби турлардан ҳамда тизимли ва амалий дастурлаш интерфейсидан ташкил топган.

**НАУЧНЫЙ СОВЕТ DSc.03/04.06.2021.FIL.72.09**
**ПО ПРИСУЖДЕНИЮ УЧЕНЫХ СТЕПЕНЕЙ ПРИ**
**БУХАРСКОМ ГОСУДАРСТВЕННОМ УНИВЕРСИТЕТЕ**

**БУХАРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**ТОИРОВА ГУЛИ ИБРАГИМОВНА**

# ТЕОРЕТИЧЕСКИЕ И ПРАКТИЧЕСКИЕ ВОПРОСЫ СОЗДАНИЯ НАЦИОНАЛЬНОГО КОРПУСА УЗБЕКСКОГО ЯЗЫКА

**10.00.01 – Узбекский язык**

**АВТОРЕФЕРАТ**
**диссертации доктора филологических наук (DSc)**

**Бухара – 2021**

**Тема диссертации доктора филологических наук зарегистрирована в аттестационной комиссии при Кабинете Министров Республики Узбекистан за B2019.3.DSc/Fil180.**

Диссертация выполнена в Бухарском государственном университете.

Автореферат диссертации на трёх языках (узбекском, русском, английском (резюме)) размещён на веб-странице Бухарского государственного университета (www.buxdu.uz) и информационно-образовательном портале «ZiyoNet» (www.ziyonet.uz).

| | |
|---|---|
| **Научный руководитель** | **Менглиев Бахтиёр Ражабович**<br>доктор филологических наук,профессор |
| **Официальные оппоненты:** | **Мухаммедова Саодат Худойбердиевна**<br>доктор филологических наук, профессор |
| | **Каримов Суюн Амирович**<br>доктор филологических наук, профессор |
| | **Хакимов Мухаммадхон Ходжаханович**<br>доктор филологических наук, профессор |
| **Ведущая организация:** | **Термезский Государственный Университет** |

Защита диссертации состоится «___» _____ 2021 года в _____ часов на заседании Научного совета DSc.03/04.06.2021.FIL.72.09 при Бухарском государственном университете. (Адрес: 200118, город Бухара, улица М.Икбола, дом-11. Тел.: +99865 221-29-14; факс: +99865 221-27-07, e-mail: buxdu_rektor@buxdu.uz).

С диссертацией можно ознакомиться в Информационно-ресурсном центре Бухарского государственного университета (зарегистрирована под номером _____). (Адрес: 200118, город Бухара, улица М.Икбола, дом-11. Тел.: +99865 221-29-14).

Автореферат диссертации разослан «____» _____2021 года.

(Реестр протокола рассылки №____ от «_____»_____2021 года).

**Д.З. Ражабов**
Заместитель председателя Научного совета по присуждению учёных степеней, доктор филологических наук (DSc), доцент

**Х.П.Эшонкулов**
Научный секретарь Научного совета совета по присуждению учёных степеней, доктор филологических наук (DSc), доцент

**Н.Г.Муродов**
Председатель Научного семинара при Научном совете по присуждению учёных степеней, доктор филологических наук (DSc), профессор

## ВВЕДЕНИЕ (аннотация диссертации доктора философии (DSc))

**Значимость и актуальность темы диссертации.** В мировой лингвистике ко второму десятилетию двадцать первого века создание языковых корпусов в Интернете является основным средством поддержания определенного языка, расширения области его исследований и демонстрации языковых навыков. В частности, компьютерные технологии, являющиеся великим изобретением двадцатого века, открывают двери широкому спектру возможностей для лингвистики, а также для других областей и ставят важные задачи перед компьютерным языком, появление компьютерной лингвистики имеет решающее значение для успеха естественных языков.

В глобальных языковых исследованиях изучение лингвистического моделирования языка, разработка алгоритмов лемминга слов и тегов, а также электронное использование устных и письменных памятников, образцов духовного наследия, созданных на определенном языке, с целью увеличения использования национального и культурного наследия. Особое внимание уделяется обработке информации с помощью компьютерных технологий, разработке необходимого методического и программного обеспечения для внедрения информационных ресурсов, развития языкового корпуса в Интернете и, исходя из этого, научным и теоретическим аспектам национального языка.

В узбекском языкознании проводились различные исследования по автоматическому переводу, развитию лингвистических основ авторского корпуса, обработке лексикографических текстов и лингвостатистическому анализу. Особый акцент был сделан на «совершенствовании системы образования и увеличении возможностей предоставления качественных образовательных услуг». Учитывая, что повышение международного статуса узбекского языка до уровня мирового языка общения, изучения и преподавание узбекского языка за границей, расширение возможностей и полировка нашего национального языка может быть достигнуто непосредственно через национальный корпус «Теоретические и практические вопросы узбекского национального корпуса». В этом смысле необходимо дальнейшее углубление исследований лингвистических основ текстового корпуса и национального корпуса, технологии создания его программного обеспечения.

Данное диссертационное исследование в определенной степени служит выполнению задач, предусмотренных в Указах Президента Республики Узбекистан за № УП-4997 «О создании Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои» от 13 мая 2016 года, № УП-4947 «О Стратегии дальнейшего развития Республики Узбекистан» от 7 февраля 2017 года, № УП-5850 «О мерах по кардинальному повышению престижа и статуса узбекского языка как государственного» от 21 октября 2019 г., в Постановлении Президента Республики Узбекистан за № ПП-2789 «О мерах по дальнейшему совершенствованию деятельности Академии наук, организации, управления и финансирования научно-

исследовательской деятельности» от 17 февраля 2017 г., в Постановлениях Кабинета Министров Республики Узбекистан № ПКМ-984 «Об утверждении положения о Департаменте развития государственного языка» от 12 декабря 2019 года, № ПКМ-40 «О мерах по организации деятельности комиссии по терминам при кабинете министров Республики Узбекистан» от 29 января 2020 года, а также в других нормативно-правовых документах, принятых в данной сфере.

**Соответствие диссертационного исследования приоритетным направлениям развития науки и технологий республики.** Диссертация выполнена в соответствии с приоритетным направлением развития науки и технологий Республики Узбекистан: I. «Формирование и реализация системы инновационных идей в социальном, правовом, экономическом, культурном, духовно-образовательном развитии информатизированного общества и демократического государства».

**Обзор зарубежных исследований по теме диссертации[31].** Исследования по формализации естественных языков и созданию компьютерных моделей их процессоров на основе математического аппарата, изучение корпуса и его типов проводится ведущими мировыми исследовательскими центрами и университетами, в том числе: Принстонский университет (США); Библиографический институт (Германия); Языковой центр Оксфордского университета (Великобритания); Университет Монпелье (Франция); Университет Упсалы (Швеция); Карлов университет (Прага); Институт лингвистических исследований (РАН), Компьютерные лингвистические лаборатории, МГУ им. М.В.Ломоносова (РФ); Южно-Казахстанский государственный университет им. Ауэзова (Казахстан), а также Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои, Бухарский государственный университет, Каршинский государственный университет, Научно-исследовательский институт языка, литературы и фольклора Российской академии наук (Узбекистан).

Изучение создания корпусов в мировой компьютерной лингвистике началось задолго до того, как появилась наука корпусной лингвистики. Примерами этого являются библейские исследования восемнадцатого века (например, Cruden), словари (Johnson, Oxford English Dictionary, Webster Dictionary), языковая подготовка (частотный корпус Thorndike, 1921) и Quirk Corpus (Survey of English Usage). Первое корпусное строение – это Brown Corpus, построенное в 1960-х годах в Университете Брауна в США. Основываясь на принципах Брауна, Проект Банка Англии и Британский национальный корпус учредили Upsal Corpus (Университет Упсалы, Швеция) (BNC). На основе британских идеалов был создан национальный корпус многих европейских языков (испанского, итальянского, хорватского).

[31] Обзор зарубежных исследований по теме диссертации www.university ofcalifornia.edu, www.harvard.edu, www.indiana.edu, www.uni-bonn.de, www.krugosvet.ru.valentnost, www.scicenter.syntagmatika, www.philol.msu.ru/, https://iling.spb.ru/grammatikon, www.princeton.edu, https://bigenc.ru/linguistics/text/, www.navoiy-uni.uz и др. источников.

Проведено исследование Ноама Хомского по формализации естественных языков и построению компьютерных моделей их процессоров на основе математического аппарата. Шарлотта Тейлор заметила существование общего и особого типов корпусов, что конкретные типы корпуса различаются по жанрам, стилю и периоду, и что обе категории корпусов существуют в диахронической и синхронной формах.

Сегодня в мировой лингвистике проводится ряд исследований по созданию национального корпуса, в том числе по следующим приоритетным направлениям: Польский национальный корпус, Алма-Атинский корпус казахского языка, Восточно-армянский корпус, Таджикский национальный корпус и другие. Чарльз Меер (Cambridge University Press, 2004) предполагает, что методологический подход к конкретным мелкомасштабным исследованиям в корпусной лингвистике также можно назвать корпусом. Дуглс Бибер (Оксфордский университет, 2015) утверждает, что аналитическая технология корпуса должна включать количественные (статистические) и качественные характеристики. М.З.Курди (Великобритания, 2016) делит тематический охват корпусов на сбалансированные типы корпусов на уровне возможностей в соответствии с отраслевой классификацией текстов.

**Степень исследованности вопроса.** Анализ таких тем, как создание национального корпуса в мировой лингвистике, его аналитическая техника и предмет корпусной лингвистики отображены в трудах следующих авторов Г.Луч, Р.Пятровский, Л.Блумфилд, Ч.К.Фрэнсиса, Ч.Ф.Меера, Г.Г.Кучера, Дж.Синклера и М.З.Курди и заслуживают особого внимания[32].

В русской лингвистике проводили целевые исследования в области корпуса В.Г.Бритвин, В.П.Захаров, И.А.Мельчук, А.Б.Кутузов, Р.Г.Котов, Л.И.Беляева, В.А.Чияковский, Е.Недошивина, В.В.Рыков, В.Плунгян изучали большой набор массивных текстов, принципы формирования корпуса, лингвистической базы данных[33].

---

[32] Chomsky N., The logical basis for linguistic theory, Proc. IXth Int. Cong, of Linguists, 1962; Leech G. The State of Art in Corpus Linguistics // English Corpus Linguistics / Aimer K., Altenberg K. (eds.) – London, 1991. – P. 8–29; Блумфилд Л. Язык. – М.: «Прогресс», 1968. – 608 с.; Fries Ch.C. The structure of English. An introduction to the construction of English sentences. – L., 1969; Bongers H. The history and principles of Vocabulary control. – Woerden: WOCOPI, 1947; Френсис Н., Кучера Г. Вычислительный анализ современного американского варианта английского языка. – М., 1967; Синклер Д. Предисловие к книге «Как использовать корпуса в преподавании иностранного языка» / Д.Синклер [Электронный ресурс]. – Режим доступа: http://www/ruscorpora.ru/corpora-infro.html, свободный; Charlez Meyer English corpus linguistics: An introduction. Cambridge University Press, 2004. 168 p.; Mohamed Zakaria Kurdi. Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax, Great Britain, USA: Wiley-ISTE, 2016, 300 p.

[33] Бритвин В.Г. Прикладное моделирование синтагматической семантики научно-технического текста (на примере автоматического индексирования). КД. – М.: МГУ, 1983; Мельчук И.А. Порядок слов при автоматическом синтезе русского слова (предварительные сообщения) // Научно-техническая информация. 1985, № 12. – С. 12–36; Захаров В.П. Корпусная лингвистика: учебник для студентов гуманитарных вузов. – Иркутск, 2011. – 161 с.; Кутузов А.Б. Корпусная лингвистика. – [Электрон ресурс]: Лицензия Creative commons Attribution Share-Alike 3.0 Unported [Электрон ресурс] // lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf; Котов Р.Г. Лингвистические аспекты автоматизированных систем управления. – Москва: Наука, 1977; Беляева Л.И., Чижаковский В.А. Тезаурус в системах автоматической переработки текста. – Кишинев, 1983; Недошивина Е.В. Программы для работы с корпусами текстов: обзор основных корпусных менежеров. Учебно-методическое пособие. – Санкт-Петербург. − 2006. 26 с.; Рыков В.В. Курс лекций по корпусной лингвистике. URL: http://rykov-cl.narod.ru/c.html; Плунгян В. Зачем мы делаем

Интересны также исследования авторов Х.Исхакова, С.Мухамедов, С.Риза о лингвистико-статистическом анализе текста в узбекском языкознании, лексикографической обработке, лингвистическом обеспечении программы автоматического редактирования, лингвистических модулях редакционно-аналитической программы, синонимичной лексике национального корпуса, лингвистические основы авторского корпуса. Работы С.Мухамедовой, Б.Менглиева, Д.Уринбаевой, А.Норова, А.Пулатова, Ю.Дысимовой, Г.Валиевой, Г.Джуманазаровой, Н.Абдурахмановой, Ш.Хамроевой, М.Абджаловой, А.Эшмуминова заслуживают внимания[34].

При написании диссертации были учтены имена и научные исследования ряда других узбекских и мировых лингвистов. В отличие от работы, проделанной в этом направлении, в нашем исследовании излагаются теоретические и практические вопросы создания национального корпуса узбекского языка.

**Связь диссертационного исследования с планами научно-исследовательских работ высшего образовательного учреждения, где выполнена диссертация.** Диссертация выполнена в рамках «Концепции развития узбекского языка и совершенствования языковой политики» на 2020–2030 годы и научно-исследовательского плана Бухарского государственного университета на 2017–2021 годы по теме: «Проблемы изучения взаимосвязи языка, личности и общество в узбекском языкознании» (2017–2021).

**Цель исследования** – разработать теоретические и практические основы формирования лингвистической базы национального корпуса узбекского языка.

**Задачи исследования**:

---

Национальный корпус русского языка. [Электрон ресурс] «Отечественные записки» 2005, № 2. http://magazines. russ.ru/oz/ 2005/2/2005_2_20-pr.html

[34] Исхакова Х.Ф. Исследования в области формальной морфологии тюркских языков (на материале татарского литературного языка в сопоставлении с турецким и узбекским). Канд. Дис. … филол. наук. – М., 1972; Муҳаммедов С.А. Статистический анализ лексико-морфологической структуры узбекских газетных текстов: Автореф. Дисс. … канд. фил. наук. – Ташкент, 1980; Ризаев С. Ўзбек тилининг лингвостатистик тадқиқи: Фил. фан. док. Дис. … автореф. – Тошкент, 2008; Муҳаммедова С. Ҳаракат феъллари асосида компьютер дастурлари учун лингвистик таъмин яратиш. Методик қўлланма. – Тошкент, 2006; Ўринбоева Д.Б. Ўзбек фольклори матнларининг лингвостатистик тадқиқи. – Тошкент: Фан, 2010; Пўлатов А. Компьютер лингвистикаси. – Тошкент: Akademnashr, 2011. – 500 б; Дысимова У. Матндаги феълларни автоматик таҳрир қилувчи дастурнинг лингвистик таъмини (расмий-идоравий услубдаги матнлар асосида). Магистрлик дисс. – Тошкент, 2002, 56 б.; Валиева Г. Расмий-идоравий услубнинг лисоний бирликларини моделлаштириш. Магистрлик диссер. –Тошкент, 2003, 60 б.; Норов А. Компьютер лингвистикаси асослари. – Қарши, 2017. – 136 б.; Жуманазарова Г.У. Фозил Йўлдош ўғли достонлари тилининг лингвопоэтикаси: Фил. фан. док. Дис. … автореф. – Тошкент, 2017; Менглиев Б. Ўзбек тили миллий корпуси. 2018 йил, 26 апрель, http://marifat.uz/marifat/ruknlar/fan/1241.htm; Абдурахмонова Н.З. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (Содда гаплар мисолида). Филол. фан. бўйича фалсафа доктори (PhD) … дис. автореф. – Тошкент, 2018; Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филол. фан. бўйича фалсафа доктори (PhD) … дис. автореф. – Қарши, 2018; Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (Расмий ва илмий услубдаги матнлар таҳрири дастури учун). Филол. фан. бўйича фалсафа доктори (PhD) … дис. автореф. – Фарғона, 2019; Эшмўминов А. Ўзбек тили миллий корпусининг синоним сўзлар базаси. Филол. фан. бўйича фалсафа доктори (PhD) … дис. автореф. – Қарши, 2019.

изучить опыт созданных мировые корпуса, включая Русский национальный корпус, Таджикский национальный корпус, Американский национальный корпус, и обосновать общие принципы формирования Узбекского национального корпуса;

описать процесс подготовки текстов кейса, обосновать способы моделирования существующих стандартов и технологий тегирования;

выделение схожих аспектов национального корпуса и электронных ресурсов, таких как подключение к компьютеру и Интернету, копирование необходимой информации на диск или бумагу, а также автоматический анализ слов, их заполнение, исправление, редактирование, наличие поисковой системы;

формирование базы данных национального корпуса и определение типа архитектуры системы, например, гибридный поиск по словоформе или лемме;

оцифровка устных и письменных рукописей, созданных на узбекском языке, всех образцов научного, теоретического, практического и духовного наследия, созданного на узбекском языке, с целью расширения использования национального и культурного наследия;

сохранение языка и расширение области его изучения, обосновать – основным средством демонстрации возможностей языка является создание языкового корпуса в системе Интернет.

**Объектом исследования** выбран национальный корпус узбекского языка.

**Предметом исследования** явились лингвистическая база, интерфейс корпуса, единицы корпуса, моделирование макета корпуса, леммы и теги при создании национального корпуса.

**Методы исследования.** В процессе исследования использовались рационально-типологический, сравнительный, содержательный, дискурсивный методы анализа.

**Научная новизна исследования:**

теоретически обоснована технология создания национального корпуса узбекского языка на принципах грамматического описания лексических единиц узбекского языка;

на основе теоретических выводов обоснованы существующие стандарты маркировки национального корпуса на основе языка SGML / XML, простые текстовые, заглавные и поэтические образцы фрагментов, регламентирующие траектории моделирования;

формат разметки текстов, требования к кодированию лексической информации и возможность использования лингвистических моделей в словообразовании, составное словообразование, дополняющие лингвистическую базу национального корпуса узбекского языка;

в процессе подготовки текстов, дополняющих лингвистическую основу национального корпуса узбекского языка: формирование системы управления корпусом и интерфейса корпуса посредством таких этапов, как предварительная обработка текста, его разметка, предоставление доступа к

корпусу, их быстрый многопараметрический поиск и статистическая обработка;

в связи с созданием национального корпуса – небольшой онлайн-проект, основанный на анализе обработки информации с помощью компьютерных технологий, создающий высокоэффективную виртуальную среду образовательного процесса.

**Практические результаты исследования:**

разработан теоретический аспект дизайна интерфейса национального корпуса;

раскрыты теоретические и практические вопросы морфологической, семантической, синтаксической разметки и лингвистического моделирования слов в узбекском языке («Толковый словарь узбекского языка»);

созданы национальный корпус-менеджер, база данных и проект системной архитектуры;

синтаксическое обозначение, меры по установлению синтаксических связей и практика присоединения определенных синтаксических символов к словам или фразам;

корпус основан на необходимости для лингвистов изучать вопросы как необходимый инструмент для лингвистов, устные и письменные памятники, источник информации, отражающий национальное и культурное наследие, средство защиты национального языка от исчезновения, средство внедрения нация в мире.

**Достоверность результатов исследования** обосновывается использованием теоретических методов и подходов, правильностью проведенных исследований с методологической точки зрения, основываясь на устную и письменную рукопись на узбекском языке, создание национального корпуса узбекского языка в электронном виде всех образцов научного, теоретического, практического и духовного наследия, созданного на этом языке, а также исторический, описательный, лингвокультурологический корпуса, основанные на методах компонентного анализа, внедрения теоретических идей и выводов в практику, полученные результаты подтверждены компетентными органами.

**Научная и практическая значимость результатов исследования.** Научная значимость результатов исследования определяется тем, что теоретические выводы исследования развития узбекского языка как электронной базы данных в области компьютерной лингвистики могут быть использованы в качестве источника в лингвистике.

Практическая значимость результатов исследования в создании лаборатории «Компьютерная лингвистика» на основе теоретических обобщений и анализа, преподавании специальных курсов, таких как «Компьютерная лингвистика», «Машинный перевод», «Корпоративное переводческое дело» способствующие повышению международного статуса узбекского языка до уровня мирового языка. Лабораторию «Компьютерная лингвистика» можно использовать для изучения и преподавания узбекского

языка за рубежом, для расширения возможностей нашего национального языка и для создания национального корпуса посредством отбеливания.

**Внедрение результатов исследований.**

На основании научных результатов, полученных в процессе определения теоретических и практических аспектов национального корпуса узбекского языка достигнуто повышение международного статуса узбекского языка, доведение его до уровня мирового языка общения на основе создания высокоэффективной виртуальной среды образовательного процесса в связи с созданием национальной сети электронного обучения и сделаны следующие выводы, такие как:

на основе теоретических выводов обоснованы существующие стандарты маркировки национального корпуса на основе языка SGML / XML, внедрение дистанционного электронного обучения использовалось в фундаментальном проекте № А5-037 «Разработка системы электронного дистанционного обучения для профессиональных колледжей в сфере ИКТ» (2015–2017) (справка № 89-03-763 Министерство высшего и среднего специального образования Республики Узбекистан от 9 февраля 2021 г.). В результате были сделаны выводы о повышении потенциала качественных образовательных услуг и создании необходимого программного и методического обеспечения внедрения информационных ресурсов;

в процессе подготовки текстов, дополняющих лингвистическую основу национального корпуса узбекского языка: формирование системы управления корпусом и интерфейса корпуса посредством таких этапов, как предварительная обработка текста, способы моделирования, использованные в фундаментальном проекте ОТ-F1-002 «Психологические механизмы формирования национальные идеи и идеологический иммунитет молодежи» (2017–2020) (справка № 89-03-763 Министерства высшего и среднего специального образования Республики Узбекистан от 9 февраля 2021 г.). В результате, в целях расширения использования национального и культурного наследия устные и письменные памятники, созданные на определенном языке, служили для выявления особенностей электронизации образцов духовного наследия;

были сделаны теоретические выводы, формат разметки текстов, требования к кодированию лексической информации и возможность использования лингвистических моделей в словообразовании, составное словообразование, дополняющих лингвистическую базу Узбекского национального корпуса F1-FA-0-13229 и использован в фундаментальном проекте «Функциональное словообразование в современном каракалпакском языке» (2012–2016) (справка № 292/1 Каракалпакского отделения Российской академии наук от 21 января 2021 г.). В результате проект обогатился новой научно-теоретической информацией;

монография «Теоретические и практические вопросы создания национального корпуса узбекского языка», основанная на теории и практике технологии создания национального корпуса узбекского языка, была использована в лекционных и практических занятиях по теме «Введение в

корпусную лингвистику» по № 5A120102 – Языкознание (узбекское языкознание) и справка № 89-03-763 Министерства среднего специального образования от 9 февраля 2021 г.). В результате он служил для описания процесса подготовки текстов для корпуса, для обоснования способов моделирования существующих стандартов и технологий тегирования;

система высшего образования на основе терминов корпусной лингвистики и смежных областей 5A120102 – «Словарь терминов корпусной лингвистики» (ISBN 978-620-0-61316-5) для специалистов в области лингвистики (узбекского языкознания) (справка № 89-03-763 Высшего и среднее специального образования Республики Узбекистан от 9 февраля 2021 г.). В результате на основе собранных материалов был создан словарь, содержащий 257 слов для создания национального корпуса узбекского языка и преподавания курса «Корпусная лингвистика» и служащий для пополнения фонда узбекской лексикографии;

такие выводы, как критерии создания банка текстов на узбекском языке и грамматическое описание лексических единиц, были использованы в учебнике «Узбекский язык» для учащихся № 5120100 – Филология и преподавание языка (русский) (справка № 89-03-763 Министерство высшего и среднего специального образования Республика Узбекистан 9 февраля 2021 г.). В результате была разработана теоретическая база для создания национального корпуса узбекского языка;

устные и письменные памятники, созданные на узбекском языке, компьютеризация образцов духовного наследия, обработка информации компьютерными технологиями, машинный перевод, разработка электронной лексикографии, создание тезаурусов, создание языкового корпуса в Интернете. окружающая среда», «Актуальная тема» (Сценарий № 1-450 Национальной телерадиокомпании Узбекистана от 11 декабря 2020 г.). В результате информация об устных и письменных памятниках, созданных на этом языке, электронизация духовного наследия, создание корпуса национального языка, превращение узбекского языка в язык «понятный» в Интернете, создали основу поразмышлять над темой для радиокоманды.

**Апробация результатов исследования.** Результаты исследования были обсуждены на 20 конференциях, в том числе на 8 международных и 12 национальных научных конференциях.

**Публикация результатов исследования.** По теме диссертации опубликовано всего 44 научных работ, в том числе 12 статей в научных изданиях, рекомендованных Высшей аттестационной комиссией Республики Узбекистан для публикации основных научных результатов докторских диссертаций (8 национальных, 6 зарубежных), в том числе 2 научных статьи в международных изданиях и 4 научные статьи в журналах на базе Scopus. Результаты представлены в 1 учебнике и 1 учебном пособии, 1 словаре, 1 монографии и 4 методических пособиях.

**Структура и объем диссертации.** Диссертация состоит из введения, четырех основных глав, заключения, списка использованной литературы и приложений, общий объем работы составляет 251 страниц.

## ОСНОВНОЕ СОДЕРЖАНИЕ ДИСЕРТАЦИИ

**Вводная глава** основана на актуальности и необходимости темы диссертации, описывает цели и задачи, объект и предмет исследования, его актуальность для научных и технологических приоритетов, описывает научную новизну и практические результаты исследования, а также раскрывает научные и Практическая значимость результатов исследования. информация о введении, опубликованных работах и структуре диссертации.

**Первая глава диссертации**, озаглавленная «Национальный корпус - как электронный лингвистический источник узбекского языка», трансформация узбекского языка в Интернет и электронный язык, необходимость совершенствования электронных ресурсов национального языка (узбекский корпус, электронные словари, тексты на сайтах). , искусственный интеллект и электронные ресурсы, понятие корпуса и его отличие от электронной библиотеки, анализ общих и различных аспектов корпуса, созданного в мире национального корпуса. Первая глава этой главы, озаглавленная «Искусственный интеллект и электронные ресурсы», показывает, что современные информационные технологии открыли дверь к широкому спектру удобств в использовании языка с помощью искусственного интеллекта, что он может выполнять многие функции, которые человеческий разум может выполнить. Было высказано предположение, что он создан, чтобы облегчить бремя людей. Развитие компьютерных технологий привело к созданию электронных ресурсов, таких как электронные словари, порталы переводов, банк терминологических данных, виртуальная (электронная) библиотека, электронный корпус текстов, электронное правительство, электронные публикации, электронные учебники и руководства. Искусственный интеллект состоит из алгоритмов и программных систем, предназначенных для выполнения множества задач, и он может выполнять ряд задач, которые может выполнять человеческий разум.

Если мы сравним систему человеческого мышления и искусственный интеллект, мы можем сделать вывод, что человеческое мышление обладает преимуществами творческого, гибкого, способного использовать эмоциональное восприятие, используя всеобъемлющие, всеобъемлющие знания. Имеет следующие недостатки: сложное кондуктивное (выразительное), неспособное быстро документировать, нестабильное мышление человека. Система поиска информации имеет преимущества последовательности, простоты представления и единообразия, легкого документирования. У него также есть ряд недостатков, таких как то, что он искусственный, ограниченный, заранее запрограммированный, разумеется, с использованием символического восприятия и специальных знаний. Путем анализа преимуществ и недостатков обеих систем в работе доказано, что основные преимущества человеческого мышления, в том числе во многих областях, таких как творчество, изобретательность, передача информации и контент в целом, превосходят искусственный интеллект.

Во второй главе тома «Анализ корпусных терминов в лингвистике корпусов» корпус представляет собой полуавтоматический процесс обработки существующих текстов, документов и данных в электронной форме, включая их автоматический анализ, т. Е. Морфологический, синтаксический и др. семантический анализ, морфологический анализ и синтез. Подчеркивается, что режим - это устройство, проверяющее уровень адаптации достоверного речевого материала без искажения смысла. Лингвистический корпус - это не только представление доступной информации в виде текста, но и анализ текста, который по своей аналитической способности считается лучше электронной библиотеки.

Отличия электронной библиотеки от языкового корпуса: единицей поиска электронной библиотеки является текст всего произведения. В нем можно искать конкретное произведение. Источником, обеспечивающим автоматический поиск информации в Интернете, является электронная или виртуальная библиотека. Электронная библиотека носит другое название. Например: виртуальная библиотека, электронная библиотека, электронная библиотека, электронная библиотека. В такой библиотеке книги, журналы и газеты будут размещаться в памяти компьютера, а не на книжных полках. Он поставляется в виде набора данных, хранящихся в цифровом формате на компьютере или специальном устройстве на компьютере. Такие данные могут включать в себя печатные, аудио-, видео- и мультимедийные данные. Нет необходимости в специальном месте для хранения книг, так как на сайте собрана различная информация электронной библиотеки. На эту страницу регулярно заходят, собирают и заполняют специалисты специального центра в библиотеках. Единица поиска корпуса может быть в виде языковой единицы и речевой единицы. Такой текст можно использовать не только для чтения, но и из-за наличия различных грамматических интерпретаций этих текстов, над ними можно производить лингвистические операции. Он отличается от тезауруса тем, что в тезаурусе поиск основан на понятии, а в корпусе ищется слово и его использование. Корпус важен как лексикографическая единица, содержащая различные словари (частотные, топонимы, грамматические слова, словосочетания и т. Д.). Корпус имеет большое значение в современной лексикографии. Поэтому он служит ресурсом для составления словарей большого объема. Со временем корпус становится большим (обширным) информационным ресурсом, становясь важным для множества лингвистических направлений. Корпоративные словари создаются и обрабатываются быстрее, чем когда-либо прежде. Тексты, доступные в операционной системе корпуса, имеют функцию сортировки. Исследователь сможет отличить нужный для себя пример от всех текстов, а не только от того, который важен для исследования. Электронная библиотека не имеет перечисленных функций.

*Третья глава тома «О лингвистическом статусе национального корпуса в узбекской лингвистике»* представляет собой научную основу корпуса как средства защиты национального языка от исчезновения и представления нации миру.

Корпус - это набор текстов в электронной форме, которые находят значение слов, словосочетаний, грамматических форм через определенную поисковую систему. Есть разные типы корпусов. Например, корпус авторов, корпус книг (включая первый корпус Библии). Национальный корпус данного языка включает в себя все аспекты, жанры, методы, территориальные и социальные варианты этого языка.

Как лингводидактик, языковой корпус одинаково важен при изучении родного и иностранного языков. Это открывает дверь к новым возможностям повышения эффективности обучения. Очень легко найти слово, фразу или фразу, которые редко используются в корпусе, либо проблема с их использованием и орфографией (орфографией) решается в очень короткие сроки. Следует отметить, что информация в языковом корпусе не такая, как описано в грамматике или учебнике, а такая, как в обществе. Это самый продуктивный инструмент в изучении народного и литературного языка. Сегодня, не только грамматику, среднему исследователю необходимо знать статус, уровень применения того или иного слова, фразы или конструкции, кто их использовал, когда и для какого стиля. Корпус ориентирован на решение схожих задач.

Национальный корпус необходим для изучения лексики и грамматики существующего языка. Еще одна задача корпуса - предоставить актуальную информацию об уровнях и областях лингвистики (лексикология, акцентология, история языка). Электронный корпус языка полезен не только лингвистам, но и всем людям, использующим узбекский язык: специалистам в различных областях, ученым, политикам, лексикографам, исследователям. Это комплексная универсальная система поиска информации, которую можно использовать для различных целей.

Создание национального корпуса - метод статистического исследования, компьютерный перевод, синтез речи и ее распознавание, осуществление лингвистической деятельности, такой как проверка орфографии, поможет реализовать следующий этап развития корпусной лингвистики.

Третья глава тома, озаглавленная «Общие и различные аспекты современных корпусов мира», анализирует описания 19 языковых корпораций, доступных в Интернете.

Критериями создания корпораций, созданных в мире, являются: создание и наполнение текста, синхронизация, представление разных жанров, сортировка отдельных текстов по соотношению чисел и специальных вероятностных операций, простота компьютерного анализа (размещение специальных символов для передачи интертекстуальности) .

Существующие корпорации используются для таких целей, как статистический анализ использования языка, программное обеспечение для обработки естественного языка (NLP), создание лексических ресурсов, преподавание или изучение языков. Следует отметить, что Л.Абьялова создала лингвистические модули редактирования и анализа программ обработки естественного языка, изучила процессы графического,

морфологического и синтаксического анализа текстов [35]. Тексты, представленные в корпусе, важны при изучении динамического состояния языка или при анализе предмета различных отраслей языкознания.

Распределение мировых корпораций и корпусов, созданных на протяжении многих лет, основные периоды создания корпуса текстов, корпуса английского и русского языков, их различные классификации отражены в исследованиях по корпусной лингвистике.

Исследования в области узбекской лингвистики и компьютерной лингвистики предоставили информацию о некоторых созданных мировых корпусах, но классификации в исследовании полностью не охвачены. Следовательно, исходя из характера темы в этой главе, расположение созданного языкового корпуса, включая 19 мировых корпусов в Интернете, язык использования, год создания, количество используемых слов, статистический анализ, обработка естественного языка (НЛП) анализируются программное обеспечение, лексические ресурсы, общие и различные аспекты, такие как преподавание или изучение языка.

Вторая глава работы, озаглавленная «Общие принципы построения национального корпуса и технология подготовки данных», состоит из трех глав. Он содержит общие правила национального корпуса, принципы представления информации в корпусе, научные аспекты технологии подготовки текстов для корпуса.

Размещение материалов, собранных в первой главе тома «Общие правила формирования Национального корпуса», на автомобильных хранилищах (компьютерах); Теоретически было изучено, что специфические символы и репрезентативность, которые позволяют осуществлять электронный поиск (на морфологическом, синтаксическом уровне), являются важным фактором в корпусе (полное отражение оригинальности многих жанров в языке). Даны предложения по структуре корпуса, интерфейсу программы, алгоритму работы программы, технологии получения результатов.

О технологическом процессе построения корпуса, обеспечивающего этапы технологического процесса Рыков В.В., Ю.Н. Марчук, Ю.Н. Марчук, И. Мельчук, Ш.Хамроева[36]:

1. Этап предварительной обработки текста. На этом этапе все тексты из разных источников корректируются и редактируются. Текст подготовлен к библиографическому и экстралингвистическому описанию.

а) этап преобразования и графического анализа. Большинство текстов рассматриваются на начальном этапе. В частности, удаляются элементы (рисунки, таблицы), которые не нужны для кодирования и автоматического анализа языка для компьютерного формата, а также подчеркивания в тексте.

---

[35] Абжалова М. Таҳрир ва таҳлил дастурларининг лингвистик модуллари: Монография – Т., 2020. – 176 б.

[36] Рыков В.В. Курс лекций по корпусной лингвистике. URL: http://rykov-cl.narod.ru/c.html.; Марчук Ю.Н. Основы компьютерной лингвистики. - М.: Изд-во МПУ, 2000.; Мельчук И.А. Порядок слов при автоматическом синтезе русского слова (предварительные сообщении) / Научно-техническая информация. 1985, № 12. – С.12-36; Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филол.фан.бўйича фалсафа доктори (PhD)…диссер.–Қарши, 2018. -Б.45.

б) этап автоматической маркировки. Это делается путем автоматического исправления результатов маркировки, т. Е. Исправления и разделения ошибок (ручного или полуавтоматического).

2. Этап разметки текста. На этом этапе вводятся необходимые данные корпуса (метаданные). Мета-описания корпусных текстов включают в себя: библиографическую информацию, символы, описывающие жанровые и стилистические особенности текста, информацию об авторе и многое другое. Эта информация обычно вводится вручную. Компоненты текста (абзацы, предложения, выбор слов) и чисто лингвистическая запись часто выполняются автоматически.

3. Этап предоставления доступа к делу. Корпусный дисплей выглядит так: он может распространяться на CD-ROM и доступен в режиме глобальной сети. У разных категорий пользователей будут разные права и разные возможности.

4. Заключительный этап - внесение изменений (corpus manager) в структуру специализированной лингвистической информационной системы, обеспечивающей быстрый многопараметрический поиск и статистическую обработку размеченных текстов.

Конечно, состав и количество ступеней в каждом случае могут отличаться от перечисленных выше, а фактическая технология может быть более сложной.

Во второй главе тома «Принципы передачи данных в корпусе» Р.Г.Пятровский, Д.Н. Лавров и его ученики разработали принципы выделения текста, передачи данных из корпуса, основываясь на своем опыте в формате кодирования данных для узбекского языка. национальный корпус[37]. описаны требования. В этой главе основные требования к поисковой системе Национального корпуса узбекского языка следующие:

1) поиск слов и словосочетаний по их особенностям (грамматическим, смысловым и т. д.);

2) учитывать расстояние между текстом (целым отрывком речи или произведения) и словами;

3) поиск метатекстовой информации;

4) расширенные языковые требования, включая логические ссылки, круглые скобки и текстовые операторы;

5) эффективность индексации;

6) быстро найти ответ на самый сложный вопрос;

7) широкий диапазон, употребление слов до самого большого размера (использование сотен миллионов слов).

Кодирование данных корпуса основано на самых авторитетных стандартах. Например, TEI (Text Encoding Initiative), XCES (XML Corpus

[37] Шаров С.А. Представительный корпус русского языка в контексте мирового опыта. (Электрон ресурс. https//lamb.viniti.ru) Лавров Д.Н., Харламова М.А., Костюшина Е.А. Модель представления экстралингвистической и тематической разметки в корпусе народной речи // У1-я Междунар. науч. конф. «Математическое и компьютерное моделирование», посвящ. памяти проф. Б.А. Рогозина. 23 ноября 2018. — С. 115-118.; http://ruscorpora.ru/new/sbornik2005/11polyakov.pdf

Encoding Standard), EAGLES (European Advisory Group on Language Engineering Standards). При представлении данных Национального корпуса форматирование текста, которое несет лингвистическую информацию, основано на Язык SGML / XML.

В кейсе есть два основных типа текстовой информации:

А. Текстовая информация большого массива. Включает символы, которые полностью представляют текст: имя автора, пол, дату рождения, заголовок текста, время создания текста, размер слова, тему, тип текста, стиль, область применения и т. д.

V. Лексическая информация. Лексическая информация включает в себя следующие символы: представляет отдельные слова, т.е. может использовать словоформу в определенном месте тела текста. Это включает:

V.1. Морфологические особенности:

• лексема (словоформа);

• грамматические особенности лексемы (группа слов, живые существа, преходящие события);

• грамматические особенности словоформы (число, договор, наклон, время, лицо).

V.2. Семантические символы:

семантическая классификация, таксономический класс, мереология, оценка, причинно-следственная связь, словообразовательные отношения и т. д [38].

В теле текст состоит из последовательности абзацев, абзацы состоят из предложений, а предложения состоят из слов. В этом случае основной единицей анализа является слово, а единицей текста - предложение. С помощью поисковой системы в корпусе можно найти слова и фразы, относящиеся к определенному символу, относящемуся только к этому предложению. Результатом поиска является список предложений, в которых найденные слова выделены отдельным шрифтом. При необходимости поисковый текст может быть расширен до границы абзаца, но не более того.

Таким образом, можно выделить основные структурные единицы в теле: слово, предложение, абзац, текст. Он не использует единицы, которые представляют структурное деление текста (части, главы, разделы), единицы, которые находятся за пределами абзаца, и единицы, которые представляют синтаксическую структуру предложения (предложения, группы).

Третья глава тома называется «Методы подготовки текстов для корпуса». Научно обосновано первое оформление текста, отобранного для национального корпуса, типы текстов, входящих в корпус, внешний вид языкового знака.

---

[38] Аброскин А. А. Поиск по корпусу: проблемы и методы их решения // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. –277–282 с.; ПоляковА.Е. Технология подготовки информации в национальном корпусе русского языка. http://www.ruscorpora.ru/new/corpora-biblio.html; Кустова Г. И., Ляшевская О. Н., Падучева Е. В., Рахилина Е. В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. –М., 2005.– С.155–174.

«Узбекская компьютерная лингвистика строится на основе особенностей узбекского языка, которые полностью отличаются от английского. Это показывает, что до создания узбекской компьютерной лингвистики необходимо было в совершенстве систематизировать и формализовать узбекский язык. Чтобы довести до уровня компьютерного решения богатые, обширные и глубоко разработанные языковые вопросы, такие как узбекский, требуется гораздо больше работы, чем английский », - сказал А. Пулатов[39].

Согласившись с учёным, можно опираться на его основные идеи, хотя напрямую использовать английскую компьютерную лингвистику при создании узбекской компьютерной лингвистики невозможно. При подготовке лингвистической базы и банка национальных текстов для создания языкового корпуса узбекского языка была сделана ссылка на исследовательскую работу по национальному корпусу русского языка. В исследовании, основанном на наблюдениях В.П. Захарова[40], А.Е. Полякова[41], процесс подготовки текстов для корпуса делится на следующие части:

1) первая верстка текста в минимальном формате HTML;

2) определение морфологических отметин и омонимии (в части тела);

3) разметка метатекста;

4) Изменить формат вывода для Яндекс-сервера.

Кодирование лексической информации в электронном теле адаптировано к правилам HTML / XML. Это открывает широкий спектр возможностей для быстрой обработки текста в программах различного типа, поискового индекса, морфологического парсера, конвертеров, этапов редактирования и автоматизации разметки в теле. Тексты для Национального корпуса импортируются из разных источников и представлены в разных форматах, таких как обычный текст, HTML, RTF, PDF.

В процессе подготовки текста из текста удаляются следующие элементы, не принадлежащие автору или не важные для изучения языка: номера страниц, заголовки столбцов, титульные листы, оглавление, выходные данные, систематическое написание, аннотации, комментарии редактора (сохраняются комментарии, написанные автором), рисунки, диаграммы, формулы (но подписи хранятся под ними);

Лингвистическая и экстралингвистическая маркировка - единственные форматы выражения данных, которые облегчают обмен информацией в корпусе.

Технологический процесс национального корпуса состоит из: создания словаря повторов лексем и словоформ на основе выбранных текстов; просмотреть текст на предмет любой единицы полученного словаря повторений; разделить графическое слово на слоги и составить словарь повторов слогов; сортировка словесных ресурсов; одновременная обработка

---

[39] Пўлатов А. Қ. Компьютер лингвистикаси / А.Қ.Пўлатов; масъул муҳаррир: А.А.Абдуазизов, М.М.Орипов. - Т.: Akademnashr, 2011. - 520 б. (-Б. 7.)

[40] Захаров В.П. Корпусная лингвистика. Учебно-методическое пособие. – Санкт-Петербург, 2005. – 48 с.

[41] Поляков А. Е. Технология подготовки информации в Национальном корпусе русского языка Текст. / А.Е. Поляков // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. – М., 2005. –С. 192.

неограниченного количества файлов; создавать текстовые корпуса с внешними символами; создаваемый текст - это корпус и расчет статистических данных для отдельных текстов, включенных в корпус.

*Третья глава диссертации, озаглавленная «Формирование лингвистической базы национального корпуса узбекского языка»,* состоит из пяти глав, в которых рассматриваются такие теоретические вопросы, как обеспечение, материал, важность моделирования и использование моделей в анализе слов.

Первая глава, озаглавленная «Лингвистическая база данных, обеспечивающая текстовый корпус», описывает базу данных, ее компоненты, особенности информационной системы, основные элементы базы данных при создании национального корпуса: таблицы, запросы, схемы данных, формы, отчеты. , макросы и модули. найдено Также описан план формирования базы данных для создания национального корпуса узбекского языка, а также этапы моделирования базы данных.

База данных - информация, программное обеспечение, обеспечивающее хранение, обновление, поиск и доставку данных[42]. Это автоматизированная система, представляющая собой комбинацию оборудования и персонала. Развитие этой технологии и создание подобных источников в лингвистике решают следующие задачи:

1) Проблема структуры и первичного анализа эмпирического материала позволяет создавать законченные тексты, начиная с функции единиц языкового уровня (грамматики, словари, фонетические базы данных). С одной стороны, завершение и определение структурной модели языковой системы, с другой - создание национальных моделей дискурсивных регионов и модели общеязыковой системы;

2) задача поиска новых способов установки и хранения языковой информации, а также организации доступа к этим материалам;

3) задача поиска новых способов обработки материала для оптимизации исследований и получения новых результатов;

4) решает задачу проверки результатов исследования, ссылаясь на большой материал.

Лингвистическая поддержка - это совокупность языковых инструментов, обеспечивающих адекватное функционирование языка в определенной области. При создании электронного корпуса узбекского языка его программное обеспечение создается только при безупречном лингвистическом обеспечении.

Вторая глава тома называется «Киберлексикография как лингвистический материал национального корпуса». Определенные достижения в области лексикографии в этой главе также сосредоточены на

[42] ДейтК. Дж. Введение в системы баз данных. 8-е издание.: Пер. с англ. – М.: Издательскийдом "Вильяме", 2005. –1328с.

новых теоретических проблемах лексикографии, основанных на требованиях времени. Обсуждалась также проблема кибер-словаря киберлексикографии[43].

Сегодня он популярен в контексте «виртуальных словарей, технологий для работы с ними и их создания». В то же время необходимо понимать отрасль в Интернете. В этом смысле термин киберлексикография относится к теоретической базе создания электронных словарей в Интернете - общих и специальных типов академических, энциклопедических и лингвистических словарей[44]. Создание корпуса, который является ярким примером кибер-словарей, подняло лексикографию на новый уровень. Известно, что корпус представляет собой электронный сборник письменных текстов, созданный на основе введения набора словарей, собранных с помощью специально разработанных компьютерных программ[45]. Корпус охватывает все аспекты языка в форме компьютерной программы. Таким образом, корпусная лингвистика ведет к переоценке языка с точки зрения его характеристик.

Корпус, который представляет собой набор словарей, встроенных в информационно-поисковую систему, является основным источником киберлексикографического корпуса. Киберлексикографические корпорации реализуются через специально разработанную компьютерную программу, которая предполагает отображение слов по мере необходимости. Что наиболее важно, программа корпуса исследует заданное целевое слово, определяет количество образцов в корпусе и вычисляет частоту связывания, отображает примеры, характерные для целевой части, из которых пользователь сможет продолжить дальнейшее исследование. Создание национальных киберсловаров узбекского языка является актуальной задачей, которая позволяет подключать и кормить мировой виртуальный мир узбекской лексикографией, национальными интернет-словарями как ее продуктом. Развитие киберлексикографии требует создания полной и полной базы данных киберсловаров на узбекском языке, автоматического редактирования, создания отличных программ, которые переводят с узбекского на другой язык или наоборот.

Третья глава, озаглавленная «Важность лингвистического моделирования при построении лингвистической базы данных», содержит теоретическую информацию и анализ важности метода моделирования для построения лингвистической базы и разработки специальных алгоритмов для маркировки каждой группы слов.

В компьютерной лингвистике важную роль играет термин «лингвистический модуль». Например, перевод естественного языка на

---

[43] Карпова О. М., Менагаришвили О. В. Электронные словари и кибернетическая лексикография : метод. рекомендации к спецкурсу. –Иваново: Иван. гос. ун-т, 2002. – 45 с.; T.Valiyev. Kibernetik leksikografiya va til korpusi muammolariga doir. SamDU. Ilmiy axborotnoma filologiya 2016-yil, 2-son. -B.67-70

[44] http://://studfile.net/preview/1619320/page:3/ Кибернетическая лексикография Захаров В. 1 А -10 ФИЯ ЧГУ имени И. Н. Ульянова.[Электрон ресурс]. http://www.myshared.ru/slide/10492/

[45] Сивакова Н.А.Лексикографическое описание английских и русских фитонимов в электронном глоссарии. дисс. .д-ра филол. наук в форме науч. докл. Тюмень., 2004. - 72 с..; Саженин, И. И. Словарный корпус: проблемы определения и структурной организации / И. И. Саженин; отв. ред. И.П. Матханова. // Проблемы интерпретационной лингвистики: типы восприятия и их языковое воплощение: межвузовский сборник научных трудов. – Новосибирск: Изд-во НГПУ, 2013. – С. 294 – 298

компьютерный язык, то есть создание способов обработки текста через компьютерную систему. Для этого используйте расширенные переводы программ, созданных на другие языки. Лингвистический модуль является самостоятельным компонентом таких программ. Например, если лексический модуль окружен словарным слоем (словами), грамматический модуль редактирует символы, знаки препинания, буквы и другие символы, правила орфографии орфографического модуля, морфологический модуль анализа слов (анализ слова-лексемы) и синтез (лексемообразование), суперсинтаксическая единица в синтаксическом модуле - анализируется феномен взаимосвязи предложений или слов. При создании национального корпуса узбекского языка его алгоритм основан на специфике языка.

Национальный корпус узбекского языка должен иметь возможность автоматически анализировать лексические единицы, имеющиеся в узбекском языке, включая синонимы, антонимы, омонимы, ассимиляционные слова, ранжирование слов, морфологическую структуру слова, словообразование, значение слова, его морфологические особенности. То есть в процессе составления, леммаджинга, разметки корпуса необходимо на основе индивидуальных поисков найти такие слова, входящие в корпус в текстах, и интерпретировать их конкретно. Для этого необходимо выполнить приведенный выше алгоритм лингвистического моделирования. Исследование М.Абджаловой «Лингвистические модули программы редактирования и анализа текстов на узбекском языке»[46], исследование А.Эшмуминова по лексическим единицам «Синонимическая база слов Узбекского национального корпуса»[47], автоматический анализ морфологических особенностей слов Ш.Хамроевой необходимо использовать отдельные части исследования «Лингвистические основы авторского корпуса»[48], исследование Н. Абдурахмановой «Лингвистическое обеспечение программы перевода английских текстов на узбекский язык»[49] по вопросам, связанным с переводом лексических единиц с узбекского языка. «Словарь синонимов узбекского языка», «Толковый словарь узбекских слов», «Словарь устаревших слов узбекского языка», «Словарь синонимов узбекского языка», «Словарь слов узбекского языка», которые доступны в узбекском языкознании для обозначения лексических единиц. Лингвистической опорой могут служить «Словарь противоречивых слов узбекского языка», «Словарь классификации слов узбекского языка», «Учебный этимологический словарь узбекского языка», «Учебный топонимический словарь узбекского языка». Только такие словари должны

[46] Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (Расмий ва илмий услубдаги матнлар таҳрири дастури учун)**.**Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф. –Фарғона, 2019.

[47] Эшмўминов А.Ўзбек тили миллий корпусининг синоним сўзлар базаси. Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф. – Қарши, 2019.

[48] Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф. –Қарши, 2018.

[49] Абдурахмонова Н.З. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (Содда гаплар мисолида). Филол.фан.бўйича фалсафа доктори (PhD)…дис. автореф. – Тошкент, 2018.

быть переработаны, леммы слов, разграничивать их количество в зависимости от характера слов и связывать члены ряда лемм друг с другом.

Только тогда исправленный словарь может стать основой программного обеспечения для программиста. Лингвистическое моделирование маркировки целесообразно, поскольку в лингвистической модели морфологический тег принимает форму условного сокращения. Для обозначения каждой группы слов разработаны специальные лингвистические модельные формы. Необходимо разработать алгоритм морфологической разметки языковой базы. Необходимо определить способы снабжения лингвистической базы семантической разметкой. Лингвистическая маркировка имеет большое значение в создании национального корпуса и формировании его языковой базы.

В четвертой главе тома, озаглавленной «Использование моделей в анализе фраз», добиться понимания моделей фраз, сформированных на основе грамматических правил и выраженных различными группами слов в терминах адаптации и спряжения, которые единица, к которой подходит модель, для различения морфологической формы, а также грамматических особенностей частей речи дается модель фразы.

Их первую букву мы берем за основу при создании шаблонов для компонентов фраз. То есть заглавные буквы [П] для придаточного предложения и заглавные буквы [Д] для доминантного слова выбираются в качестве общей модели словосочетания. Знак [ $\Longrightarrow$ ], указывающий направление, используется в качестве модели зависимости, чтобы указать, что части взаимосвязаны в зависимости. Тот факт, что эта модель ориентирована от подчиненного к доминирующему, обусловлен лидерским компонентом доминирующего компонента и особенностью подчинения подчиненного компонента. Для большей ясности мы приводим образцы фразы в таблицах.

<table>
<tr><th>№</th><th>Части речи</th><th>Модели</th><th>структура фразы в соответствии с ее расположением</th><th>Модель по месту нахождения</th></tr>
<tr><td>1</td><td>Подчиненное слово</td><td>П</td><td rowspan="3">Подчиненный<br>$\Longrightarrow$<br>Доминирующий</td><td rowspan="3">П $\Longrightarrow$ Д</td></tr>
<tr><td colspan="2">Подчинение</td><td>$\Longrightarrow$</td></tr>
<tr><td>2</td><td>Доминирующее слово</td><td>Д</td></tr>
</table>

Применение метода моделирования во фразеологическом анализе важно для формирования синтаксических типов словосочетания. В этом процессе первые буквы этих единиц берутся за образец, основанный на традиционном методе. Например, адаптивная составная фраза [Ap]; связанная со спряжением фраза [Cp]; управленческая фраза делится на группы по типу грамматических средств. То есть соединения, связанные конъюнктивными суффиксами; краткая управленческая фраза [CMRp];

фраза, связанная со вспомогательными средствами; вспомогательная предложная фраза оформлена символами [APp]. Образцы фраз также имеют разный внешний вид в зависимости от типа грамматических правил и грамматических показателей.

В пятой главе тома, озаглавленной «Использование лингвистических моделей в составном словообразовании», рассматривается проблема словообразования (словообразования) в узбекском языкознании, 3 различные формы аффиксации, несколько визуальных моделей словообразования в зависимости от группы слов. предложены и доказаны на примерах.

При создании лингвистической базы «Национального корпуса узбекского языка» важно создавать модели искусственных слов. В этом случае, используя лингвистические модули, предложенные М. Абджаловой, можно предложить следующую модель словообразования в 3-х различных формах методом аффиксации:

Отсюда: B = Основа, DW = производное слово

1. DW = база + «ли»; DW = асос + «ля»; DW = база + «размер»; DW = база + «lik»;

DW = база + «ци»; DW = база + «ксон»; DW = база + «дон» .....

2. В словообразовании с методом аффиксации аффиксы обычно добавляются после основания. Соответственно, искусственные слова, образованные этим методом, имеют форму «основа + суффикс» (например, «вкус + меньше», «угнетать + ред»).

3. Искусственные слова также имеют форму «префикс + основание». Это явление проявляется в основном в словообразовании с помощью аффиксов, заимствованных из таджикского языка: нечестный, неудобный, недостойный.

4. DW = be + base + lik; DW = нет + база + lik; DW = ветчина + база + lik; DW = bad + base + lik .....

За суффиксом могут следовать суффикс и префикс: бе-сабр-лик, бе-парво-лик, бе-саранджом-лик, бе-сариште-лик, хам-ярал-лик, хам-нафас-лик, но- инсоф-лик, но-мард-лик, но-макул-чилик, но-махрам-лик, но-аник-лик, бад-бахт-лик.

Можно сказать, что использование метода моделирования в области лингвистики, особенно в процессе грамматического анализа, на первый взгляд кажется движением от простоты к сложности, но служит повышению уровня понятности и точности изучаемого предмета.

Четвертая глава диссертации «Технология создания национального корпуса-менеджера узбекского языка» состоит из четырех глав. В ней рассматривается система корпус-менеджер: архитектура и модель корпуса данных, парсинг, алгоритм формирования интерфейс фрагмента корпуса, структура фрагмента онлайн-версии национального корпуса узбекского языка.

Корпус менеджеров (обозреватель корпуса или система запросов корпуса) - это инструмент для многоязычного анализа корпуса и обычно представляет собой сложную систему, используемую для поиска языковых

форм и последовательностей корпуса менеджеров. Он может предоставить информацию о тексте или информацию, предоставленную вызывающим, в терминах расположения свойства данных (например, леммы и тега). Это называется «конкорданс». Другие действия включают поиск по совместному размещению, статистику частоты и метаданные, которые обрабатываются в тексте. Относительно краткое описание корпуса менеджеров относится к серверу или механизму запросов корпуса. В этом случае аспекты, относящиеся к стороне клиента, называются пользовательскими интерфейсами. Корпус менеджера может быть представлен как программа на персональном компьютере или как веб-сервис[50].

Неотъемлемой частью понятия «текстовый корпус» является система управления текстовыми или лингвистическими данными. В последнее время его больше называют менеджером корпуса. Corpus Manager - это специализированная поисковая система. Он включает в себя собственное программное обеспечение для поиска данных, сбор статистических данных и удобную для пользователя доставку результатов.

Например: поиск по Русскому национальному корпусу осуществляется на базе Яндекс.Сервер Профессионал. Яндекс.Сервер, напротив, занимается поиском скрытых функций и отдельной табличной информации в грамматике и метатексте. Данные поиска формируются через Яндекс.Сервер. Обеспечивает полнотекстовый поиск информации с учетом морфологических особенностей русского языка на веб-сервере в корпоративной сети. Поиск ведется с учетом морфологии русского, английского, украинского языков. Он также работает на Яндексе в Интернете. Если вы введете слово «идти», вы увидите документы, содержащие слова «идти», «идет», «шелл», «шла». Результатом поиска будут документы, отсортированные по релевантности. Они учитывают не только количество документов, но и контраст слов, частоту их использования и расстояние между словами.

Запросы анализируются с точки зрения их предмета и формального содержания и интерпретируются в глоссарии научных терминов, которые работают с корпусом. Поиск состоит из сравнения отдельных элементов корпуса по порядку и определения их совместимости. В этом случае тексты корпуса считаются актуальными и рекомендуются к отправке [51].

Модель языка запросов Узбекского национального корпуса обычно включает в себя следующие элементы:

1) элементы прямого поиска (термины и информационные запросы);

2) средства морфологической стандартизации элементов текстового запроса;

3) операторы (конъюнкция, дизъюнкция, отрицание);

---

[50] Suleymanov D., Nevzorova O., Gatiatullin A., Gilmullin R., Khakimov B. National corpus of the Tatar language “Tugan Tel”: grammatical annotation and implementation. Procedia-Social and Behavioral Sciences, 2013, vol. 95, pp. 68–74. DOI: 10.1016/j. sbspro.2013.10.623.

[51] Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Suchomel V. The Sketch Engine: ten years on. Lexicography, 2014, no. 1, pp. 7–36. DOI: 10.1093/ijl/ecw029.

4) инструменты линейной грамматики (операторы расстояния и положения);

5) дополнительные условия поиска:

-поиск в отведенных для этого местах корпуса (например, в тегах);

-ограничение области поиска (по произведениям некоторых авторов, некоторым документам и их видам);

6) квалификационные (рейтинговые) требования по полученным результатам;

7) требования к форме и виду результатов.

На первом этапе разработки важно выбрать систему резервного копирования данных и управления базами данных (СУБД), необходимую для поисковой системы. Возможность использования СУБД и резервного копирования данных позволяет быстро и надежно получать доступ к большим томам в реальном времени и должна соответствовать следующим критериям:

- оперативность (1 запрос в секунду, включая скорость базы данных, включая таблицу на 100 миллионов строк);

- масштабируемость (применение требований к функциональности системы в соответствии с процессами, распределенными на нескольких машинах);

- стоимость корпуса (в анализ входит бесплатная коммерциализация и хранение данных);

- Взаимодействие с ПО (поддержка возможности работы с такими системами, как PHP и Unix);

- Наличие документов (полное наличие документов на русском, английском и татарском языках);

- Перспективы развития (динамика развития проекта, сообщество пользователей, планы разработчиков);

Архитектура базы данных и системы предназначена для ответов на следующие типы вопросов:

- для прямого поиска по словоформе или лемме;

- повторно изучить морфологические особенности явлений, представленных в виде союзов, дизъюнкций, форм отрицания, таких как and, or, yo;

- для типа гибридного поиска словоформы и морфологических признаков в лемме.

Использование архитектуры, созданной для Корпуса-Менеджера, позволяет решить множество задач. В будущем эту архитектуру можно будет легко применить для интеграции лингвистического анализа данных, включая морфологический анализатор, многозначное решение морфологического модуля и различные другие сервисы.

Такой подход к решению проблемы операционных систем для лингвистического корпуса. Эта специально разработанная система может

быть использована не только для работы электронного корпуса татарского языка, но и при внесении изменений в узбекский корпус [52].

Во второй главе тома, озаглавленной «Синтаксический анализ текстов в корпусе (синтаксический анализ)», теоретически анализируется синтаксический анализ, его функции, морфологический анализатор и т. Д.

Синтаксический анализ - это компьютерное определение синтаксического анализа. Для этого создается математическая модель для сравнения токенов, описываемых одним из языков программирования, с официальной грамматикой. Например, PHP, Perl, Ruby, Python. Когда человек читает, с точки зрения филологической науки, он синтаксически анализирует слова (токены), которые он видит на бумаге, сравнивая их в своем словаре (официальная грамматика). Машиночитаемый «сценарий» - это программа (сценарий), которая позволяет сравнивать предлагаемые слова со словами в Интернете. Сфера применения таких программ очень широка, но все они работают по практически одному алгоритму. Алгоритм работы с парсингом следующий: независимо от того, на каком официальном языке программирования он написан, алгоритм его обработки остается неизменным. Доступ в Интернет, доступ к коду веб-ресурса (доступ) и его загрузка; чтение, получение и обработка данных; представить полученные данные в обработанном виде - файлы в форматах .txt, .sql, .xml, .html и других форматах.

Парсинг решает следующие задачи:

1. Лексический анализ - это разделение текста на предложения и словосочетания.

2. Морфологический анализ слов (токенизация и лемматизация) - для определения части речи, спряжения, типа (родини) и других грамматических особенностей слов с учетом текста (приоритет смысла).

3. Синтаксический синтаксический анализ - для определения отношения слов в предложениях, для поиска притяжательного и причастия, для разделения предложений на группы по притяжательному, дополнительному и падежному типу.

4. Упрощенный синтаксический анализ (chunking) - разделение сложного текста на подпункты.

Они выполняются для всего вышеперечисленного, включая слова, которых нет в словаре. Также можно будет подключить режим коррекции орфографии морфологического анализатора.

Синтаксический анализ предназначен для быстрого анализа очень больших объемов текста (десятки килобайт или сотни мегабайт). Для достижения высокой эффективности синтаксического анализа весь словарь вводится в его оперативную память при запуске. Анализ не ограничивает размер анализируемого текста или время, необходимое для его использования.

[52] Хакимов Б.Э., Гильмуллин Р.А., Гатауллин Р.Р. Разрешение грамматической многозначности в корпусе татарского языка // Ученые записки Казанского ун-та: Гуманитарные науки. 2014. Т. 156. № 5. С. 236–244.

Токены - это объекты морфологического анализатора, функция которых заключается в выполнении стемминга, лематизации и морфологического анализа. Узбекский язык по своей структуре является агглютинативным, что влияет на алгоритм работы морфологического анализатора. Анализ был основан на словарном подходе. В нем идентифицируется каждая основа слова, и определяется, к какой парадигме принадлежит то или иное слово. Грамматический словарь основан на «Толковом словаре узбекского языка», который является единственным лексикографическим академическим исследованием.

Следовательно, синтаксический анализ - это процесс установления синтаксических связей и присоединения определенных синтаксических символов к словам или фразам.

Третья глава главы называется «Этапы формирования интерфейса корпуса». Анализирует интерфейс и его типы, дизайн.

Слово «интерфейс» происходит от английского языка и означает «внешний вид». Это слово часто используется в компьютерных технологиях. Компьютер - единственная система связи, которая обеспечивает разнообразный обмен информацией между человеком и машиной. Интерфейс - это два элемента единой системы и связующее звено, которое работает с этой системой. Интерфейс - это система связи между различными узлами и сложными аппаратными блоками, а также технологией и пользователем. Он выражается в виде логического (система представления информации) и формального (свойства информации). Он используется для выдачи команд для определенных задач. Такой интерфейс называется пользовательским интерфейсом. Интерфейс любого устройства делится на внешний и внутренний вид в зависимости от выполняемых им функций. У пользователя не будет прямого доступа к внутреннему интерфейсу, у него есть приватная опция. С помощью внешнего интерфейса пользователь может напрямую общаться и использовать его для управления устройством. Эти два типа интерфейсов всегда вписываются в одно устройство и обеспечивают его работу, они не могут существовать по отдельности. Пользовательский интерфейс можно разделить на 2 части. Например, это часть, которая отвечает за ввод данных на устройстве, а пользователь отвечает за их вывод. Если мы говорим о простом рабочем компьютере, то в первой категории у нас есть все, что работает на компьютере. Соответственно, все относится ко второй категории, через которую компьютер передает информацию пользователю в ответ на команды, подаваемые той же клавиатурой, мышью и другими устройствами ввода, то есть мониторами, динамиками, гарнитурами, принтерами, плуторами и т. д. Используемые интерфейсы в компьютерной технике бывают следующих видов:

**Визуальный.** Стандартный компьютерный интерфейс, который передает данные с помощью визуальных изображений, отображаемых на мониторе.

**Жест**. Как правило, он служит интерфейсом для телефонов или планшетов. В большинстве случаев это сенсорная панель, которая реагирует

на движения пальцев человека, управляющего системой, и реагирует в определенной степени на каждое конкретное движение. Его можно назвать упрощенной версией простого визуального интерфейса.

**Звук.** Этот тип интерфейса появился сравнительно недавно. Позволяет управлять системой с помощью голосовых команд. Система, в свою очередь, отвечает посредством взаимодействия с пользователем. Интересно, что современные технологии позволяют контролировать не только звук телефонов или компьютеров, но также звук бытовой техники и даже бортовых компьютеров.

Одна из новейших тенденций в этой области - сенсорный интерфейс. Принцип его работы основан на физическом взаимодействии пользователя и машины, которое осуществляется через определенные объекты.

Интерфейс национального корпуса отличается другим дизайном, структурой, доработка которой возложена на автора, создавшего корпус. Потому что интерфейс имеет привлекательный общий вид, который производит первое впечатление на теле. В интерфейсе должны быть учтены украшения, отражающие национальный колорит, а также символы, отражающие классику или современность.

Следовательно, важно, чтобы интерфейс был идеально и систематически разработан, чтобы продемонстрировать его удобные и наиболее эффективные возможности при использовании национального корпуса. Поэтому интерфейс должен быть создан в удобном для пользователя формате, отвечающем требованиям современного программного обеспечения.

*Четвертая глава тома называется «Структура фрагмента онлайн-версии Национального корпуса узбекского языка».* Фрагмент онлайн-фрагмента приведен в качестве примера в этой главе. Вы можете искать слово или фразу с помощью кнопки «Искать из основного текста» окна основного поиска (интерфейса). Если ввести поисковый запрос по слову «небо», «земля» или «время», результат в новом окне будет выглядеть следующим образом:

«Небо» - 1. Голубое пространство, видимое над землей в виде купола; 2. Космос вокруг Земли (астрономический космос);

Или же:

«Земля» - 1. Третья планета после Солнца. 2. Земная часть той же планеты (при размещении напротив покрытой водой части). 3. Внешний слой земной коры. 4. Место, место, место, занятое чем-либо;

Или же:

«Время» - 1. Период, время. Форма последовательного обмена состояниями материи (объекта) и событий (процессов); время, имеющее общие свойства, такие как непрерывность, неповторение, необратимость. 2. Когда идет слово; время, время, период, время в целом.

Кроме того, если слово включено в текст, включенный в лингвистическую базу данных, оно выдается как результат поиска в другом окне (в отдельном представлении).

С помощью кнопки «Лексико-грамматический поиск» тела вы можете узнать особенности слова, такие как значение, грамматический статус, синонимия, омонимия, паронимия, антонимия, вариант слова, период использования слова, стиль использования слова. Если ввести слово «тонкий» в кнопке «Лексико-грамматический поиск», в новом окне результат будет выглядеть следующим образом:

• Значение слова: 1. Поперечное сечение меньше нормы; 2. Кричащий, резкий (о звуке);

• грамматическая позиция: [качество];

• синонимия: мягкий;

• омонимия: нет;

• паронимия: нет;

• антоним: толстый;

• вариант: нет;

• срок применения: современный;

• способ применения: нейтральный;

В окне «Синтаксический поиск» вы можете увидеть типы предложения по назначению: Декларативное предложение. Повелительное предложение. Вопросительное предложение. Например, если вы наберете «декларативное предложение» в кнопке «Синтаксический поиск», результат в новом окне будет выглядеть следующим образом:

Группа моих друзей, как обычно, собралась и отправилась в Шайхантахур (Ойбек);

Я не знал, какое выражение было у него на лице, когда он сидел напротив лунного света (С. Ахмад);

Мой брат Афзалхан долго сидел, но медленно выпрямлялся (О. Хошимов);

Молодой человек больше походил на студента, чем на одежду (А. Каххор).

"Что такое корпус?" Кнопка содержит информацию о концепции кейса и кейса, созданных на данный момент.

Кнопка «Аналитика» предоставляет информацию о словообразовании (около 200 слов), усвоении слов (более 200 слов), ранжировании слов (более 150 слов), словосочетаниях (более 100 слов).

Итак, общий вид национального корпуса узбекского языка разделен на несколько окон и правую и левую колонки. В нем будут следующие окна: «Лексический поиск», «Морфологический поиск», «Синтаксический поиск». Кнопки поиска автоматически анализируют слова и фразы за считанные секунды.

## ЗАКЛЮЧЕНИЕ

1. Корпусная лингвистика - наиболее продвинутая отрасль лингвистики, а корпус - необходимый инструмент лингвистов; устные, письменные памятники являются источником информации, отражающим

национально-культурное наследие. Корпус - это набор текстов, подлежащих поисковой программе, и четко определенный корпус служит стабильной лингвистической базой для обеспечения эффективности лингвистических исследований. В качестве продукта искусственного интеллекта лингвистический корпус включает электронный словарь, портал переводов, терминологическую базу данных, виртуальную (электронную) библиотеку, электронное правительство, электронные публикации, электронные учебники и руководства. Лингвистические электронные источники, являющиеся продуктом искусственного интеллекта, считаются сырьем для создания определенного языкового корпуса.

2. Необходимо поднять международный статус узбекского языка, поднять его до уровня мирового языка общения, изучать и преподавать узбекский язык за рубежом, расширять возможности нашего национального языка и работать напрямую через национальный корпус. Языковой корпус - бесценное сокровище как средство защиты национального языка от исчезновения и представления нации миру.

3. Языковой корпус - это устройство, которое проверяет степень адаптации достоверного речевого материала без искажения смысла в полуавтоматическом режиме, наряду с обработкой существующих текстов, документов, данных, их автоматическим анализом, т. Е. Морфологическим синтаксическим и семантическим анализом. , морфологический анализ и синтез. Она превосходит электронную библиотеку возможностью анализа текста в виде автоматического анализа, синтеза.

4. Создание Национального корпуса осуществляется в два этапа: определение списка источников и оцифровка текстов (преобразование в компьютерную форму). Его технологический процесс состоит из: создания словаря повторов лексем и словоформ на основе выбранных текстов; просмотреть текст на предмет любой единицы полученного словаря повторений; разделить графическое слово на слоги и составить словарь повторов слогов; сортировка словесных ресурсов; одновременная обработка неограниченного количества файлов; создавать текстовые корпуса с внешними символами; расчет статистических данных для корпуса созданных текстов и отдельных текстов, входящих в корпус; работа с исходными текстами в формате txt, doc i rtf, автоматическая установка кодировки и др.

5. Выбраны наиболее эффективные стандарты кодирования корпусных данных. Представление его данных основано на текстовом макете SGML / XML. При кодировании лексическая информация адаптируется к правилам HTML / XML. Тексты, отобранные для Национального корпуса, взяты из разных источников и представлены в разных форматах: обычный текст, HTML, RTF, PDF.

6. Лингвистическая и экстралингвистическая маркировка создается в едином формате выражения данных в узбекском национальном корпусе, а также в мировых языковых корпорациях. Будет проведен пересмотр теоретических основ морфологической и синтаксической разметки на основе академической грамматики, будет проведена практическая работа по

сокращению системы тегов семантической разметки. Важность розетки в корпусе несравнима, ведь ширина или узость доступа к корпусу зависит от розетки корпуса. Безупречная планировка - залог широкого выбора вариантов, универсальности жилья.

7. Автоматизированная система, которая выполняет функцию информации, программного обеспечения, обеспечивая хранение, обновление, поиск и доставку данных, представляет собой базу данных. Это лингвистическая база, которая обогащает искусственный интеллект нашего родного языка, оригинальные тексты лучше всего обрабатываются на естественном языке, служат для скорости и ясности. Аннотированный словарь узбекского языка, а также произведения различных стилей, созданные на узбекском языке, могут служить лингвистической опорой национального корпуса узбекского языка.

8. Киберлексикография может быть передовым разделом современной лексикографии как лингвистический материал национального корпуса узбекского языка.

9. Corpus Manager - это специализированная поисковая система, которая объединяет программное обеспечение для поиска данных в корпусе, собирая статистику и предоставляя результаты пользователям удобным способом. Прямой поиск и гибридный поиск будут разработаны в архитектуре базы данных и системы National Corus.

10. Интерфейс - это интерфейс системы связи, который отвечает требованиям современного программного обеспечения, прост в использовании, прост в эксплуатации. Он состоит из таких типов, как визуальный, жестовый, звуковой, а также систематизированный и практичный интерфейс программирования.

**SCIENTIFIC COUNCIL FOR AWARDING SCIENTIFIC DEGREES DSc.03/04.06.2021.FIL.72.09 AT BUKHARA STATE UNIVERSITY**

**BUKHARA STATE UNIVERSITY**

**TOIROVA GULI IBRAGIMOVNA**

# THEORETICAL AND PRACTICAL ISSUES OF CREATING THE NATIONAL CORPUS OF THE UZBEK LANGUAGE

**10.00.01 – Uzbek language**

**DISSERTATION ABSTRACT**
**for the doctor of philological sciences (DSc)**

**Bukhara - 2021**

**The topic of the Doctor of Philological Sciences (DSs) dissertation has been registered with the Higher Attestation Commission of the Cabinet of Ministers of the Republic of Uzbekistan under the number B2019.3.DSc/Fil180.**

The dissertation was completed at Bukhara State University.

The dissertation abstract is available in three languages (Uzbek, Russian, English (resume)) on the website of Bukhara State University (www.fdu.uz) and on the information and educational portal "ZiyoNet" (www.ziyonet.uz).

| | |
|---|---|
| **Scientific supervisor:** | **Mengliev Bakhtiyor Rajabovich,**<br>doctor of philological sciences, professor |
| **Official opponents:** | **Mukhammedova Saodat Khudoyberdievna**<br>doctor of philological sciences, professor |
| | **Karimov Suyun Amirovich**<br>doctor of philological sciences, professor |
| | **Khakimov Mukhammadkhon Khodzhakhanovich**<br>doctor of philological sciences, professor |
| **Leading organization:** | **Termez State University** |

The defense of the dissertation will be held at the meeting of the Academic Council of Bukhara State University DSc.03/04.06.2021.FIL.72.09 at 2021 "____" _____ hours _____. (Address: 200118, Bukhara, M.Iqbol street, 11. Tel .: + 99865221-29-14; fax: + 99865221-27-07, e-mail: buxdu_rektor@buxdu.uz).

The dissertation is available at the Information Resource Center of Bukhara State University (registered under number ____).

(Address: 200118, Bukhara, M.Iqbol street, 11. Tel .: + 99865221-29-14.

The abstract of the dissertation was distributed on 20_ "___" _____.

(Protocol of the register number _____ in _____, 2021).

**D.Z. Rajabov**
Chairman of the Scientific Council for awarding scientific degrees, Doctor of Philological Sciences (DSc), Docent

**H.P.Eshonqulov**
Scientific Secretary of the Scientific Council for awarding scientific degrees, Doctor of Philological Sciences (DSc), Docent

**G.N. Murodov**
Chairman of the Scientific Seminar at the Scientific Council for awarding scientific degrees, Doctor of Philological Sciences (DSc), Professor

## Introduction (annotation of the doctoral dissertation (DSc)

**The purpose of the research** is to develop a theoretical and practical basis for the formation of the linguistic base of the national corpus of the Uzbek language.

**The object of the research** has been chosen as the national corpus of the Uzbek language.

**The subject of the research** has been selected as the linguistic base, corpus interface, corpus units, corpus layout modeling, lemmas and tags in the creation of the national corpus.

**Scientific novelty of the research work consists of the following:**

theoretically substantiated the technology of creating a national corpus of the Uzbek language on the principles of grammatical description of lexical units of the Uzbek language;

on the basis of theoretical conclusions, the existing standards for marking the national corpus based on the SGML / XML language were substantiated, as well as simple text, capital and poetic patterns of fragments that regulate modeling trajectories;

text markup format, requirements for coding lexical information and the possibility of using linguistic models in word formation, compound word formation, complementing the linguistic base of the national corpus of the Uzbek language;

in the process of preparing texts that complement the linguistic basis of the national corpus of the Uzbek language: the formation of a corpus control system and interface of the corpus through such stages as preprocessing the text, its layout, providing access to the corpus, their fast multi-parameter search and statistical processing;

in connection with the creation of the national corpus - a small online project based on the analysis of information processing using computer technologies, creating a highly effective virtual environment for the educational process.

**Implementation of the results.** The scientific significance of the research results is determined by the fact that the theoretical findings of the study of the development of the Uzbek language as an electronic database in the field of computer linguistics can be used as a source in linguistics.

The practical significance of the research results in the creation of a laboratory "Computer Linguistics" from theoretical generalizations and analysis, teaching special courses such as "Computer Linguistics", "Machine Translation", "Corpus-Based Translation Studies", raising the international status of Uzbek language, raising it to the level of world language. It can be used to study and teach the Uzbek language abroad, to expand the capabilities of our national language and to create a national corpus through bleaching.

**Implementation of research results**. Based on the scientific results obtained in the process of determining the theoretical and practical aspects of the national corpus of the Uzbek language:

on the basis of theoretical conclusions, the existing standards for marking the national corpus based on the SGML / XML language were substantiated, the introduction of distance e-learning was used in the fundamental project No. 89-03-763 Ministry of Higher and Secondary Specialized Education of the Republic of Uzbekistan dated February 9, 2021). As a result, conclusions were drawn about increasing the potential of high-quality educational services and creating the necessary software and methodological support for the implementation of information resources;

in the process of preparing texts that complement the linguistic basis of the national corpus of the Uzbek language: the formation of a corpus control system and interface of the corpus through such stages as text preprocessing, modeling methods used in the fundamental project OT-F1-002 "Psychological mechanisms of formation of national ideas and ideological immunity youth ”(2017–2020) (reference No. 89-03-763 of the Ministry of Higher and Secondary Specialized Education of the Republic of Uzbekistan dated February 9, 2021). As a result, in order to expand the use of the national and cultural heritage, oral and written monuments, created in a particular language, served to identify the features of the electronicization of samples of spiritual heritage;

theoretical conclusions were made, the format of text markup, requirements for coding lexical information and the possibility of using linguistic models in word formation, compound word formation, complementing the linguistic base of the Uzbek national corpus F1-FA-0-13229 and used in the fundamental project "Functional word formation in the modern Karakalpak language "(2012–2016) (reference No. 292/1 of the Karakalpak branch of the Russian Academy of Sciences dated January 21, 2021). As a result, the project was enriched with new scientific and theoretical information;

the monograph "Theoretical and practical issues of creating a national corpus of the Uzbek language", based on the theory and practice of the technology of creating a national corpus of the Uzbek language, was used in lectures and practical classes on the topic "Introduction to corpus linguistics" according to No. 5A120102 - Linguistics (Uzbek linguistics) and certificate No. 89-03-763 of the Ministry of Secondary Special Education dated February 9, 2021). As a result, it served to describe the process of preparing texts for the corpus, to substantiate ways of modeling existing standards and tagging technologies;

system of higher education based on the terms of corpus linguistics and related areas 5A120102 - "Dictionary of terms of corpus linguistics" (ISBN 978-620-0-61316-5) for specialists in the field of linguistics (Uzbek linguistics) (reference number 89-03-763 Higher and secondary specialized education of the Republic of Uzbekistan dated February 9, 2021). As a result, on the basis of the collected materials, a dictionary was created containing 257 words for creating a national corpus of the Uzbek language and teaching the course "Corpus linguistics" and serving to replenish the fund of Uzbek lexicography;

such conclusions as the criteria for creating a bank of texts in the Uzbek language and the grammatical description of lexical units were used in the textbook "Uzbek language" for students No. 5120100 - Philology and language

teaching (Russian) (reference No. 89-03-763 Ministry of Higher and Secondary Specialized of the Republic of Uzbekistan on February 9, 2021). As a result, a theoretical basis was developed for the creation of a national corpus of the Uzbek language;

oral and written monuments created in the Uzbek language, computerization of samples of spiritual heritage, information processing by computer technologies, machine translation, development of electronic lexicography, creation of thesauri, creation of a language corpus on the Internet. environment ”,“ Hot topic ”(Scenario No. 1-450 of the National TV and Radio Company of Uzbekistan dated December 11, 2020). As a result, information about oral and written monuments created in this language, the electronicization of spiritual heritage, the creation of a corpus of the national language, the transformation of the Uzbek language into a language "understandable" on the Internet, created the basis for a radio team to reflect on the topic.

**The structure and scope of the dissertation.** The dissertation consists of an introduction, four main chapters, a conclusion, a bibliography and annexes, the total volume of the work is 251 pages.

**ЭЪЛОН ҚИЛИНГАН ИШЛАР РЎЙХАТИ**
**СПИСОК ОПУБЛИКОВАННЫХ РАБОТ**
**LIST OF PUBLISHED WORKS**

**I бўлим (I часть; part I)**

1. Тоирова Г. Ўзбек тили миллий корпусини яратишнинг назарий ва амалий масалалари. Халқаро монография. Германия: «GlebeEdit» номли халқаро нашриёт, 2020. – 169 б.
2. Тоирова Г. Корпус лингвистикасининг атамалар луғати. Изоҳли луғат. Германия: «GlebeEdit» номли халқаро нашриёт. 2020, –68 б.
3. Toirova G. Actual problems of Uzbek linguistic research. // International Scientific Journal Theoretical & Applied Science. Philadelphia, USA. Issue: 02 Volume: 75. 07 (75). 2019. – P. 169-172. (Impact Factor –8,758)
4. Toirova G. The Role Of Setting In Linguistic Modeling. //International Multilingual Journal of Science and Technology. ISSN: 2528-9810 Vol. 4 Issue 9, September – 2019, –P.722-723 (scopus)
5. Toirova G., Astanova G., Rahimova N. Artistic Expressions of a Situational Pragmatic System. // International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019. – P.4591-4593 (scopus)
6. Toirova G.,Yuldasheva M., Elibaeva l. Importance of Interface in Creating Corpus. // International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2S10, September 2019. –P.352-355. (scopus)
7. Toirova G., Jurayeva O., Abulova Z., Norova M., Norova F. Application of Innovative Technologies in Teaching Process. // International Journal of Psychosocial Rehabilitation, Vol. 24, Special Issue 1, 2020. ISSN: 1475-7192.– P.386-390. (scopus)
8. Тоирова Г. Важность интерфейса в создание корпуса. International Scientific Journal «Internauka», // Международный научный журнал «Интернаука». – 2020. – №7. Онлайн журнал. https://doi.org/10.25313/2520-2057-2020-7-5944(Impact Factor –8,758)
9. Тоирова Г., Рахимова Н. “Hozirgi o‘zbek adabiy tili” fanini o‘qitishda masofaviy ta’limning imkoniyatlaridan foydalanish. // Бухоро Давлат университети илмий ахбороти. – Бухоро, 2019. –№3. – Б.263-269. (10.00.00; №1).
10. Тоирова Г. Ўзбек тили миллий корпусини яратишда лингвистик разметканинг ўрни. // Бухоро Давлат университети илмий ахбороти. – Бухоро, 2020. – №1 (77). – Б.125-132. (10.00.00; №1)
11. Toirova G. Importance of interface in creating corpus. // Хоразм Маъмун академияси ахборотномаси. –Урганч, 2020, – №2/2. –Б. 49-52. (10.00.00; №21)
12. Тоирова Г. Миллий корпусни яратишда интерфейснинг аҳамияти. //Қарақалпақ давлат университети хабаршысы. – Нукус, 2019, ISSN 2010-9075–№ 4(45). –С.195-198. (10.00.00; №12)

13. Тоирова Г. Миллий корпус яратишнинг технологик жараёни хусусида //Ўзбекистонда хорижий тиллар. Электрон илмий-методик журнал. – Тошкент. 2020, –№ 2 (31), –Б.57– 64. (10.00.00; №17)

14. Тоирова Г. Ўзбек тили миллий корпусни яратишда интерфейснинг аҳамияти. // Сўз санъати халқаро журнали, – Тошкент, 2020, № 3, – Б.100-105. (10.00.00; №31)

15. Toirova G. The importance of linguistic module forms in the national corpus// Zamonaviy fan, ta'lim va tarbiyaning dolzarb muammolari (Mintaqada zamonaviy fan, ta'lim va tarbiyaning dolzarb muammolari) (Elektron ilmiy jurnal), – Урганч. 2020, –№ 5 , –Б.155– 166. (10.00.00; №22)

16. Toirova G. The importance of linguistic models in the development of language bases. // Бухоро Давлат университети илмий ахбороти. – Бухоро, 2020. –№6. – Б.98-106. (10.00.00; №1)

17. Тоирова Г. Лингвистик базани тузишда модуллаштиришнинг аҳамияти // Наманган давлат университети илмий ахборотномаси. – Наманган, 2021. – №3. -Б.377-386. (10.00.00; №26)

18. Toirova G. Dil Modelsel Modellemede Ayarın Rolü. // International conference on academic studies in philology (bicoasp). – Bandırma, 26-28 September, 2019, – P.27. ISBN: 978-605-69052-7-8 26-28

19. Тоирова Г. Эффективность использования инновационных технологий в обучении родному языку. Реалізація компетентнісно орієнтованого навчання в освіті: теоретичний і практичний аспекти. Збірник наукових праць за матеріалами міжнародної науково-практичної конференції –Київ, 4 листопада 2019.– С. 310-312. ISBN 978-966-644-517-2

20. Тоирова Г. Масофавий таълимнинг имкониятлари. «Замонавий таълимда рақамли технологиялар: Филология ва педагогика соҳасида замонавий тенденциялар ва ривожланиш омиллари» мавзусида халқаро илмий-амалий масофавий конференция тўплами. – Тошкент, 2020 йил – Б.34-37. DOI 10.26739/conf_02/05/2020

21. Тоирова Г. "Ўзбек миллий ва таълимий корпусларини яратишнинг назарий ҳамда амалий масалалари" мавзусидаги халқаро илмий-амалий конференция материаллари –Тошкент, 2021 йил 7 май. –Б.90-93.

## II бўлим (II часть; part II)

22. Abuzalova M., Toirova G. «Hozirgi o'zbek tili» Morfologiya. O`quv qo`llanma. O`zRVM Muvofiqlashtiruvchi kengashning 892-030 raqamli guvohnomasi. 2019-yil, 4-oktyabr 892-sonli buyruq. –Toshkent: Navro`z, 2019. – 105 b. 6 b.t.

23. Toirova G. O`zbek tili. Darslik. O`zRVM Muvofiqlashtiruvchi kengashning 892-030 raqamli guvohnomasi. 2020-yil, 30-iyun 359-sonli buyruq. 2020. –279 b.

24. Toirova G. Systematic and informative in uzbek discourse // Journal of Social Sciences and Humanities Research Vol 5 Issue 2 June 2017. – P.1-6.

25. Тоирова Г. Нутқий мулоқотда лисоний ва нолисоний омиллар / Филология ва методика масаллари. Илмий мақоллар тўплами. –Тошкент: Наврўз, 2018, –Б.153-155

26. Тоирова Г. Нутқий мулоқотнинг лисоний вербал ва лисоний новербал воситалари хусусида // БухДУ илмий ахбороти. 2018, – № 1. –Б.67-74.

27. Юлдашева Д.,Тоирова Г. Лингвоперсонология ҳақида // «Linguistics, Translatology, Linguodidactics-NUU 2018» International conference, 4-8 october, 2018, –S. 157- 160

28. Тоирова Г., С.Ҳ.Саъдуллаева (талаба) Талаба-ёшларга таълим-тарбия беришда ахборот коммуникацион технологияларнинг ўрни // “Талаба –ёшлар тарбиясида инновацион ёндашув: тарбиянинг янги методлари ва унда ахборот коммуникацион технологияларнинг ўрни” мавзусидаги республика илмий-амалий конференция материаллари. –Тошкент, 2018, –Б.225 227.

29. Тоирова Г., Зарипова А.(талаба). Хайрлашиш – нутқий мулоқотнинг хотимаси // Педагогик маҳорат. Илмий-назарий ва методик журнал. –Бухоро, 2019, –№3. –Б.130-133.

30. Тоирова Г., Зарипова А.(талаба) Фатика и единицы фатичного общения // International euroasia Congress on Scientific Researches and Recent Trends-V. –Baku, December 16-19, 2019. –P.207. ISBN 978-625-7029-48-3.

31. Тоирова Г., Максудова М. Лингвистические свойства этикетки // Materiály XV mezinárodní vědecko – praktická konference zprávy vědecké ideje – Praha, Volume 8, 22- 30 října, –2019.–P. 20-22. ISBN 978-966-8736-05-6.

32. Toirova G., Jabborova D., Raximova N. The role of setting in linguistic modeling // Proceedings of the icecrs “generating knowladge through research”. Sidoarjo university (Indonesia), universiti utara malaysia (Malaysia), global research network (USA) publishing 4 aprel, –2019.–p.53-55.

33. Тоирова Г., Зарипова А.(талаба). Нутқий мулоқотнинг якуний босқичи ҳақида // O‘zbek tilshunosligining dolzarb masalalari (O‘zbek tiliga Davlat tili maqomi berilganligining 30 yilligiga bag‘ishlangan respublika ilmiy-amaliy konferensiya materiallari) –Buxoro, 2019-yil 19-aprel. –B. 98-101.

34. Тоирова Г., Чориқулова Н.И. (талаба). Маҳсулот ёрлиғи ва унинг тил хусусиятлари // O‘zbek tilshunosligining dolzarb masalalari (O‘zbek tiliga Davlat tili maqomi berilganligining 30 yilligiga bag‘ishlangan respublika ilmiy-amaliy konferensiya materiallari) –Buxoro, 2019-yil 19-aprel. –B.101-103.

35. Тоирова Г. «Ўзбек тилининг миллий корпус»ни яратишнинг технологик жараёни хусусида // Филологиянинг долзарб масалалари. Республика илмий-услубий конференция материаллари. –Қўқон 2020 йил 25 апрель. –Б.92-95

36. Qahhorov O., Yuldasheva D.,Toirova G. Davlat tiliga e’tibor – mustaqillikka sadoqat ramzi // Buxoro davlat universiteti ilmiy axboroti. – Buxoro, 2020. –№5. – Б.73-76.

37. Toirova G. Milliy korpusining fragmenti interfeysini shakllantirish algoritmi // Davlat tili - taraqqiyot va milliy yuksalish mezoni. (O‘zbek tiliga

Davlat tili maqomi berilganligining 31 yilligiga bag‘ishlangan respublika ilmiy-amaliy konferensiya materiallari) – Buxoro, 2020 -yil 16-oktabr. –B.401-408.

38. Тоирова Г. Ўзбек тилининг миллий корпусни яратишнинг технологик жараёни хусусида // «Филологиянинг долзарб масалалари» мавзусида бўлиб ўтадиган Республика илмий-услубий конференция материаллари. – Қўқон, 2020 йил 25 апрель. – Б.92-94.

39. Toirova G. O‘zbek tili milliy korpusining interfeysini shakllantirish algoritmi // «O‘zbek tilini dunyo miqyosida keng targ‘ib qilish bo‘yicha hamkorlik istiqbollari» mavzusidagi xalqaro ilmiy-amaliy anjuman materiallari. – Toshkent, 2020 -yil 19-20- oktabr. –B.401-408.

40. Toirova G., Hamroyeva N. The importance of linguistic models in the development of language bases // Sciences of Europe. vol 2, No 59 (2020) ISSN 3162-2364. –P. 57-64

41. Тоирова Г., Ҳамроева Н. Нутқ одоби Навойи назмида //«Башарият маънавий юксалиши ва ёшлар тарбиясида Алишер Навоий меросининг аҳамияти» мавзуидаги республика онлайн илмий-амалий конференция материаллари. – Бухоро, 2021 йил 5 февраль.–Б.260-263.

42. Тоирова Г., Хайруллаева Г. Нутқни тўғридан-тўғри разметка қилиш тизимини моделлаштириш (А.Навоий асарлари мисолида) // «Башарият маънавий юксалиши ва ёшлар тарбиясида Алишер Навоий меросининг аҳамияти» мавзуидаги республика онлайн илмий-амалий конференция материаллари. – Бухоро, 2021 йил 5 февраль. –Б.263-266.

43. Toirova G., Jahonova N. Natural Language Modeling Issue // Middle european scientific bulletin. volume 10 March 2021 – P. 247-255. ISSN 2694-9970

44. Тоирова Г. Корпус учун матнларни тайёрлаш технологияси хусусида // «Замонавий филологияда интеграцион жараёнлар» мавзусидаги Республика миқёсидаги онлайн илмий-амалий анжуман материаллари. –Бухоро, 2020 йил 25 апрель. – Б.56-58.

Автореферат “Дурдона” нашриётида таҳрирдан ўтказилди ва ўзбек, рус ҳамда инглиз тилларидаги матнларнинг мослиги текширилди.

Босишга рухсат этилди: 12.08.2021. Бичими 60x84 1/16. Рақамли босма усулида босилди. Times New Roman гарнитураси. Шартли босма тобоғи: 4.5. Адади 100 нусха. Буюртма №.258

Гувоҳнома AI № 178. 08.12.2010.
“Sadriddin Salim Buxoriy” МЧЖ босмахонасида чоп этилди.
Бухоро шаҳри, М.Иқбол кўчаси, 11-уй. Тел.: 0(365) 221-26-45.