**ҚАРШИ ДАВЛАТ УНИВЕРСИТЕТИ ҲУЗУРИДАГИ
ИЛМИЙ ДАРАЖАЛАР БЕРУВЧИ
PhD.30.05.2018.Fil.70.01 РАҚАМЛИ ИЛМИЙ КЕНГАШ**

БУХОРО ДАВЛАТ УНИВЕРСИТЕТИ

ХАМРОЕВА ШАҲЛО МИРДЖОНОВНА

ЎЗБЕК ТИЛИ МУАЛЛИФЛИК КОРПУСИНИ ТУЗИШНИНГ ЛИНГВИСТИК АСОСЛАРИ

10.00.01 – Ўзбек тили

**ФИЛОЛОГИЯ ФАНЛАРИ БЎЙИЧА ФАЛСАФА ДОКТОРИ (PhD)
ДИССЕРТАЦИЯСИ АВТОРЕФЕРАТИ**

Қарши − 2018

УДК:410

Фалсафа доктори (PhD) диссертацияси автореферати мундарижаси

Оглавление автореферата диссертации доктора философии (PhD)

Contents of Dissertation Abstract of the Doctor of Philosophy (PhD)

**ҚАРШИ ДАВЛАТ УНИВЕРСИТЕТИ ҲУЗУРИДАГИ
ИЛМИЙ ДАРАЖАЛАР БЕРУВЧИ
PhD.30.05.2018.Fil.70.01 РАҚАМЛИ ИЛМИЙ КЕНГАШ**

БУХОРО ДАВЛАТ УНИВЕРСИТЕТИ

ХАМРОЕВА ШАҲЛО МИРДЖОНОВНА

ЎЗБЕК ТИЛИ МУАЛЛИФЛИК КОРПУСИНИ ТУЗИШНИНГ ЛИНГВИСТИК АСОСЛАРИ

10.00.01 – Ўзбек тили

**ФИЛОЛОГИЯ ФАНЛАРИ БЎЙИЧА ФАЛСАФА ДОКТОРИ (PhD)
ДИССЕРТАЦИЯСИ АВТОРЕФЕРАТИ**

Қарши − 2018

Филология фанлари бўйича фалсафа доктори (PhD) диссертацияси мавзуси Ўзбекистон Республикаси Вазирлар Маҳкамаси ҳузуридаги Олий аттестация комиссиясида № B2018.3.PhD/Fil504 рақам билан рўйхатга олинган.

Диссертация Бухоро давлат университетида бажарилган.

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

Илмий раҳбар:	**Менглиев Бахтиёр Ражабович** филология фанлари доктори, профессор
Расмий оппонентлар:	**Муродова Нигора Кулиевна** филология фанлари доктори, профессор
	Каримов Суюн Амирович филология фанлари доктори, профессор
Етакчи ташкилот:	**Урганч давлат университети**

Диссертация ҳимояси Қарши давлат университети ҳузуридаги Илмий даражалар берувчи PhD.30.05.2018.Fil.70.01 рақамли Илмий кенгашнинг 2018 йил «___» ___соат ____ даги мажлисида бўлиб ўтади (Манзил: 180103, Қарши шаҳри, Кўчабоғ кўчаси, 17. Тел.: (0 375) 225-34-13; факс: (0375) 221-00-56; e-mail: qarshidu@umail.uz).

Диссертация билан Қарши давлат университетининг Ахборот-ресурс марказида танишиш мумкин (_____ рақами билан рўйхатга олинган). (Манзил: 180103, Қарши шаҳри, Кўчабоғ кўчаси, 17. Тел.: (0 375) 225-34-13; факс: (0375) 221-00-56; e-mail: qarshidu@umail.uz). Қарши давлат университети, Ўзбек филологияси факультети фаоллар зали.

Диссертация автореферати 2018 йил «_____» _____ да тарқатилди.
(2018 йил _____ даги _____ рақамли реестр баённомаси).

Н.Н.Шодмонов
Илмий даражалар берувчи
илмий кенгаш раиси, ф.ф.д.

Г.Н.Тожиева
Илмий даражалар берувчи
илмий кенгаш илмий котиби,
ф.ф.ф.д. (PhD)

Д.Тўраев
Илмий даражалар берувчи
илмий кенгаш қошидаги илмий
семинар раиси, ф.ф.д., профессор

КИРИШ (Фалсафа доктори (PhD) диссертацияси аннотацияси)

Тадқиқот мавзусининг долзарблиги ва зарурати. Жаҳон тилшунослигида компьютер ва корпус лингвистикаси муаммоларини ўрганиш XX асрнинг 40-йилларида бошланиб, бу соҳада дастлабки илмий фаразлар айтилди. Хусусан, ўтган асрнинг 60-йилларида мазкур жараён жадаллашди, XXI аср бошларида ўзида миллионлаб сўзларни акс эттирувчи юзлаб тил корпуслар пайдо бўлди. Сунъий интеллектнинг автоматик таржима, компьютер таҳлили, таҳрири, тезаурус, электрон луғат сингари имкониятлари кенгайди, илмий-назарий асослари яратилди, амалиётда қўллаш мумкин бўлган илк намуналари қўлланила бошлади. Фандаги бу янгиланишлар ахборот технологияларини тилшуносликка татбиқ этиш билан боғлиқ истиқболли илмий йўналишлар пайдо бўлишига йўл очди. Бу эса корпус, корпус лингвистикаси, унинг шаклланиши, тараққиёти, бугунги ҳолати ва корпус тузишнинг умумий тамойилларини ўрганиш заруратини ҳамда мавзу долзарблигини белгилаб беради.

Дунё тилшунослигида XXI асрга келиб, корпус лингвистикасини илмий-назарий жиҳатдан ўрганиш ҳаракати кучайди. Такомиллашиб бораётган компьютер лингвистикаси йўналишида автоматик таржима сифатини яхшилаш, тилни лингвистик моделлаштириш, ҳар бир тилга оид сўзларни леммалаш назарияси, алгоритмини яратиш ҳамда муайян тилнинг кўп асрлик миллий-маданий меросдан фойдаланиш имконини ошириш мақсадида уларни электронлаштириш жаҳон тилшунослигида долзарб масалага айланди. Тилшуносликда, хусусан, компьютер лингвистикаси соҳасида корпус яратиш, мавжуд корпуслар ҳажмини кенгайтириш, матнни автоматик қайта ишлайдиган дастурларни ишлаб чиқиш кабилар ечимини кутаётган муҳим масалалардан бири бўлиб турибди.

Истиқлол йилларида компьютер лингвистикасида автоматик таржима, сунъий интеллектнинг ўзбек тилини тушуниш ва қайта ишлашига эришиш борасида қатор тадқиқотлар амалга оширилган бўлса ҳам, корпус лингвистикаси яхлит тарзда, монографик планда ўрганилмаган. Бинобарин, барча илмий йўналишлар қаторида тилшуносликда ҳам "...илмий ва ижодий изланишларни ҳар томонлама қўллаб-қувватлаш, улар учун зарур шарт-шароитлар яратиш вазифа"сининг[1] белгиланиши фанлар интеграцияси бўйича чуқур изланишлар олиб бориш зарурлигини кўрсатади. Мамлакатимизда тилга эътибор маънавиятга эътиборнинг устувор йўналишларидан бири даражасига кўтарилди. Шу боисдан она тилимизни асраб-авайлаш, бойитиш, ундан амалий фойдаланиш самарадорлигини ошириш билан бирга, ўзбек тилининг замонавий ахборот-коммуникация тизимида кенг қўлланишига эришиш кечиктириб бўлмайдиган, долзарб вазифага айланди. Корпус лингвистикаси истиқболли илмий йўналиши

[1]Мирзиёев Ш.М. Эркин ва фаровон, демократик Ўзбекистон давлатини биргаликда барпо этамиз. Ўзбекистон Республикаси Президенти лавозимига киришиш тантанали маросимига бағишланган Олий Мажлис палаталарининг қўшма мажлисидаги нутқ. – Тошкент: Ўзбекистон, 2016. – Б.13.

сифатида ўзбек тили миллий корпусини яратиш, муаллифлик корпуси тузишнинг лингвистик асосларини ишлаб чиқиш, лингвистик моделларни тузиш сингари масалаларни замонавий илмий тамойиллар асосида тадқиқ этиш фанимиз олдида турган долзарб вазифалардан биридир.

Ўзбекистон Республикаси Президентининг 2016 йил 13 майдаги "Алишер Навоий номидаги Тошкент давлат ўзбек тили ва адабиёти университетини ташкил этиш тўғрисида"ги ПФ-4997-сон, 2017 йил 7 февралдаги "Ўзбекистон Республикасини янада ривожлантириш бўйича Ҳаракатлар стратегияси тўғрисида"ги ПФ-4947-сон Фармонлари, 2017 йил 17 февралдаги "Фанлар академияси фаолияти, илмий тадқиқот ишларини ташкил этиш, бошқариш ва молиялаштиришни янада такомиллаштириш чора-тадбирлари тўғрисида"ги ПҚ-2789-сон Қарори, 2017 йил 13 сентябрдаги "Китоб маҳсулотларини нашр этиш ва тарқатиш тизимини ривожлантириш, китоб мутолааси ва китобхонлик маданиятини ошириш ҳамда тарғиб қилиш бўйича комплекс чора-тадбирлар дастури тўғрисида"ги ПҚ-3271-сонли Қарори ҳамда мазкур фаолиятга тегишли бошқа меъёрий-ҳуқуқий ҳужжатларда белгиланган вазифаларни амалга оширишда ушбу диссертация иши муайян даражада хизмат қилади.

Тадқиқотнинг Ўзбекистон Республикаси фан ва технологиялар тараққиётининг устувор йўналишларига мослиги. Тадқиқот республика фан ва технологиялар ривожланишининг I."Ахборотлашган жамият ва демократик давлатни ижтимоий, ҳуқуқий, иқтисодий, маданий, маънавий-маърифий ривожлантириш, инновацион иқтисодиётни ривожлантириш" устувор йўналишига мувофиқ бажарилган.

Муаммонинг ўрганилганлик даражаси. Жаҳон тилшунослигида корпус лингвистикаси ўтган асрнинг 60-йилларида ўрганиш объектига айланган. "Ишонарли лингвистик маълумотлар катта массивли матнлар мажмуасидангина олиниши мумкин" деган қараш ўтган асрнинг 60-йилларида Р.Г.Пиатровский томонидан айтилган[2] бўлса-да, корпус соҳасидаги мақсадли тадқиқотлар 40-йилларда Блумфильд, Фрайс ва Бонджерслар томонидан бошланган[3]. Браун корпуси (1961-1964) тузувчилари Нилсон Френсис ва Генри Кучера илк марта корпус тузиш принципларини ишлаб чиққан. Бу борада Инглиз тили банки (1980) лойиҳаси муаллифи Жон Синклер ишлари ҳам эътиборга сазовор[4]. Рус тилшунослигида В.П.Захаров, А.Б.Кутузов, Е.В.Недошивина, В.В.Риков, В.Плунгянлар корпус, унинг турлари, ўзига хос хусусияти, корпуснинг ижтимоий аҳамияти, корпус тузиш

[2] 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.

[3] 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.

[4] 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.

тамойиллари борасида тадқиқот олиб боришган[5]. Муаллифлик корпуслари О.В.Кукушкина, А.А.Поликарпов, Е.В.Суровцевалар томонидан тадқиқ этилган[6]. Ўзбек тилшунослигида компьютер лингвистикаси, матнга лексикографик ишлов бериш ва лингвостатистик таҳлил этиш борасида бирмунча тадқиқотлар амалга оширилган. А.Пўлатов, Ҳ.Орзиқулов, С.Муҳамедов, М.Айимбетов, С.Муҳамедова, С.Каримов, Г.Жуманазарова, А.Бабанаров, Д.Ўринбоева, Н.Абдураҳмонова, А.Норов ва бошқаларнинг кузатишларини ана шундай ишлар сифатида қайд этамиз[7]. Бу тадқиқотлар матнни инновацион ёндашув – компьютер лингвистикаси ютуқлари ёрдамида лексикографик ва лингвостатистик тадқиқ этишнинг замонавий усулларини тавсия этганлиги билан долзарблик касб этган бўлса-да, уларнинг ҳеч бирида ўзбек тили корпусларини яратиш масаласи кун тартибига қўйилган эмас. Аммо улар миллий тилимиз корпусини яратиш йўлида қўйилган тамал тошлари бўлганлигини таъкидлаш жоиз. Диссертацияни тайёрлаш жараёнида юқорида санаб ўтилган тилшуносларнинг тадқиқотлари чуқур ўрганилди, муносабат билдирилди ва улардан тадқиқотда фойдаланилди.

Тадқиқотнинг диссертация бажарилган олий таълим ёки илмий-тадқиқот муассасасининг илмий-тадқиқот ишлари режалари билан боғлиқлиги. Диссертация Ф-1-06 “Истиқлол даври ўзбек адабиётида Ғарбу Шарқ адабий анъаналари синтези” (2012-2016) фундаментал лойиҳаси доирасида бажарилди.

Тадқиқот мақсади. Тадқиқотда корпус, унинг ўзига хос хусусияти, ижтимоий, лексикологик, таълимий ва бошқа соҳалардаги аҳамияти, корпус лингвистикаси тарихи, корпус турлари, муаллифлик корпусининг лингвистик қимматини ўрганиш, ўзбек тили муаллифлик корпусини яратишнинг лингвистик асосларини ишлаб чиқиш мақсад этиб белгиланган.

---

[5] 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; Недошивина Е.В. Программы для работы с корпусами текстов: обзор основных корпусных менеджеров. 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 Национальный корпус русского языка? “Отечественные записки” 2005, №2. http://magazines.russ.ru/oz/2005/2/2005_2_20-pr.html

[6]Кукушкина О.В., Поликарпов А.А., Суровцева Е.В. (Под ред. В.В.Дубичинский) Электронный корпус текстов художественных произведений А.П.Чехова: принципы организации и возможности лексикографического использования// Слово и словарь. Vocabulum et vocabularium. Сборник научных трудов по лексикографии. Вып. 12. – Харьков-Клагенфурт, 2011. – 416 с.

[7] Муҳаммедов С.А. Статистический анализ лексико-морфологической структуры узбекских газетных текстов: Автореф. дисс… канд. фил. наук.- Ташкент, 1980.; Бабанаров А. Разработка принципов построения словарного обеспечения турецко-русского машинного перевода: Автореф. дисс… канд. фил. наук. - Л., 1981.; Айымбетов М.К. Опыт лингвостатистического анализа лексики и морфологии каракалпакского публицистического текста: Автореф. дисс… канд. фил. наук.- Ташкент, 1987.; Каримов С., Қаршиев А., Исроилова Г. Абдулла Қаҳҳор асарлари тилининг луғати. Алфавитли луғат. Частотали луғат. Терс луғат. - Тошкент, 2007.; 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.; Жуманазарова Г.У. Фозил Йўлдош ўғли достонлари тилининг лингвопоэтикаси: Фил. фан. док. дис…автореф. - Тошкент, 2017.; Абдурахмонова Н.З. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (Содда гаплар мисолида). Филол.фан.бўйича фалсафа доктори (PhD)…дис. афтореф. – Тошкент, 2018.; Пўлатов А. Компьютер лингвистикаси. – Тошкент: Akademnashr, 2011.; Норов А. Компьютер лингвистикаси асослари. – Қарши, 2017. – 136 б.

Тадқиқотнинг вазифалари:

корпус, корпус лингвистикаси, унинг шаклланиши, тараққиёти, корпус лингвистикасининг бугунги ҳолати ва корпус тузишнинг умумий тамойилларини ўрганиш;

муаллифлик корпуси тузишнинг лингвистик асосларини тадқиқ этиш;

муаллифлик корпусларининг муштарак ва фарқли жиҳатларини аниқлаш;

А.П.Чехов, А.С.Пушкин, Ф.М.Достоевский, А.С.Грибоедов, У.Шекспир, Фирдавсий, Румий, Саъдий Шерозий, Лойиқ Шерали асарлари корпуси мисолида муаллифлик корпуси тажрибасини ўрганиш;

муаллифлик корпуси тузиш тамойилларини ишлаб чиқиш;

Абдулла Қаҳҳор корпуси интерфейсининг дизайн, қидирув тизими, маълумотлар базаси ойнаси мазмуни каби хусусиятларини тавсифлаш;

Абдулла Қаҳҳор асарлари мисолида сўзларни морфологик теглаш ва лингвистик моделлаштириш йўлларини асослаш;

семантик гуруҳ, тўда ва майдон таснифининг матнни семантик теглашдаги аҳамиятини ёритиш;

Абдулла Қаҳҳор асарлари мисолида синтактик теглашнинг лингвистик асосларини яратиш.

Тадқиқотнинг объекти сифатида корпус, унинг турлари ҳамда муаллифлик корпуси танланган.

Тадқиқотнинг предметини Абдулла Қаҳҳор асарлари корпуси интерфейси, корпус бирликлари, теглаш муаммолари ташкил этади.

Тадқиқотнинг усуллари. Тадқиқот мавзусини ёритишда таснифлаш, тавсифлаш, қиёслаш, статистик, когнитив таҳлил методларидан фойдаланилди.

Тадқиқотнинг илмий янгилиги қуйидагилардан иборат:

ўзбек тилшунослигида корпус, унинг лингводидактика ва лингвистикада маълумотларни марказлаштириш, қайта ишлаш имконияти ҳамда корпусдан фойдаланиш аҳамияти очиб берилган;

корпус яратиш уч босқичда тараққий этганлиги, биринчи авлоди электрон кутубхона шаклидалиги, иккинчи авлоди матнни қайта ишлай олиши, янги даврда эса ҳажман катта, аммо содда тегли корпуслар яратилаётганлиги, рус ва инглиз тили корпуслари тарихи, корпус лингвистикасининг бугунги ҳолати, замонавий рус, инглиз, турк, тожик тили корпусларининг жаҳон корпуслари орасидаги ўрни, уларнинг муштараклиги корпуснинг умумий талаблари даражасидалиги, фарқли томонлари эса сўз миқдори ва лингвистик таҳлил имкониятида эканлиги аниқланган;

корпусни лойиҳалаш ва тузиш технологик жараёнлари ўрганилган, тегнинг корпусда матнни қайта ишлашни белгилаб бериши, тег сўзшаклларнинг грамматик белгиларини кўрсатиши асосланган, корпус менежерининг қидирув, саралаш, фильтрлаш вазифаси ва унинг BONITO, XAIRA, SARA, SQR, DDS ва бошқа турлари келтирилган;

муаллифлик корпусининг илмий, амалий, таълимий мақсад ва вазифалари, миллий, параллел, таълимий, мультимодал корпуслардан фарқи,

муаллифлик корпусининг тузилиши, таркиби, ҳосиласи, рус ва инглиз тилида мавжуд муаллифлик корпусларининг ўхшаш ҳамда фарқли жиҳатлари очиб берилган;

муаллифлик корпусини тузишнинг лойиҳалаш, теглаш ҳамда қидирув тизими (корпус менежери)ни танлаш тамойиллари ишлаб чиқилган.

Тадқиқотнинг амалий натижалари қуйидагилардан иборат:

Абдулла Қаҳҳор корпуси интерфейси лойиҳалаштирилган;

ўзбек тилидаги сўзларни морфологик, семантик, синтактик теглаш ва лингвистик моделлаштириш усуллари очиб берилган;

Абдулла Қаҳҳорнинг “Бемор” ҳикояси матни морфологик, семантик, синтактик тегланган ва натижалар умумлаштирилган;

Абдулла Қаҳҳор муаллифлик корпуси лойиҳаси яратилган.

Тадқиқот натижаларининг ишончлилиги ўрганилган материалларнинг ўзбек тили табиатидан келиб чиққан ҳолда хулосалар қилишга ёрдам берганлиги, уларнинг асосли эканлиги, методологик мукаммаллиги, муаллифлик корпуси тамойилларини яратишда амалда исботланган манбаларга таянилганлиги билан изоҳланади.

Тадқиқот натижаларининг илмий ва амалий аҳамияти. Тадқиқот ўзбек тили миллий ва муаллифлик корпусларини яратишнинг назарий асосларини ишлаб чиқишда, компьютер лингвистикаси йўналишида тадқиқотлар яратишда илмий-назарий манба сифатида хизмат қилади.

Тадқиқотнинг амалий аҳамияти амалий филология бўлимларидан бири − корпус лингвистикасининг фан сифатида ўқитилиш жараёнида дастур, режалар тузиш ҳамда мавзуларни баён этишда манба вазифасини ўташи, ўзбек тилида турли типдаги корпуслар ҳамда бошқа муаллифлик корпусларини яратишда намуна бўла олиши билан изоҳланади.

Тадқиқот натижаларнинг жорий қилиниши. Тадқиқотда илмий асосланган корпус тузишнинг умумий тамойиллари: корпусни лойиҳалаш ва тузиш босқичининг технологик жараёни, тегнинг корпус тузишдаги аҳамияти ва лингвистик восита эканлиги, корпус менежерининг қидирув, саралаш, фильтрлаш хусусияти; унинг турлари тавсифи асосида:

ўзбек тилшунослигида корпус, унинг лингводидактика ва лингвистикада маълумотларни марказлаштириш, қайта ишлай олиш хусусияти, назарий асослари, тил корпусининг лингвистик, амалий ва таълимий аҳамияти ҳақидаги назарий маълумотлар; корпус лингвистикасининг бугунги ҳолати, замонавий рус, инглиз, турк, тожик тили корпусларининг жаҳон корпуслари орасидаги аҳамияти, уларнинг муштарак ва фарқли томонлари қиёсланиши натижаларидан “Корпус лингвистика атамаларининг қисқача изоҳли луғати” яратишда фойдаланилган (Олий ва ўрта махсус таълим вазирлигининг 2018 йил 26 октябрдаги 89-03-3647-сон маълумотномаси). Илмий натижалар асосида корпус лингвистикаси атамалари сираси мукаммаллашишига эришилган;

диссертациянинг сўзшаклларни теглаштириш, уларнинг лингвистик моделларини тузиш, тил корпусида ясама сўзни таҳлил қилиш тамойиллари борасидаги таҳлил ва хулосалари натижаларидан Самарқанд давлат чет

тиллар институтида 2008-2011 йилларда бажарилган ОТ-Ф8-062 рақамли “Тил тараққиётининг деривацион қонуниятлари” мавзусидаги фундаментал лойиҳада фойдаланилган (Олий ва ўрта махсус таълим вазирлигининг 2018 йил 26 октябрдаги 89-03-3647-сон маълумотномаси). Тадқиқот натижаларини қўллаш деривация ҳодисаларини изоҳлашда хизмат қилган;

корпусни лойиҳалаш ва тузиш технологик жараёнлари, тегнинг корпусда матнни қайта ишлашни белгилаб бериши, тег сўзшаклларнинг грамматик белгисини кўрсатиши; корпус менежерининг қидирув, саралаш, фильтрлаш вазифаси, унинг BONITO, XAIRA, SARA, SQR, DDS турлари тавсифи; муаллифлик корпусининг электрон луғат, электрон кутубхонадан фарқли жиҳатлари қиёси; муаллифлик корпусини тузишнинг лойиҳалаш, теглаш, корпус менежери танлаш тамойилларини белгилаш натижаларидан Камолиддин Беҳзод номидаги Миллий рассомлик ва дизайн институтида давлат илмий-техника дастури доирасида 2014-2016 йилларда бажарилган ЁА1-ФҚ-0-07289 сонли “Миллий амалий ва тасвирий санъат атамаларининг қисқача ўзбекча, русча, инглизча изоҳли луғати” мавзусидаги фундаментал тадқиқот лойиҳасида фойдаланилган (Олий ва ўрта махсус таълим вазирлигининг 2018 йил 26 октябрдаги 89-03-3647-сон маълумотномаси). Натижада амалий ва тасвирий санъат атамаларининг қисқача изоҳли луғати илмий-оммабоплиги таъминланган ва янги манбалар билан бойиган;

корпус интерфейси лойиҳалаштирилган; асарлар матни теглари лингвистик моделидан муаллифлик корпусининг off-line варианти фрагментини яратишда фойдаланилган (Олий ва ўрта махсус таълим вазирлигининг 2018 йил 26 октябрдаги 89-03-3647-сон маълумотномаси; Ўзбекистон Республикаси Интеллектуал мулк агентлиги ҳузуридаги IP-CENTERнинг №000895 ҳамда №000986 рақамли гувоҳномалари). Натижада “Абдулла Қаҳҳор асарлари корпуси” off-line варианти фрагменти яратилган.

Тадқиқот натижаларнинг апробацияси. Мазкур тадқиқот натижалари бўйича 2 та халқаро ва 5 та республика илмий-амалий анжуманларида маърузалар қилинган.

Тадқиқот натижаларнинг эълон қилинганлиги. Диссертациянинг асосий мазмуни муаллиф томонидан чоп этилган 1 та луғат, 2 та муаллифлик гувоҳномаси, 13 та илмий мақола (уларнинг 5 таси Ўзбекистон Республикаси ОАК тасарруф этган илмий журналларда) ва тезисларда ўз ифодасини топган.

Диссертациянинг тузилиши ва ҳажми. Диссертация кириш, уч асосий боб, умумий хулоса, фойдаланилган адабиётлар рўйхати ва иловадан иборат. Диссертациянинг умумий ҳажми 165 саҳифани ташкил этади.

ДИССЕРТАЦИЯНИНГ АСОСИЙ МАЗМУНИ

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

Диссертациянинг **“Корпус лингвистикаси шаклланиши, тараққиёти ва назарий асослари”** деб номланувчи I бобида корпус, корпус лингвистикаси, унинг шаклланиши, тараққиёти, бугунги ҳолати ўрганилган ҳамда ушбу масалаларга муносабат билдирилган. Бобнинг **“Корпус ва корпус лингвистикаси ҳақида”** деб аталувчи биринчи бўлимида корпус лингвистикаси, унинг предмети изоҳланган, илмий адабиётларги таърифлари тил корпуси маълум тилнинг белгиланган даврдаги, хилма-хил жанр, ранг-баранг услуб, ҳудудий ҳамда ижтимоий вариантдаги матнларнинг электрон шаклдаги махсус дастурий таъминот асосидаги йиғиндиси деган фикрда умумлаштирилган[8]. Жаҳон тилларининг жуда кўпчилиги мукаммаллик даражаси, матнни қайта ишлаш имконияти билан фарқ қилувчи ўз миллий корпусларига эгалиги, бугунги кунда лингвистик тадқиқот ва амалий топшириқлар ечими учун тил корпуслари замонавий тилшуносликнинг инкор этиб бўлмас иш қуролига айланганлиги, корпуснинг оддий электрон кутубхонадан фарқи, корпус аннотацияси, конкорданс (қидирув тизимининг нисбатан оддий кўриниши) ёки корпус менежери, унинг қидирув имконияти, корпус менежерига қўйилган умумий талаблар ёритиб берилган. Корпуснинг лексикография, лексикология, синтактис, услубиятни ўрганишдаги лингвистик аҳамияти, лингводидактика, она тили, хорижий тил таълимидаги ўрни ёритиб берилган.

Тилнинг миллий корпуси ахборот манбаи сифатида қуйидаги қулайликларни яратишини таъкидлаш жоиз:

1) корпуслашган тилда яратилган оғзаки, ёзма ёдгорликлар, миллий, маданий мерос намуналари электрон кўринишда Интернет тармоғидан жой олади;

2) корпус табиий (реал) тилнинг электрон шаклдаги, қидирув дастурига жойлаштирилган матнлар йиғиндиси; бир марта тузилиб, мукаммал тегланган корпус лингвистик тадқиқотлар самарадорлигини таъминлашда барқарор лингвистик база вазифасини бажаради;

3) корпус электрон кутубхона, луғат, грамматикалар яратишга асос бўлади. У кенг кўламли бўлганлиги учун маълумотнинг ўзига хослигини кафолатлаб, тил ҳодисаларининг барча қирраларини тўлиқ намойиш этишни таъминлайди;

[8] 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.

4) турли маълумотлар тил корпусида ўзининг табиий контекстуал шаклида жойлашади, бу эса уларни ҳар томонлама, чуқур, объектив ўрганишга асос бўлади;

5) корпус – тилни тадқиқ этиш (сўзнинг ўзгариши, историзм, неологизмларнинг вужудга келиши, маъно кенгайиши, торайиши, янги фразеологизмларнинг пайдо бўлишини кузатиш), тил ўрганиш, луғат тузишда энг замонавий, кенг имкониятли дастурлаштирилган тизим.

“Корпус лингвистикасининг шаклланиши ва тараққиёти” бўлимида компьютер асригача ва ундан кейинги корпус шакллари, тил корпус яратишнинг 4 асосий даври: 1960, 1970, 1980, 2000 йилларда яратилган корпуслар, уларнинг хусусиятлари, инглиз, рус корпус лингвистикаси тарихи ҳақида батафсил маълумот берилган. Корпуснинг пайдо бўлиши, ривожланиши икки босқич: компьютер асригача бўлган ва компьютер асри корпуслари давридан иборат. Биринчи давр корпуси – картотекалар тўпламидан иборат, гарчи бугунги корпус кўринишида бўлмаса-да, лингвистик тадқиқот учун материал вазифасини ўтаган. Компьютер асрида эса улар электрон шаклга кирган, дастурлаштирилган. Том маънодаги корпус лингвистикасига ўтган асрнинг 60-йилларида инглиз корпуси (Браун корпуси) муаллифлари асос солишган. Инглиз тили, унинг вариантлари учун турли мақсадни кўзловчи 100дан ортиқ корпус тузилган. Рус тилшунослигида корпус лингвистикасининг тамал тоши 1980 йилларда Швециянинг Упсала университетида қўйилган.

“Корпус лингвистикасининг бугунги ҳолати” бўлимида корпус турлари, унинг репрезентативлик хусусияти, таркибига кирувчи жанрлари, ёзма, оғзаки, бир тилли, кўп тилли корпус; балансга келтирилган, ҳажман у қадар катта бўлмаган, алоҳида тадқиқий топшириққа хизмат қилувчи махсус матнлар корпусининг ўзига хос хусусиятлари ёритиб берилган.

Корпус тузилиши, мақсади, турғун/ўзгарувчанлиги каби жиҳатларига кўра қанча кўринишга эга бўлмасин, уларни икки томондан гуруҳлаш мақсадга мувофиқ:

а) бутун тилга (одатда, тилнинг маълум даврига) ёки унинг маълум бир воқеланиш тури(жанр, услуб, бирор ижтимоий ёки ёшга оид гуруҳ, ёзувчи/олим тили)га оидлигига кўра;

б) лингвистик тег турига кўра.

“Муаллифлик корпуси тузишнинг умумлингвистик асослари” деб номланувчи II бобда корпус тузишнинг умумий тамойиллари, муаллифлик корпуси тузишнинг лингвистик асослари, муштарак ҳамда ўзига хос жиҳатлари тадқиқ этилган.

“Корпус тузишнинг умумий тамойиллари” бўлимида корпусни лойиҳалаш ва тузиш босқичининг технологик жараёни ёритилган. В.П.Захаров, С.Ю.Богдановалар корпусни лойиҳалаштириш жараёнининг муҳим жиҳати сифатида хронология масаласини келтиришади. Масалан, тилнинг замонавий корпуси деганда нима тушунилиши лозим? Турли жанрларда корпуснинг хронологик чегараси турлича бўлиши табиий. Шу билан бирга, матн таркибида мавжуд бўлган расмлар тил материалига

тегишли бўлмаганлиги сабабли корпус таркибига кирган матндан чиқариб ташлашни, жадвалларни корпусга мослаб қайта ишлаш, иқтибос, кўчирма гаплар, ўзлашма бирлик (атама)лар, ўлчов бирликлари ҳам алоҳида эътиборни талаб қилади. Санаб ўтилган масалаларнинг баъзиси лойиҳалаштириш босқичида маълум принцип асосида ҳал этилса, айрими корпус тузиш жараёни ёки корпусдан фойдаланишда ҳал этилади. Кузатиш ва таҳлиллар натижасида корпусни яратиш қуйидаги босқичлардан иборат эканлиги аниқланди:

1. Таҳлил, матнга дастлабки ишлов бериш босқичида турли манбалардан қабул қилинган матнлар филологик текширув – таҳрирдан ўтади.

2. Конверсиялаш, графематик таҳлил жараёнида баъзи матнлар қайта кодлаштириш жараёни амалга ошадиган илк машина ишловидан қайта-қайта ўтади, номатний қисмлар (расм, жадвал) ўчирилади ёки ўзгартирилади. Матндаги бўғин кўчириш, чегаралар (MS-DOS матнларида) бекор қилинади, тире ва бошқа белгилар бир хиллигига эришилади. Графематик таҳлилда корпусга кирувчи матн қисмлар(сўз, боғловчи)га ажратилади, номатний элементлар ўчирилади.

3. Ностандарт (нолексик) элементни белгилаш, расмийлаштириш, махсус матний элемент(қисқартма асосида ёзилган ном (исм, фамилия), бошқа алифбода ёзилган ўзлашма лексема, расмга берилган ном, изоҳ, зарварақ, адабиётлар рўйхати)ни бир хил мезон асосида қайта кўриб чиқиш амаллари автоматик равишда матн муҳаррири томонидан бажарилади.

Ушбу бўлимда тегнинг корпус тузишдаги аҳамияти, лингвистик восита эканлиги, матнни автоматик теглаш, корпуснинг қидирув тизими – корпус менежери масалаларига алоҳида ўрин ажратилган.

Тег икки: лингвистик ва экстралингвистик турга ажратилади[9]. Экстралингвистик тегнинг қуйидаги кўринишлари фарқланади:

1. Матн форматининг ўзига хослигини акс эттирувчи (боб, хат боши, қисм ва ҳ.) тег.

2. Матн ва унинг муаллифига тегишли маълумотни ифодаловчи тег.

Лингвистик тегнинг морфологик, синтактик, семантик, анафорик, просодик кўринишлари мавжуд. Корпусни теглаш (инг. tagging) дастурлаштирилган йўл билан амалга оширилади. Бунда, аввало, вақтни тежаш, меҳнатни камайтириш назарда тутилса, иккинчидин, матнга автоматик ишлов бериш муаммосига ечим топилади. Ҳозирча анафорик, просодик теглаш қийинлигича қолиб кетяпти; теглаш фақат қўлда бажариляпти, кейинчалик бу ҳам дастурлаштирилади, албатта. Морфологик, синтактик теглаш теггер, парсинг ёрдамида амалга оширилса ҳам, бу дастурларнинг аксарияти автоматик теглашдан кейинги тузатишни талаб қилади. Чунончи, морфологик омонимия (кўпроқ флектив тилларга хос), синтактик кўпмаънолилик ҳолатида дастур хулосанинг бир неча кўринишини таклиф қилади, тадқиқотчи кераклисини танлайди. Янги авлод корпуслари

[9]http://rykov-cl.narod.ru/c.html.

ҳажмининг фавқулодда катталашганлиги мутахассислар олдига теглашнинг тўлиқ автоматлаштирилган турига ўтиш, янги, мукаммал теггер, парсинглар яратиш вазифасини қўяди. Автоматик морфологик таҳлил (теггер) ёрдамида ҳар бир лексик бирликка (сўз туркуми, лемма, граммема гуруҳи) алоҳида грамматик характеристика (шахс-сон, келишик, бошқа грамматик категория) берилади.

В.П.Захаровнинг фикрича, лингвистик теглашнинг барча (морфологик, синтактик, семантик, анафорик, просодик) тури қуйидаги тамойиллар асосида амалга оширилади[10]:

1) тег схемасини тавсифлаш (асослаш);
2) умумий лингвистик тушунчалар тизимини аниқлаш;
3) фойдаланувчи учун маълум бўлган таҳлил схемасини шакллантириш;
4) тег схемасининг назарий анъанавийлигига эришиш;
5) халқаро андозаларга амал қилиш.

Тил корпусининг ажралмас, асосий қисми унинг қидирув тизими– корпус менежери; у матн ва лисоний бирликларни бошқарувчи система. Корпус менежери − корпус маълумотлари устида ишлашга мўлжалланган махсус қидирув тизими; статистик маълумот, қидирув натижасини фойдаланувчига қулай кўринишда кўрсатиб берувчи дастурий таъминот. Е.В.Недошивина корпус менежерига қўйиладиган талабларни санар экан, уларнинг энг асосийси сифатида матннинг калит сўзлари рўйхати, тўлиқ конкорданс рўйхатни ярата олиш; фақат сўзни эмас, балки сўз бирикма ҳолидаги сўровга ҳам жавоб бера олиш; шаблон асосида (мураккаб сўров) қидирувни амалга ошириш; олинган натижа(чиқарилган рўйхат)ни бир неча мезон асосида саралай олиш; сўзшаклга берилган сўровни чекланмаган миқдордаги контекстда акс эттириш; корпуснинг алоҳида элементлари бўйича статистик маълумот бера олиш; корпус тегидан келиб чиққан ҳолда лемма, сўзшаклнинг морфологик хусусияти ҳамда метаахборот (библиографик, типологик)ни тўлиқ ифодалай олиш; натижаларни сақлаш, чоп этиш; файл ва корпуснинг чекланмаган ҳажми билан ишлай олиш; қидирувни тез амалга ошириш, натижаларни чиқариш; турли матн форматларини (txt, doc, rtf, html, xml ва б.) ўқий олиш ҳамда шу формат билан ишлаш; малакали ҳамда янги фойдаланувчи учун бирдек қулай бўлиш каби хусусиятларга алоҳида тўхталади[11]. Демак, корпус тузиш унинг тамойилларини ишлаб чиқишдан бошланади; бунда асосий эътиборни корпусни лойиҳалаштириш, теглаш, унга мос қидирув тизими (корпус менежери)ни танлашга қаратиш лозим. Корпусда тегнинг аҳамияти тенгсиз, чунки корпусдан фойдаланиш имконининг кенг ёки торлиги корпусдаги тег хусусиятига боғлиқ.

Мукаммал тег – кенг имкониятли, универсал корпус гарови. Корпус тузиш учун теглаш дастурлари − парсинг, таггинглар муҳим восита. Шу

[10] Захаров В.П., Богданова С.Ю. Корпусная лингвистика.– Иркутск: ИГЛУ, 2011. – С.76.

[11] Недошивина Е.В. Программы для работы с корпусами текстов: обзор основных корпусных менеджеров. работа с системой DDC.// Языковая инженерия: в поиске смыслов. – (Электрон ресурс) : https//docplayer.ru.

сабабли бугунги кунда сунъий интеллектнинг ўзбек тилини “ўқиш, тушуниш, қайта ишлаш, унга мос жавоб қайтариш”га қодир дастурлар яратиш дастурчилар олдида турган долзарб вазифадир. Сунъий интеллект ўзбек тилини “билмайди”, уни “тушунадиган” махсус парсинг, таггинг дастурлари ишлаб чиқиш лозим.

II бобнинг **“Муаллифлик корпуси тузишнинг лингвистик асослари”** бўлими муаллифлик корпуси мақсад ва вазифалари, идеографик теглаш ва корпус интерфейси муаммоларига бағишланган. Корпус интерфейси – фойдаланувчини унинг бутун мазмуни билан таништирувчи, мундарижа вазифасини бажарувчи жуда муҳим таркибий қисм. Муаллифлик корпуси интерфейси – муаллиф ижодий мероси жанрий таснифини бир қарашдаёқ тушуниш учун яхши имконият; корпуснинг бошқа маълумотлар базасидан фарқи ҳам шундай имкониятнинг мавжудлигида. Муаллифлик корпусининг муҳим таркибий қисми тегнинг қай даражада серқирра ва мукаммаллигидан қатъи назар, интерфейсдир. Мукаммал ишланган интерфейс – фойдаланувчи учун тушунарлилик ва ишлашни қулайлаштирувчи омил. Муаллифлик корпусининг яна бир муҳим белгиси тегланганлик даражаси ҳамда тури. Лингвистик тег асосида частотали, терс луғат тузиш мумкин бўлса, муаллиф асарларини идеографик теглаш идеографик луғат яратишга асос бўлади.

Бобнинг **“Муаллифлик корпусларининг муштарак ва ўзига хос жиҳатлари”** бўлими муаллифлик лексикографияси ва муаллифлик корпуси муносабати, унинг ўзига хос хусусияти: тузилиши, таркиби ва ҳосиласи, А.С.Пушкин, Ф.М.Достоевский, А.С.Грибоедов муаллифлик корпусларининг ўхшаш ва фарқли томонлари тадқиқига бағишланади. Муаллифлик лексикографияси, муаллифлик луғатини тузишнинг назарий асослари, муаллифлик идиостилини ўрганиш тажрибаси, муаллифлик электрон луғати тузишнинг ўзига хос жиҳати, миллий ва адабий тил ички корпусидан муаллифлик лексикографиясида фойдаланиш тажрибаси, янги турдаги луғатлар тузиш ғоялари, муаллифлик лексикографияси асосида замонавий адабий тилдаги ўзгаришларни қиёсий-таҳлилий тадқиқ этиш муаммоларига бағишланган[12]. А.П.Чехов бадиий асарлари корпуси мисолида муаллифлик корпусининг ўзига хос хусусияти: тузилиши, таркиби ва ҳосиласи борасидаги изланишлар асосида корпуснинг шартли равишда “ИСТОК”[13] деб аталувчи ахборот-тадқиқий тизими, унинг асосий вазифаси, мавжуд имкониятлари; А.П.Чехов корпуси асосида *А.П.Чехов бадиий асарлари тили грамматик-семантик частотали луғати* яратилганлиги аниқланган. Луғат ва электрон корпус йиғиндиси – янги типдаги лексикографик маҳсулот. Бу маҳсулотнинг аҳамияти луғат билан ишлашнинг янгича услуби, қулайлиги, фойдаланувчининг ёзувчи ижоди тил хусусиятлари бўйича кейинчалик турли хил мустақил тадқиқотлар олиб бориш имконининг мавжудлиги билан

[12] www.ruslang.ru/seminar_aut_lexocorg020413 сайтидаги “Теория и практика авторской лексикографии” семинар материаллари.

[13] “ИСТОК” – “Исследование Словаря, Текстовых Особенностей, Конкордансов” сўзларининг бош ҳарфидан олинган қисқартма.

белгиланади[14]. Муаллифлик корпуси муаллифлик лексикографияси билан чамбарчас боғлиқ. Шу сабабли муаллифлик лексикографиясининг тараққиёт даражаси корпус лингвистикаси билан икки томонлама муносабатда: бирининг тараққиётини иккинчисининг ривожисиз кўриб бўлмайди. Афсуски, тилшуносликнинг бу икки йўналишини жаҳон тилшунослигида ҳам ривожланган соҳа дейиш қийин; муаллифлик лексикографияси йўналишида анча ишлар қилинган бўлса ҳам, бугунги кунда ечимини кутиб турган масалалар кўп. Муаллифлик корпуси тараққиёти эса турғун ҳолатда. Корпус имкониятларининг кенгайиб бориши муаллифлик корпусларининг мукаммаллашувига таъсир этмаяпти. Мавжуд муаллифлик корпуслари (Европа тиллари корпуси) жуда содда тегланган. Бу борада рус корпус лингвистикасида эришилган ютуқлар ҳавас қилгулик. А.П.Чехов, А.С.Пушкин, Ф.М.Достоевский каби буюк ёзувчи, шоирларнинг корпуси, Британия миллий корпуси таркибидаги Шекспир, Манҳейм университети корпусидаги Гёте корпуси ҳамда форс-тожик мумтоз, муосир адабиёти вакиллари корпуси муаллифлик корпусининг ютуғи; улар соҳанинг кейинги ривожи учун намуна бўлиб хизмат қилади. Муаллифлик корпуси устидаги изланишлар қуйидаги хулосаларга олиб келди:

корпус мақсадига кўра тадқиқий ва иллюстратив бўлади; муаллифлик корпуси ҳар икки мақсадда ҳам тузилиши мумкин. Тадқиқий корпус тузувчиси маълум лингвистик муаммони ечиш учун ўз корпусини тузади, шу асосда қилаётган илмий тадқиқоти учун хулоса чиқаради; иллюстратив корпус муаллифи эса корпусни кенг фойдаланувчилар турли амалларни (таълимий топшириқ тузиш, корпус асосида тажриба ўтказиш, таржима, матн тузишда лингвистик база сифатида фойдаланиш) бажаришига мўлжаллаб тузади. Демак, муаллифлик корпуслари – *тадқиқий* ва *иллюстратив* корпус кўриниши;

корпуснинг ўзгарувчанлигига кўра турғун ва динамик тури фарқланиб, муаллифлик корпуси *турғун корпус*, тўлиқ/фрагмент кўриниши бўйича *тўлиқ матнли корпус* сирасига киради. Чунки муаллифлик корпуси миллий корпусда бўлганидек доимий янгилаб боришни талаб қилмайди: бир марта тузилади, тўлдириш, тузатиб боришга эҳтиёж қолмайди. Тўлиқ матнли корпус сифатида эътироф этилишига сабаб шуки, муаллифлик корпусида бир муаллиф ижоди тўлиқ қамраб олинади ҳамда бутун ижодий мероси (асар матни тўлиқ олинади) киритилади. Нутқ турига кўра *ёзма*, параллеллигига кўра *бир тилли*, матннинг ихтисослашувига кўра *аралаш* (чунки бир муаллифнинг барча жанрдаги асарлари киритилади), кириш усулига кўра (корпус тузувчисининг ихтиёрига қараб) ҳар хил: *пуллик, эркин, ёпиқ* бўлиши мумкин. Изоҳ хусусиятига кўра *морфологик, синтактик, семантик тегланган*, матн ҳажмига кўра *тўлиқ матнли*, хронологик жиҳатига кўра

[14] Кукушкина О.В., Поликарпов А.А., Суровцева Е.В. (Под ред. В.В.Дубичинский). Электронный корпус текстов художественных произведений А.П.Чехова: принципы организации и возможности лексикографического использования// Слово и словарь. Vocabulum et vocabularium. Сборник научных трудов по лексикографии. Вып. 12. Харьков- Клагенфурт, 2011. – С.216.

диахрон, умумийлигига (тузувчилар сонига) кўра *якка ва умумий муаллифли* корпус сифатида қаралади.

“Ўзбек тили муаллифлик корпусини тузишнинг хусусий лингвистик асослари (Абдулла Қаҳҳорнинг “Бемор” ҳикояси мисолида)” деб номланувчи III бобда Абдулла Қаҳҳор корпуси интерфейси хусусиятлари, ўзбек тилидаги сўзларни морфологик теглаш ва лингвистик моделлаштириш йўллари, семантик гуруҳ, тўда ва майдон таснифининг матнни семантик теглашдаги аҳамияти масалалари таҳлил этилади.

“Абдулла Қаҳҳор корпуси интерфейсининг ўзига хос хусусиятлари” бўлими асосини муаллифлик корпуси тузишнинг умумий тамойиллари асосида Абдулла Қаҳҳор асарлари корпуси тузиш тамойилларини ишлаб чиқиш ташкил этади. А.Қаҳҳор асарлари корпусини тузишда дунё муаллифлик корпуслари тажрибаси, ўзбек тили ҳамда А.Қаҳҳор ижодининг ўзига хос томонларига алоҳида эътибор бериш талаб қилинади. Муаллифлик корпуси таркибий қисмларининг тўғри жойлаштирилиши, корпус мазмуни, материали, теглар тизимининг мукаммаллиги фойдаланувчи/тадқиқотчи учун муҳим омиллар бўлганлиги сабабли Абдулла Қаҳҳор корпусини тузишнинг муҳим босқичлари сифатида корпус интерфейсини лойиҳалаш (1), корпусга материал танлаш (2), материални қайта ишлаш (3), корпусни теглашнинг умумий тамойилларини ишлаб чиқиш (4), корпуснинг морфологик, семантик, синтактик теглар мажмуини ишлаб чиқиш (5), корпуснинг дастурий таъминотини яратиш (6) кабиларни алоҳида санаб ўтиш жоиз. Корпус интерфейси дизайни, тузилиши, имкониятларини ишга солувчи қидирув ойналари, уларнинг таркиби, корпус материаллари корпус лингвистикаси мутахассиси томонидан тайёрланади, корпус дастурий таъминоти эса дастурловчи зиммасида қолади. Корпуснинг энг асосий белгиси уни мунтазам мукаммаллаштириб, тўлдириб бориш, бу хусусият корпусни бошқа электрон маҳсулотлардан ажратиб туради. Бунинг учун корпуснинг асосий базаси у тузилган вақтда тўғри режалаштирилган бўлиши лозим. А.Қаҳҳор асарлари корпуси учун А.Қаҳҳор асарларининг беш жилдлик мукаммал нашри асосий манба бўлади. Тузилиши лойиҳалаштирилган А.Қаҳҳор корпусидан ҳам электрон кутубхона, ҳам матнни қайта ишлай оладиган, сўровга жавоб берадиган тегланган корпус сифатида фойдаланиш мумкин.

“Ўзбек тилидаги сўзларни морфологик теглаш ва лингвистик моделлаштириш йўллари” бўлимида леммалаш ҳамда тегнинг лингвистик модели ўрганилди. Ўзбек тилидаги матнларни морфологик теглашнинг ўзига хос тамойилларини ишлаб чиқиш объект сифатида танланган “Бемор” ҳикоясини теглаш учун назарий асос вазифасини бажара олади. Морфологик тег тизимига *сўзшакл, лемма* ва *тег* киради. Сўзшакл – танланган матндаги морфологик бирлик. Сўзшаклни теглашнинг биринчи босқичи уни леммалаш, яъни сўзшаклнинг лексема шаклини келтириш. Флектив тилларни теглашда энг қийин босқич – бу леммалаш (лемматизация), яъни сўзнинг лексема шаклини сўзшаклга тег сифатида бириктириш. Чунки флектив тилларда сўзшаклдаги грамматик маъно сўз ўзагига қоришиб кетган бўлади. Флектив тиллардан фарқли ўлароқ, агглютинатив тилда леммалаш жараёни

анча осон. Сўзшаклнинг грамматик шаклсиз қисми (ўзак ёки негиз) леммага тенг. Тегда лемма <*> белгиси ичида берилади. Барча сўз туркумларида леммалаш шу асосга, яъни "сўзнинг ўзак-негиз қисми леммага тенг" тамойилига асосланилса, феъл туркумида феъл-лемма II шахс буйруқ-истак майли шаклида берилади. Луғатларда феълнинг ҳаракат номи шаклида берилиши одат тусига кирган: <бормоқ>. Аммо бу шакл корпус учун мос эмас, чунки корпусдаги матнда сўзнинг <бормоқ> шакли эмас, <бор> шакли қидирилади. Шунга асосланиб, феъл-лемма **ўқитди** <ўқи>, **бўлмади** <бўл>**, кўрсатди** <кўр>**, олди <ол>** шаклида берилади. Феъл-иборанинг лемма шаклини "Фразеологик луғат"даги каби *-моқ* қўшимчаси билан бериш тўғрироқ, чунки *кўзи тиниб* каби қурилишли иборани <кўзи тин> шаклида леммалашда ҳеч қандай маъно йўқ. Фойдаланувчи ибора қидирганда морфологик эмас, балки семантик тег натижасидан фойдаланади. Шунга асосланиб, иборанинг **кўзи тиниб <**кўзи тинмоқ>, **боши айланадиган** <боши айланмоқ> тарзида тегланиши мақсадга мувофиқ. Бошқа барча сўзлар <*> белгиси ичида ўзак-негиз шаклида лексема ҳолида ёзилади: **саҳарга** <саҳар>. Сўз туркумларида ўзакнинг соддалашиш масаласи ҳам бор. Табиийки, соддалашиш жараёнида бўлган ўзак шу ҳолида лемма шаклига келтирилади[15]. От сўз туркумидаги луғавий шакл лемма таркибига киритилмайди, чунки бу шакл лексик маънога таъсир қилмайди, тегда **қизча** <қиз> кўринишида акс этади. Айрим сўзлар бундан мустасно. Чунончи, *боғча, шолча, кўрпача* сўзлари шу ҳолида леммага тенг, чунки улардаги *-ча* кичрайтириш шакли эмас. Теглаш жараёнида ҳар бир сўзшаклга 5тадан 10тагача, баъзан ундан ҳам кўпроқ морфологик тег (изоҳ) ёзиш талаб қилинади. Тегларни лингвистик моделлаштириш мақсадга мувофиқ, чунки лингвистик моделда морфологик тег шартли қисқартма шаклини олади. Ҳар бир сўз туркумини теглаш учун махсус лингвистик модел шакллари ишлаб чиқилади.

Феъл туркумини теглашнинг лингвистик моделини ишлаб чиқишда феълнинг барча грамматик категориялари ҳисобга олинади; феъл-леммага сўзшаклга қараб шу теглардан тегишлиси ёзилади. *Теглар тизими кўриниши:* **чўзилади** <чўз> [ф], [муст. ф.], [ҳар. ф.], [ў-сиз ф.], [б-ли ф.], [ўзл. н.], [сод. ф.], [т.ф.], [х.м.], [к.з.], [III ш.б.]

От туркумидаги сўзни леммалаш, теглашнинг дастлабки босқичида унинг умумий грамматик маъносига асосланилади[16]. Борлиқдаги бир турдаги предметдан бири ёки шу турдаги предметларнинг умумий номини билдиришига кўра атоқли ва турдош от маъновий гуруҳи ҳамда унинг тури аниқланади. Атоқли от *теглари:* **Абдуғанибой** <Абдуғанибой> [от], **[ат. от.], [шахс н.],** [б. к.], [бирл. с.] кўринишида бўлса, турдош от **паркда <парк>** [от], **[тур. от], [ан.о.], [ў.ж.н.],** [я.о.], [с.о.], [т.о.], [ў.п.к.], [бирл. с.] тарзда тегланади.

---

[15] "Бемор" ҳикояси матнида бундай сўз учрамаганлиги учун бу масала таҳлилдан четда қолади.

[16] Sayfullayeva R. va b. Hozirgi o`zbek adabiy tili. O`quv qo`llanma. – Тошкент: Fan va texnologiya, 2009. – Б 367.

Сифат туркумидаги сўзни леммалаш ўзак-негизни <*> белгиси билан сўзшаклдан ажратиш билан амалга оширилади, лингвистик модели “сифат = [сиф.]” шаклида белгиланади, **йироқ <**йироқ> [сиф.]. Сифат-сўзшаклни теглашда от туркумидаги каби умумий грамматик маъно белгиланади: *Теглар тизими:* **йироқ <**йироқ> **[сиф.], [ас. с.],** [о.д.], [с.с.], [т.с.], [хус. ЛМГ].

Олмош туркуми теглар мажмуи унинг маъновий гуруҳлари, тузилиши, ясалиши, келишик, эгалик, сон категорияси каби морфологик белгилардан ташкил топади. *Теглар тизими:* **шу <шу>** [олм.], [кўр.олм.], [с.олм.], [туб олм.], [б.к.], [бирл. с.].

Сон сўз туркуми теглари лингвистик модел сифатида соннинг маъновий гуруҳи, тузилишига кўра тури тавсифидан иборат бўлади. *Теглар тизими:* **25** <йигирма беш> [с.], [сан. ЛМГ], [мур.с.].

Равиш туркуми тег тизимида равишнинг ЛМГи, тузилиши, деривацияси каби категориялари мавжуд бўлади. Равиш-сўзшаклга қуйидаги теглар бириктирилади: **ҳозир** <ҳозир> [рав.], [п. рав.], [с. рав.], [туб рав.].

Ёрдамчи сўз туркумлари лингвистик моделлари боғловчида **аммо <аммо>** [боғ.], [соф боғ.], [зид. боғ.]; кўмакчида **билан <билан>** [кўм.], [соф кўм.], [вос.м.]; юкламада **ҳатто <ҳатто>** [юк.], [куч. юк.], [соф юк.] кўринишида бўлади.

“Семантик гуруҳ, тўда ва майдон таснифининг матнни семантик теглашдаги аҳамияти” бўлимида семантик тег муаммоси ўз ечимини топган. Атоқли от *семантик тег тизими*: **Сотиболдининг** <Сотиболди> [от], [ат. от], [ш.н.]. Турдош от *семантик тег тизими:* **осмон <**осмон> **[от],** [тур. от], [ан.от], [ў.ж.н.], **[4.3.**]. Ажратиб кўрсатилган тег сўзшаклнинг “Фазо. Фазовий ҳолат. Шакл” микромайдонига мансублигини ифодалайди. **табибга <табиб>** [от], [тур. от], [ан.о.], [ш. н.], [касб ЛМГ], **[11.I.3].** Ажратиб кўрсатилган тег “касб-ҳунар” микромайдонига тегишлиликни билдиради.

Феъл туркумининг семантик тег тизими − морфологик тегларга қўшимча изоҳ. *Феъл туркумининг семантик тег тизими:* **оғриб қолди** <оғриб қол> [ф], [муст. ф.], [ҳолат ф.], **[11.I.h], [14.2].** Ажратиб кўрсатилган тег “соғлиқ микромайдони” ва “инсонга хос ҳаракат, ҳолатни ифодаловчи ЛМТ”га мансублик ҳақида маълумот беради. Равиш сўз туркумининг семантик тег тизимида равишнинг маъно гуруҳлари маълумот (тег) вазифасини бажаради.

“Синтактик теглашнинг лингвистик асослари” бўлимида синтактик тег – матннинг синтактик таҳлилига асосланган теглар мажмуи, морфологик таҳлилга асосланилган парсинг натижаси эканлиги, тегнинг бу кўриниши лексик ва бошқа турли синтактик қурилмалар (содда гап, қўшма гап, кўчирма гап ва ҳ.), бирликлар орасидаги синтактик алоқани кўрсатиши[17] ҳақида гапирилган. Матн синтактик тег тизимининг энг катта ахборот базасини ташкил этувчи изоҳ – гап қурилишига оид маълумотлар йиғиндиси. Гап синтаксиси гапни қайси жиҳатдан ўрганса, теглаш жараёнида шу белгиларнинг барчасини қамраб олиш жоиз. Гап билан боғлиқ энг биринчи

[17] Захаров В.П., Богданова С.Ю. Корпусная лингвистика. – Иркутск: ИГЛУ, 2011. – С.93.

тег гапнинг тузилишига кўра турини ифодаловчи изоҳдан иборат. Тег кўриниши:

1. ***<СГ>*** *Табиб қон олди****.</СГ>***

2. ***<ҚГ>****Абдуғанибой унинг сўзини эшитиб кўп афсусланди, қўлидан келса ҳозир унинг хотинини оёққа бостириб беришга тайёр эканини билдирди, кейин сўради:* ***</ҚГ>.*** Корпус менежерига “содда гапни топиш” ёки “қўшма гапни топиш” буйруғи берилса, танланган матннинг барча СГ ёки ҚГлари натижалар ойнасида акс этади.

Гапнинг ифода мақсадига кўра турини изоҳлаш ҳам теглашнинг муҳим босқичини белгилайди: “дарак гап” = <дг>, “сўроқ гап”= <сг>, “буйруқ гап”= <бг>. Бу теглар жуфт бўлишига ҳожат йўқ. Якка ҳолда қўлланилганда ҳам ахборот бера олади. Чунки гапнинг тузилишига кўра тури гапни чегаралаган, шу чегара охирида бу маълумотни ҳам қўшиш мумкин. Тег тизими: *<СГ>Девонаи Баҳоваддинга ҳеч нарса кўтардингми?* ***<сг>,*** *</СГ>*

Гап лисоний қурилишида эганинг ифодаланиш ёки ифодаланмаслигига кўра “эгали гап” = <Е+>, “эгасиз гап” = <Е->; эгасиз гапларнинг “шахси номаълум гап” = <ш.н.г>, “атов гап” = <а.г>, “семантик-функционал шаклланган гап” = <с.фш.г> каби белгиларни қўшиш мумкин. Тег тизими: *<СГ>* ***“****Кўнгилга армон бўлмасин” деб “чилёсин” ҳам қилдиришга тўғри келди.* <дг>, **<Е->, <ш.н.г>,** *</СГ>*.

Бош ва иккинчи даражали бўлакнинг иштирокига кўра “йиғиқ гап” = <йг>, “ёйиқ гап” = <ёг> каби изоҳ ҳам тег сирасидан ўрин олади. Тег тизими:

1. *<СГ>Бемор оғирлашди*. <дг>, <Е+>, **<йг>,** *</СГ>*.

2. *<СГ>Шаҳарда битта докторхона бор*. <дг>, **<ёг>,** <Е+>, *</СГ>*.

Гап билан грамматик алоқага киришмайдиган бўлакларнинг мавжудлиги ҳам изоҳ талаб қилади ҳамда “ундалма” = <у>, </у>, “киритма” = <к>, </к> каби белгилар мажмуидан иборат бўлади; Бундай теглар жуфт қўлланилади. Чунки бирликнинг чегарасини кўрсатиш лозим бўлади. *Тегланган кўриниши:*

1. <СГ> Буларнинг ҳаммаси, **<к>** албатта, **</к>** пул билан бўлади. </СГ>

2. <СГ>, **<у>** Худоё **</у>** аямди дайдига даво бейгин <бг>, <Е+>, <ёг>, </СГ>.

Қўшма гап тегланиши:

1. <Б-сиз ҚГ> *Аллақандай бир хотин келиб толнинг хипчини билан савалади, товуқ сўйиб қонлади...* **</Б-сиз ҚГ>.**

2. <БҚГ> *Сотиболди хўжайинининг олдига арзга борди, аммо бу боришдан муддаоси нима эканини аниқ билмас эди.* **</БҚГ>**

3. <ЭҚГ >, </ЭҚГ> . “Бемор” ҳисоясида ЭҚГ учрамаганлиги сабабли мисол келтирилмади.

Кўчирма гапнинг изоҳланиши учун қуйидаги теглар керак бўлади:

1) кўчирма гап = <КГ>, </КГ>;
2) муаллиф гапи = <МГ>, </МГ>.

Тег кўриниши: **<КГ>** *Бегуноҳ гўдакнинг саҳарда қилган дуоси ижобат бўлади, уйғотинг қизингизни!* **</КГ> - <МГ>** *деди*. **</МГ>.**

Хулоса сифатида А.Қаҳҳор корпусига материал танлашнинг муҳим масала эканлиги: ёзувчи ҳаёти, ижодига оид материал йиғиш, А.Қаҳҳорнинг мукаммал асаллар тўпламидаги барча асарларини саралаш, ижодкорнинг ҳаётига оид мақола ва хотираларни танлаш кераклигини қайд этиш лозим. Корпусда бундай маълумотларнинг мавжудлиги ундан электрон кутубхона сифатида фойдаланиш имконини берса, корпус матнларининг тегланиши А.Қаҳҳор асарлари матни устида турли лингвистик амалларни бажаришда кўп функцияли ахборот манбаи бўлади.

ХУЛОСА

1. Корпус лингвистикаси тилшуносликнинг жадал ривожланаётган соҳаси, корпус эса тилшуноснинг зарурий иш қуроли; оғзаки, ёзма ёдгорликлар, миллий-маданий меросни акс эттирувчи ахборот манбаидир. Қидирув дастурига бўйсундирилган матнлар йиғиндиси, мукаммал тегга эга корпус лингвистик тадқиқотлар самарадорлигини таъминлашда барқарор лингвистик база вазифасини бажара олади. Корпус − табиий (реал) тилнинг электрон шаклдаги, маълумотни табиий контекстуал шаклида сақлайдиган, тил ҳодисаларини ҳар томонлама, чуқур, объектив ўрганишга, электрон кутубхона, луғат, грамматикалар яратишга асос бўладиган тизим. У тилни тадқиқ этиш(сўзнинг ўзгариши, историзм, неологизмнинг вужудга келиши, маъно кенгайиши, торайиши, янги фразеологизмларнинг пайдо бўлишини кузатиш)да, лингводидактика ва луғат тузишда замонавий, кенг имкониятли дастурлаштирилган тизимдир.

2. Тил корпусига энг кўп эҳтиёж сезувчи соҳа – матнни автоматик қайта ишлаш дастури (корпус орфографик, грамматик тузатишларни автоматик равишда бажаради), қидирув тизимига эга бўлган, турли функцияни бажарувчи дастурлар ҳамда таржима дастури. Корпуснинг ижтимоий аҳамияти кенг қамровли. У тилшунос, таржимон, ўқитувчи, дастурчи, журналист, муҳаррир, умуман, кундалик фаолиятида сўз билан иш кўрадиган ҳар қандай киши учун замонавий ахборот воситаси.

3. Корпуснинг пайдо бўлиши, ривожланиши икки босқич – компьютер асригача бўлган давр ва компьютер асри корпуслари даврига бўлинади. Компьютер асрига келиб у электрон тус олди. Корпус лингвистикасига 1960 йилларда Браун корпусини яратиш билан асос солинган; рус корпус лингвистикасининг тамал тоши 1980 йилларда Швециянинг Упсала университетида қўйилган. Ўтган даврда жаҳон корпус лингвистикаси катта ютуқларга эришди; бугунги кунда ҳам мунтазам ривожланиб боряпти.

4. Корпус турли томондан таснифланади: бу таснифлар қанча кўп бўлмасин, бутун тилга ёки унинг маълум бир воқеланиш турига оидлигига ҳамда лингвистик тег турига кўра гуруҳлаш мақсадга мувофиқ. Шунингдек, унинг ёзма, оғзаки, аралаш шакли; мультимодал, махсус матнлар корпуслари каби турлари ҳам бўлади. Махсус матнлар корпуси ҳажман катта бўлмайди, алоҳида тадқиқий топшириққа хизмат қилади, тузувчи режасига мувофиқ яратилади. Махсус корпус Миллий корпусдан шу жиҳати билан фарқланади.

5. Корпус тузиш тамойилларини ишлаб чиқишда асосий муаммо– корпусни лойиҳалаштириш, теглаш, унга мос қидирув тизими – корпус менежери/конкордансни танлаш. Корпус тузишнинг энг аҳамиятли босқичи − теглаш; корпусдан фойдаланиш имконининг кенг/торлиги тег даражаси ва турига боғлиқ: мукаммал тег – кенг имкониятли, универсал корпус гарови. Корпус тузиш амалиётида теглаш дастлаб қўлда қилинган, кейинчалик автоматик теглаш дастурлари (парсинг, таггинг) яратилган. Корпус тузишда дастурий таъминотнинг энг муҳим қисми парсинг ва таггинг дастурларидир.

6. Сунъий интеллектнинг ўзбек тилини “ўқиш, тушуниш, қайта ишлаш, унга мос жавоб қайтариш”га қодир дастурини яратиш, “ўзбек тили”ни теглай оладиган парсинг, таггинг дастурларини ишлаб чиқиш – дастурчилар олдида турган долзарб вазифа. Бундай дастурларсиз ўзбек корпус лингвистикаси ривожлана олмайди, чунки корпус лингвистикаси – тилшунос ҳамда дастурчининг ҳамкорлигида иш кўрадиган соҳа.

7. Корпус мақсадига кўра тадқиқий ёки иллюстратив бўлиши мумкин. Тадқиқий корпус тузувчиси маълум лингвистик муаммони ечиш учун ўз корпусини яратади ва шу асосда олиб бораётган тадқиқоти учун хулоса чиқаради; иллюстратив корпус муаллифи эса корпусни кенг фойдаланувчиларнинг турли амалларни бажаришига мўлжаллаб тузади. Шунга асосланиб айтиш мумкинки, муаллифлик корпуслари (тузувчи мақсадига қараб) икки хил: *тадқиқий* ва *иллюстратив* мақсадда тузилиши мумкин. Корпусга турли аспектда ёндашар эканмиз, муаллифлик корпусини шу параметрлар асосида қуйидагича тавсифлаш мумкин: муаллифлик корпуси турғун/динамиклигига кўра *турғун*, тўлиқ ёки фрагмент кўринишли бўлишига кўра *тўлиқ матнли.* Чунки муаллифлик корпуси, миллий корпусда бўлганидек, доимий янгиланишни талаб қилмайди: бир марта тузилади, тўлдириш, тузатишга эҳтиёж қолмайди. Муаллифлик корпусида бир муаллиф ижоди тўла қамраб олинганлиги туфайли (ижодий меросидаги асар матнлари тўлиқ олинади) *тўлиқ матнли* саналади. Шунингдек, муаллифлик корпуси нутқ турига кўра *ёзма*, параллеллигига кўра *бир тилли*, матннинг ихтисослашувига кўра *аралаш* (ижодкорнинг барча жанрдаги асари киритилади), кириш усулига кўра (корпус тузувчисининг ихтиёрига қараб) ҳар хил: *пуллик, эркин, ёпиқ* бўлиши мумкин. Изоҳ хусусиятига кўра *морфологик, синтактик, семантик тегланган*, хронологик жиҳатига кўра *диахрон*, умумийлигига (тузувчилар сонига) кўра *якка ва умумий муаллифли* бўлиши мумкин.

8. Муаллифлик корпусининг энг оптимал варианти серқирра тег, мукаммал интерфейсга эга кўринишидир. Мукаммал ишланган интерфейс фойдаланувчига тушунарли, ишлашга қулай. Муаллифлик корпуси ҳосиласи сифатида автоматик равишда частотали, терс, идеографик луғат яратиш мумкин.

9. Муаллифлик корпуси ва муаллифлик лексикографияси – чамбарчас боғлиқ ҳодиса: бирининг тараққиётини иккинчисининг ривожисиз кўриб бўлмайди. Муаллифлик лексикографияси йўналишида анча ишлар қилинган бўлса ҳам, муаллифлик корпуси тараққиёти турғун ҳолатда. Мукаммал

тузилган, чуқур тегланган муаллифлик корпуси лексикология, лексикография, тил тарихи, диалектология, социолингвистика, психолингвистика, нейролингвистика соҳалари ривожига ҳисса қўшади, бу йўналишлардаги тадқиқотлар учун янги имкониятлар яратади. Шунингдек, муаллифлик корпуси ёрдамида маълум ёзувчи/шоир тили конкордансини тузиш, маълумотлар базасини яратиш ҳам мумкин.

10. Муаллифлик корпусини тузишда глобал тармоқда мавжуд бўлган рус тилидаги А.П.Чехов, А.С.Пушкин, Ф.М.Достоевский корпуслари, инглиз тилидаги Шекспир, немис тилидаги Гёте корпуси, форс-тожик адабиёти вакиллари Саъдий, Румий, Фирдавсий, Лойиқали Шерали муаллифлик корпуслари намуна бўла олади. Санаб ўтилган муаллифлик корпуслари билан танишиб, уларнинг ўзига хос, фарқли, ўхшаш томонларини ўрганиш натижасида корпус тузиш методикаси, концепциясини ишлаб чиқиш, корпус учун манба танлаш, корпусни теглаш, луғат мақоласи тузилиши, лексемани тавсифловчи белги (бош сўз, қисқача изоҳ, грамматик характеристика, частота); сўзшаклнинг грамматик белгиси, сўз қўллаш ҳолати характеристикаси (контекст) каби масалалар муаллифлик корпуси тузишнинг асосий муаммолари эканлиги аниқланди. Муаллифлик корпуси тузишнинг умумий тамойиллари асосида Абдулла Қаҳҳор асарлари корпуси тузиш принциплари ишлаб чиқилиши керак. Бундай корпусни тузишда дунё муаллифлик корпуслари тажрибаси, ўзбек тили ҳамда А.Қаҳҳор ижодининг ўзига хос жиҳатларига алоҳида эътибор қаратиш талаб этилади. Муаллифлик корпуси таркибий қисмларининг тўғри жойлаштирилиши, корпус мазмуни, материали, корпус материалини теглаш ҳал қилувчи омиллар бўлганлиги учун Абдулла Қаҳҳор корпусини тузишнинг муҳим босқичлари корпус интерфейсини лойиҳалаш (1), корпусга материал танлаш ва уни қайта ишлаш (2), корпусни теглаш тамойилларини ишлаб чиқиш (3), морфологик, семантик, синтактик тегларни моделлаштириш (4), дастурий таъминот яратиш(5)дан иборат бўлади.

11. Абдулла Қаҳҳор асарлари корпусини тузиш учун морфологик, семантик, синтактик теглаш тамойилларини ишлаб чиқиш лозим, чунки ўзбек тилини “тушунадиган” – автоматик теглай оладиган дастур мавжуд эмас. Корпус интерфейси дизайни, тузилиши, қидирув ойналари, уларнинг таркиби, корпус материали корпус лингвистикаси мутахассиси томонидан, дастурий таъминот эса дастурчи томонидан тайёрланади. А.Қаҳҳор корпусига материал танлаш ҳам муҳим масала: ёзувчи ҳаёти ва ижодига оид материал йиғиш, А.Қаҳҳорнинг мукаммал асарлар тўпламидаги барча асарларини саралаш, ижодкорнинг ҳаётига оид мақола, хотираларни танлаб олиш лозим. Корпусда бундай маълумотларнинг мавжудлиги ундан электрон кутубхона сифатида фойдаланиш имконини берса, корпус матнларининг тегланиши А.Қаҳҳор асарлари матни устида турли лингвистик амалларни бажариш, адабиётшунослик, тарих, этнография, лингвомаданиятшунослик, лингвомаънавиятшуносликка оид тадқиқот олиб боришда кўп функцияли ахборот манбаи сифатида хизмат қила олади. Бунинг учун корпуснинг асосий базаси тузилаётган вақтда тўғри режалаштирилган бўлиши лозим. “А.Қаҳҳор

асарлари корпуси” учун А.Қаҳҳор асарларининг беш жилдлик нашри асосий манба бўлади.

12. “А.Қаҳҳор асарлари корпуси”ни тузиш учун намуна сифатида танланган матн − “Бемор” ҳикояси матнини теглашнинг лингвистик ва экстралингвистик тег моделлари тузилади. Танланган объект – “Бемор” ҳикояси матни тегланади: сўзшакл леммалаштирилади, яъни матндаги сўзшаклнинг лексема шакли белгиланади, сўзшаклга лингвистик изоҳ – теглар бириктирилади; морфологик, семантик ва синтактик теглар гуруҳланади, матн шундай гуруҳлар асосида тегланади.

НАУЧНЫЙ СОВЕТ PhD.30.05.2018.Fil.70.01
ПО ПРИСУЖДЕНИЮ УЧЕНЫХ СТЕПЕНЕЙ ПРИ
КАРШИНСКОМ ГОСУДАРСТВЕННОМ УНИВЕРСИТЕТЕ

БУХАРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ХАМРОЕВА ШАХЛО МИРДЖОНОВНА

ЛИНГВИСТИЧЕСКИЕ ОСНОВЫ СОЗДАНИЯ АВТОРСКОГО КОРПУСА УЗБЕКСКОГО ЯЗЫКА

10.00.01 – Узбекский язык

АВТОРЕФЕРАТ ДИССЕРТАЦИИ ДОКТОРА ФИЛОСОФИИ (PhD)
ПО ФИЛОЛОГИЧЕСКИМ НАУКАМ

Карши − 2018

Тема диссертации доктора философии (PhD) по филологическим наукам зарегистрирована в Высшей аттестационной комиссии при Кабинете Министров Республики Узбекистан за № B2018.3.PhD/Fil504.

Диссертация выполнена в Бухарском государственном университете.

Автореферат диссертации размещен на трех языках (узбекский, русский, английский (резюме)) в веб-странице Научного совета karshi.uz и в информационном образовательном портале «ZiyoNet»

Научный руководитель:	**Менглиев Бахтиёр Ражабович** доктор филологических наук, профессор
Официальные оппоненты:	**Муродова Нигора Кулиевна** доктор филологических наук, профессор
	Каримов Суюн Амирович доктор филологических наук, профессор
Ведущее учреждение:	**Ургенчский государственный университет**

Защита диссертации состоится «_____» 2018 года в _____ часа на заседании Научного совета PhD.30.05.2018.Fil.70.01 при Каршинском государственном университете (Адрес: 180103, город Карши, улица Кучабог,17. Тел.:(0375) 225-34-13; факс: (0375) 221-00-56; e-mail: qarshidu@umail.uz).

С диссертацией можно ознакомиться в Информационно-ресурсном центре Каршинского государственного университета. (зарегистрирован под номером _____). (Адрес: 180103, город Карши, улица Кучабог,17. Тел.:(0375) 225-34-13; факс: (0375) 221-00-56; e-mail: qarshidu@umail.uz). Каршинский государственный университет, зал заседаний факультета узбекской филологии.

Автореферат диссертации разослан «_____» _____года.
(протокол реестра № ____от _____2018 года).

Н.Н.Шодмонов
председатель Научного совета по присуждению научных степеней, доктор филологических наук

Г.Н.Тожиева
ученый секретарь Научного совета по присуждению ученых степеней, доктор философии (PhD)

Д.Тураев
председатель научного семинара при Научном совете по присуждению ученых степеней, доктор филологических наук, профессор

ВВЕДЕНИЕ (аннотация диссертации доктора философии (PhD)

Актуальность и востребованность темы исследования. Изучение проблем компьютерной и корпусной лингвистики в мировом лингвистике началось в начале 40-х годов XX века. В связи с этим в этой области были созданы первые научные труды. В частности, в 60-х годах прошлого века этот процесс ускорился, появились сотни лингвистических корпусов, которые отразили миллионы слов. Искусственный интеллект обладает потенциалом автоматического перевода, компьютерного анализа, редактирования, тезауруса, электронного словаря, научных и теоретических основ, первых примеров, которые были применены на практике. Эти обновления привели к появлению перспективных научных тенденций, связанных с лингвистическим использованием информационных технологий. Он определяет необходимость изучения общих принципов корпуса, формирования корпуса, развития, текущего состояния и создания корпуса, а также актуальности темы.

В XXI веке мировая лингвистика выросла в научно-теоретическом изучении корпусной лингвистики. В целях повышения качества автоматического перевода в области передовой компьютерной лингвистики лингвистическое моделирование языков, создание алгоритма лингвистического слова и использование многоязычного национально-культурного наследия определенного языка стали важной проблемой в мировой лингвистике. Компьютерная лингвистика, в частности создание корпусов, расширение размеров существующих корпусов и разработка программ автоматической обработки текста, являются одной из основных проблем, связанных с лингвистикой.

За годы независимости был проведен ряд исследований в области компьютерной лингвистики для автоматического перевода, искусственного интеллекта и понимания узбекского языка, но корпусная лингвистика полностью не изучалась в монографическом плане. Поэтому в лингвистике, а также во всех научных областях “...всесторонняя поддержка научных и творческих исследований и создание необходимых для них условий”[1] указывает на необходимость углубленных исследований по интеграции науки. Внимание к языку в нашей стране было одним из приоритетов нашего внимания к духовности. Поэтому наряду с сохранением, обогащением и эффективностью его использования широкое применение узбекского языка в современной информационно-коммуникационной системе стало неотложной задачей. В этой связи одной из актуальных задач нашего исследования является изучение таких вопросов, как создание национального корпуса узбекского языка, создание лингвистических основ

[1]Мирзиёев Ш.М. Эркин ва фаровон, демократик Ўзбекистон давлатини биргаликда барпо этамиз. Ўзбекистон Республикаси Президенти лавозимига киришиш тантанали маросимига бағишланган Олий Мажлис палаталарининг қўшма мажлисидаги нутқ. – Тошкент: Ўзбекистон, 2016. – Б.13.

авторского корпуса и лингвистических моделей в качестве перспективного научного направления корпусной лингвистики.

В определённой степени данная диссертация решает задачи, которые предусмотрены в следующих законах и нормативно-правовых актах как Указ Президента Республики Узбекистан от 13 мая 2016 года № УП – 4797 “Об организации Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои”, Указ Президента Республики Узбекистан от 7 февраля 2017 года № УП – 4947 “О стратегии действий по дальнейшему развитию Республики Узбекистан”, Постановление Президента Республики Узбекистан от 17 февраля 2017 года № ПП – 2789 “О мерах по дальнейшему совершенствованию деятельности Академии наук, организации, управления и финансирования научно-исследовательской деятельности”, Постановление Президента Республики Узбекистан от 13 сентября 2017 года № ПП-3271 “О программе комплексных мер по развитию системы издания и распространения книжной продукции, повышению культуры чтения”.

Соответствие исследований приоритетами научно-технического развития Республики Узбекистан. Исследование проводилось в соответствии с приоритетным направлением республиканского научно-технического развития: I. “Информационное общество и демократическое, социальное, правовое, экономическое, культурное, духовное и образовательное развитие демократического государства”.

Уровень изучения проблемы. В мировом лингвистике авторский корпус стала предметом обучения в 60-х годах прошлого века. Ещё в 60-х годах Р.Г.Петровским было сказано, что “Достоверная лингвистическая информация может быть получена из большого массива текстов”[2], но целевые исследования в области корпусной лингвистики начинались ещё в 40-е годы[3] Блумфильдом, Фрицем и Бонджерсом. Создатели Корпуса Брауна (1961-1964), Нильсон Фрэнси Генри Кучера вначале разработали принципы создания корпуса; работы Джона Синклера, автора Английского банка (1980), заслуживают особого внимания[4]. В русской лингвистике В.Захаров, А.К.Кутузов, Е.В.Недошивина, В.Рыков, В.Плунгян изучали корпус, его разновидности, особенности, социальную значимость корпуса и принципы построения корпусов[5]. Специальные исследования по делам об

[2] 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.

[3] 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.

[4] Френсис Н., Кучера Г. Вычислительный анализ современного американского варианта английского языка. – М., 1967.; Синклер Дж. Предисловие к книге “Как использовать корпуса в преподавании иностранного языка”/ http://www/ruscorpora.ru/corpora-infro.html, свободный.

[5] Кутузов А.Б. Корпусная лингвистика. − (Электрон ресурс): Лицензия Creative commons Attribution Share-Alike 3.0 Unported - //lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf; Недошивина Е.В. Программы для работы с корпусами текстов: обзор основных корпусных менеджеров. 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 Национальный корпус русского языка? “Отечественные записки” 2005, №2. http://magazines.russ.ru/oz/2005/2/2005_2_20-pr.html

авторском корпусе можно найти в работах О.Кукушкиной, А.Поликарпова и Е.Суровцевой[6].

В узбекской лингвистике проведено изучение компьютерной лингвистики, лексикографической обработки текста и лингвистического анализа. Мы можем также отметить труды А.Пулатова, С.Мухамедова, М.Айымбетова, С.Мухамедовой, С.Каримова, Г.Жуманазаровой, А.Бабанарова, Д.Уринбаевой, Н.Абдурахмановой, А.Норова и других исследователей. Эти выводы особенно актуальны в контексте инновационного подхода к текстологическим исследованиям-внедрения современных лексикографических и лингвостатических методов компьютерных достижений, но к сожалению ни один из них не включен в повестку дня создания узбекских лингвистических корпусов. Однако стоит отметить, что они были столпами создание корпусов национального языка[7].

Связь дисциплинарной деятельности с планом исследований, проводимым высшим учебным заведением или исследовательским институтом. Диссертация было выполнено в рамках фундаментального проекта Ф-1-06 “Синтез литературных традиций Востока и Запада в узбекской литературе в период независимости”.

Цель исследования. Целью исследования является изучение создание лингвистические основы корпуса узбекского языка, изучение языковые ценности лингвистического корпуса; изучение истории корпусной лингвистики и авторского корпуса, их особенностей в социальной, лексикологической, образовательной и других сферах.

Задачи исследования:

изучение авторского корпуса, его формирование и развитие, общее состояние корпусной лингвистики и состояние мировых корпусов;

изучение лингвистических основ создания авторского корпуса;

определение общих и конкретных аспектов авторского корпуса;

---

[6] Кукушкина О.В., Поликарпов А.А., Суровцева Е.В. (Под ред. В.В.Дубчинский) Электронный корпус текстов художественных произведений А.П.Чехова: принципы организации и возможности лексикографического использования// Слово и словарь. Vocabulum et vocabularium. Сборник научных трудов по лексикографии. Вып. 12. – Харьков-Клагенфурт, 2011. – 416 с.

[7] Муҳаммедов С.А. Статистический анализ лексико-морфологической структуры узбекских газетных текстов: Автореф. дисс… канд. фил. наук.- Ташкент, 1980; Бабанаров А. Разработка принципов построения словарного обеспечения турецко-русского машинного перевода: Автореф. дисс… канд. фил. наук. - Л., 1981; Айымбетов М.К. Опыт лингвостатистического анализа лексики и морфологии каракалпакского публицистического текста: Автореф. дисс… канд. фил. наук.- Ташкент, 1987; Каримов С., Қаршиев А., Исроилова Г. Абдулла Қаҳҳор асарлари тилининг луғати. Алфавитли луғат. Частотали луғат. Терс луғат, - Тошкент, 2007; 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; Жуманазарова Г.У. Фозил Йўлдош ўғли достонлари тилининг лингвопоэтикаси: Фил. фан. док. дис…автореф. - Тошкент, 2017; Абдурахмонова Н.З. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (Содда гаплар мисолида): Филол.фан.бўйича фалсафа доктори (PhD)…дис. афтореф. – Тошкент, 2018; Пўлатов А. Компьютер лингвистикаси. – Тошкент: Академнашр, 2011.; Норов А. Компьютер лингвистикаси асослари. – Қарши, 2017. – 136 б.

изучение опыта авторского корпуса по произведениям А.П.Чехова, А.С.Пушкина, Ф.М.Достоевского, А.С.Грибоедова, У.Шекспира, Фирдоуси, Руми, Саъди Шерази, Лойик Шерали;

разработка принципов авторского корпуса;

изучение характерных особенностей интерфейса корпуса Абдуллы Каххара;

на примере произведений Абдуллы Каххара, обоснование морфологической характеристики и лингвистического моделирования слов;

объяснение семантической группы, важность семантической разметки при классификации текста;

создание лингвистических основ синтаксической разметки на примере произведений Абдуллы Каххара.

Объект исследования. В качестве объекта исследования выбраны корпус, виды корпусов и авторский корпус.

Предметом исследования составляют интерфейс корпуса Абдуллы Каххара, единицы измерения корпуса и проблемы теггирование.

Методы исследования. В освещении темы исследования применены методы классификационного, статистического, сравнительного, когнитивного анализа.

Научная новизна исследования заключается в следующем:

объясняется важность использования авторского корпуса в лингвистике и в лингводидактике, способность обрабатывать и использовать централизованную информацию и возможность использования авторского корпуса;

было отмечено что создания корпуса развивалось в три этапа, первый этап в форме электронной библиотеки, второй этап обработка текста, третий этап история лингвистических корпусов русского и английского языков, современное состояние корпусной лингвистики, роль современного русского, английского, турецкого и таджикского корпусов в мировом лингвистике, также было обнаружена отличительная черта в количестве слов и в возможности лингвистического анализа;

изучены технологический процесс проектирования и построения корпуса, также определено, что тег определяет обработку текста в корпусе, маркировку грамматических фраз слова, поиск, сортировку и фильтрацию корпусного менеджера, а также его BONITO, XAIRA, SARA, SQR, DDS и другие типы;

обоснованы научные, практические, образовательные цели и задачи авторского корпуса, которые отличаются от национального, параллельного, образовательного, мультимодального корпуса, структуры, состава авторского корпуса, сходные и отличительные аспекты авторского корпуса на русском и английском языках;

разработаны принципы проектирования, теггирования для авторского корпуса и поисковой системы (corpus manager).

Практические результаты исследования заключаются в следующем:

разработан интерфейс корпуса Абдуллы Каххара;

объясняются методы морфологического, семантического, синтаксического и лингвистического моделирования узбекского языка;

проведены морфологическое, семантическое, синтаксическое теггирование в тексте рассказа “Бемор” Абдуллы Каххара и обобщенны результаты;

создан проект авторского корпуса Абдуллы Каххара;

создан словарь терминов корпусной лингвистики.

Достоверность результатов исследования связана с тем, что изученные материалы основаны на проверенных ресурсах, которые помогают делать выводы, основанные на характере узбекского языка, их обоснованности, методологическом совершенстве и принципах целостности.

Научно-практическое значение результатов исследований. Исследование имеет теоретическое значение в развитии теоретических основ создания национальных и авторских корпусов на узбекском языке. Его практическое значение имеет тот факт, что одна из отраслей прикладной филологии – роль корпусной лингвистики в преподавании предмета как науки, основа написания и написания программ и планов, основа для создания различных типов и других авторских корпусов на узбекском языке.

Внедрение результатов исследования. Основными принципами научно обоснованного корпуса в исследовании являются: технологический процесс проектирования и создание корпуса, важность тега при создании корпуса и лингвистических средств, поиск, отбор, фильтрация корпусного менеджера; исходя из его характеристик:

создан “Терминологический словарь корпусной лингвистики” по результатам сравнения общих и разных аспектов современного русского, английского, турецкого и таджикского языков среди мирового корпуса, использованы сбор и совокупность информации в лингводидактики и в лингвистике авторского корпуса в узбекской лингвистике, теоретические основы, обработка информации, теоретические сведения о практических и образовательных преимуществ, нынешнее состояние корпусной лингвистики (Справка Министерства высшего и среднеспециального образования № 89-03-3647 от 26 октября 2018 года). На основе научных результатов была достигнута улучшению описании терминов корпусной лингвистики;

были использованы результаты анализа и выводов диссертации о словоформах, их лингвистического моделирования и анализа производного слова в лингвистическом корпусе в фундаментальном проекте под номером ОТ-Ф8-062 на тему “Деривационные закономерности языкового развития” (Справка Министерства высшего и среднеспециального образования № 89-03-3647 от 26 октября 2018 года). Результаты исследования были использованы для объяснения феноменов деривации;

научные результаты о технологические процессы проектирования и построения корпуса, обозначение тега обработки текста в корпусе,

отображение грамматических признаков помеченных слов; поиск, сортировка и фильтрация корпусного менеджера и описание его видов BONITO, XAIRA, SARA, SQR, DDS; сравнение авторского корпуса от электронного словаря и электронной библиотеки; результаты проектирования, теггирование и выбора критериев отбора менеджеров использованы в фундаментальном исследовательском проекте на тему “Краткий словарь узбекского, русского и английского глоссариев национальных прикладных и изобразительных искусств” который проведен в 2014-2016 годах в Национальном институте искусств и дизайна им. Камолиддина Бехзода в рамках государственной научно-технической программы под номером ЁА1-ФҚ-0-07289 (Справка Министерства высшего и среднеспециального образования № 89-03-3647 от 26 октября 2018 года). В результате лингвистического моделирование лемм достигнута научно-популярности краткого пояснительного словаря терминов прикладного и художественного искусства и словарь наполнен новыми источниками.

разработан интерфейс корпуса, по моделям лингвистической разметки был создан off-line вариант фрагмента авторского корпуса (Справка Министерства высшего и среднеспециального образования № 89-03-3647 от 26 октября 2018 года, сертификаты №000895 и № 000986 IP-центр при Агентстве интеллектуальной собственности Республики Узбекистан). В результате был создан фрагмент off-line версии Корпуса Абдуллы Каххара.

Апробация результатов исследования. О результатах исследования были опубликованы статьи в 2-международных и 5-республиканских научно-практических конференциях.

Публикация результатов исследований. По теме диссертации было опубликовано 16 трудов. Из них 1 терминологический словарь, 2 авторское свидетельство, 13 статей в изданиях предложенных Высшей аттестационной комиссией Республики Узбекистан для публикации докторских диссертаций, из которых 4 в зарубежных журналах.

Объем и структура диссертации. Диссертация состоит из введения, трех глав, заключения, списка использованной литературы и приложения. Общий объем диссертации составляет 165 страницы.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Вводный раздел основан на актуальности темы, описывает цели, задачи исследования, объект и предмет, соответствует приоритетам развития национальной науки и техники, научных и практических результатов, научно-практическое значение результатов, предоставляется информация о результатах введение в практику, об опубликованных работах и о структуре диссертации.

В первой части диссертации названной **“Формирование, развитие и теоретические основы корпусной лингвистики”** рассматривается корпус, корпусная лингвистика, его формирование, развитие и современное состояние.

В первой части этой главы, названной **“О корпусе и корпусной лингвистике”** прокомментированы корпусная лингвистика и её предмет.

Описание в научной литературы обобщается на идее о том, что языковой корпус представляет собой сборник текстов разных жанров, разновидностей, региональных и социальных версий определенного языка на основе специального программного обеспечения[8]. Многие языки мира имеют свой национальный корпус, который отличается уровнем передового опыта и способностью обрабатывать текст, так как сегодня языковые корпуса стали незаменимым инструментом для решений практических задач и исследований современной лингвистики. Корпус отличается от простой электронной библиотеки тем, что имеет аннотацию корпуса, конкорданс (относительно простой внешний вид поисковой системы) или корпусного менеджера, также освещены возможности поиска и общие требования к корпусному менеджеру. Освещена роль корпусной лексикографии, лексикологии, синтаксиса, изучение методологии лингвистики, лингводидактики, в обучения родного и иностранного языка.

Следует отметить, что национальный корпус языка как источник информации имеет следующие преимущества:

1) образцы словесного, письменного, национального и культурного наследия, созданного на родном языке, появляются в Интернете в электронном виде;

2) корпус формируется, когда естественный (реальный) язык представляет собой набор текстов, помещенных в электронном виде, в программу поиска, отлично скоординированный корпус выполняет функцию стабильной лингвистической базы в обеспечении эффективности лингвистических исследований;

3) корпус является основой для создания электронных библиотек, словарей, грамматики. Поскольку он обширен, он обеспечивает целостность информации и обеспечивает полную демонстрацию всех аспектов языковых событий;

4) в языковом корпусе различная информация представляется в виде естественной контекстной формы, которая является основой их всестороннего, глубокого, объективного изучения;

5) корпус – является многофункциональной запрограммированной системой, которая исследует язык (изменение слова, появление историзма, неологизма, увеличение и уменьшение смысла, изучение появления новых фразеологизмов), изучает язык, а также используется при создании словарей.

В разделе **“Формирование и развитие корпусной лингвистики”** передаётся информация об 4 основных периодов формирования лингвистического корпуса до и после компьютерного века: в 1960, 1970, 1980, 2000 гг. появившиеся корпуса, структуры, их свойства, история английской и русской корпусной лингвистики. Появление и развитие

[8] 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.

корпуса состоит из двух этапов: первый этап до компьютерной революции, второй этап эра компьютерной революции. Первый период корпуса – это сборник картотек, хотя и не в форме современного корпуса, а как материал для лингвистических исследований. А в компьютерном веке они вошли в электронную форму, запрограммированы. Частичная корпусная лингвистика была построена авторами английского корпуса (Браунского корпуса) в 60-х годах прошлого века. Для разных вариантов доступно более 100 английских версий. Полная структура корпусной лингвистики в русской лингвистике была заложена в 1980-х годах в Швеции – в Упсальском университете.

В разделе **“Современное состояние корпусной лингвистики”,** выделены такая специфика как виды корпусов, репрезентативные свойства, жанры которые входят в его состав, письменный, устный, многоязычный корпус; сбалансированные, отличительные черты корпусов специальных текстов, которые не столь велики и служат отдельным исследовательским заданиям. Хотя корпус имеет множество аспектов, таких как его структура, цель, устойчивая изменчивость, желательно сгруппировать их с обеих сторон:

а) в зависимости от языка (обычно определенного периода языка) или определенного типа поведения (жанр, стиль, социальная или возрастная группа, писатель или научный язык);

б) согласно типу лингвистической разметки.

В главе II под названием **“Общелингвистические принципы создания авторского корпуса”** рассматриваются общие принципы корпусной лингвистики, лингвистические основы авторского корпуса, общие и конкретные аспекты. Раздел **“Общие принципы создания корпуса”** описывает технологический процесс проектирования и создания корпуса.

В.П.Захаров, С.Ю.Богданова внесли хронологическую задачу в качестве важного аспекта процесса проектирования корпуса. Например, что означает “современный корпус языка”? Естественно во многих жанрах хронологическая граница корпуса совсем иная. Однако из-за того, что изображения, содержащиеся в тексте, не относятся к языковому материалу, очень важно уделить внимание к тексту, который был в составе корпуса, таблицы соответственно разработать под корпус, а также нужно обратить внимание на цитаты, фразы, термины и единицы измерения.

Некоторые эти проблемы решаются на этапе проектирования, некоторые из них решается при процессе создания и использовании корпуса. Было установлено, что создание корпуса состоит из следующих этапов:

1. Анализы, тексты, полученные из разных источников при первоначальной обработке текста, подлежат филологической проверке, редактированию.

2. В процессе преобразования в графическом анализе некоторые тексты проходят процесс повторного кодирования, удаляются или заменяются не текстовые части (изображений, таблиц). Строка в тексте перекрывается, границы (в тексте MS-Word) прерываются, дубликаты и другие символы становятся однозначными. В графическом анализе текст, который входит в

корпус, делится на части (слово, словосочетание) и не текстовые элементы удаляются.

3. Идентификация нестандартных (не лексических) элементов, создание специального текстового элемента (аббревиатуры, имена, фамилии), операции поиска по одному и тому же критерию, например, заглавие текста, название, описание, список прокрутки и т.д. автоматически выполняется текстовым редактором.

В данном разделе особое место уделяется важности разметки в создании корпуса как лингвистическое средство, автоматическая разметка текста, поисковой системы – менеджеру корпуса.

Тег делится на лингвистические и экстралингвистические виды. Различают следующие типы экстралингвистического тега[9]:

1. Отражающая своеобразность текстового формата (глава, параграф, разделы и т. д.).

2. Тег формулирующий текст и информацию принадлежащую автору

Лингвистическая разметка имеет морфологические, синтаксические, семантические, анафорические и просодические виды. Разметка (аннотация) корпуса выполняется запрограммированным способом. В этом случае, прежде всего, обращают внимание на экономию времени и сокращении работ, во-вторых, решается проблема автоматической обработки текста. Так как пока анафорическая, просодическая разметка остается сложной, разметка выполняется только вручную, конечно в будущем весь этот процесс будет запрограммирован. Хотя морфологическая и синтаксическая разметка выполняется с помощью теггера и парсинга, многие из этих программ требуют дополнительной автоматической коррекции. Например, в случае морфологической неоднозначности (в более формальных языках) в контексте синтаксического мультимедиа программа предлагает несколько выводов, а исследователь выбирает правильный. Чрезвычайное увеличение размеров корпусов следующего поколения делает задачу для профессионалов переключиться на полностью автоматизированный вид, ставит такие задачи как создание совершенно нового теггера и парсинга.

При автоматическом морфологическом анализе (таггирование) каждая лексическая единица (словосочетание, лемма, граммоидная группа) имеет отдельный грамматический характер (личные местоимения, падеж или другая грамматическая категория).

По словам В.П.Захарова все типы (морфологический, синтаксический, семантический, анафорический, просодический) выполняются по следующим принципам:

1) описание (обоснование) схемы разметки;

2) определение системы общих лингвистических понятий;

3) формирование известной схемы анализа для пользователя;

4) достижение теоретической традиции схемы разметки;

5) соответствие международным стандартам[10].

[9] http://rykov-cl.narod.ru/c.html.

Важнейшей частью лингвистического корпуса является его поисковая система – менеджер корпуса, текст и система лингвистического контроля. Корпусный менеджер – специализированная поисковая система, предназначенная для работы с информацией корпуса, статистической информацией, программным обеспечением, которая отображает результаты поиска удобным для пользователя способом. балки сўз бирикма ҳолидаги сўровга ҳам жавоб бера олиш;

Е.Недошивина рассматривает требования к менеджеру корпуса как к корпусу, наиболее важным из которых является список ключевых слов текста, создание полного противоречивого списка; не только слово, но и словосочетания должны отвечать на вопрос; получать результаты (список результатов) на основе нескольких критериев; отображать нежелательное количество запросов; предоставлять статистическую информацию об отдельных элементов корпуса; полностью понять морфологический характер леммы, слова и метаинформации (библиографической, типологической), основанной на корпусном диалекте; сохранять и распечатать результаты; работа с неограниченным размером файла; быстрый поиск, вывод результатов; читать и писать различные текстовые форматы (txt, doc, rtf, html, xml и т.д.); а также удобство для профессионального и нового пользователя[11].

Таким образом, создание корпуса начинается с разработки его принципов; важно сосредоточиться на проектировании, разметки и поиске подходящего менеджера поисковых систем. В корпусе разметка имеет большое значение, потому что широкая или узкая полоса пропускания, используемая корпусом, зависит от разметки. Идеальная разметка – залог широкомасштабного и универсального корпуса. При создании корпуса программы разметки – парсинг, таггинги считаются важнейшими инструментами, демонстрирующие, то что современные программисты сталкиваются с проблемой разработки программ, способных “читать, понимать, обрабатывать, реагировать на искусственный интеллект” узбекского языка. Искусственный интеллект “не знает” узбекского языка, он должен иметь специальные парсинг, таггинг программы которые следует разработать для того чтобы его “понять”.

Во II главе в разделе **“Лингвистическая основа создания авторских корпусов”** основное внимание уделяется целям и задачам авторского корпуса, идеографической разметки и интерфейсу корпуса. Интерфейс корпуса – очень важный компонент функциональности, который знакомит пользователя со всем его содержимым. Интерфейс авторского корпуса – это отличная возможность того что с первого взгляда можно понять жанровую классификацию творческого наследия автора. Важной частью авторского корпуса является интерфейс, независимо от его совершенства.

---

[10] Захаров В.П., Богданова С.Ю. Корпусная лингвистика.– Иркутск: ИГЛУ, 2011. – С.76.

[11] Недошивина Е.В. Программы для работы с корпусами текстов: обзор основных корпусных менеджеров. работа с системой DDC.// Языковая инженерия: в поиске смыслов. – (Электронный ресурс): https//docplayer.ru.

Идеально разработанный интерфейс – это удобный и простой в использовании фактор. Еще одна ключевая особенность авторских корпусов это их вид и степень разметки. Если появляется возможность создания частотного, обратного словаря на основе разметки, то это послужит основой идеографического тега произведений автора для создания идеографического словаря.

Раздел **“Общие и специфические аспекты авторского корпуса”** посвящается исследованию авторской лексикографии и отношениям авторских корпусов, их особенность, структура, состав и обобщённость, схожесть и отличительные черты авторских корпусов А.С.Пушкина, Ф.Д.Достоевского, А.Грибоедова. Авторская лексикография занимается проблемами теоретических основ авторского словаря, опытом изучения авторской идиомы, особенностям создания авторского электронного словаря, опыта использования национального и литературного языка в авторской лексикографии, идеями создания новых словарей, сравнительно-аналитическое изучение современных литературных текстов на основе авторской лексикографии[12].

На основе творческой работы А.П.Чехова, изучены особенности авторского корпуса: структура, состав и функция условной информационно-исследовательской системы корпуса, называемой “ИСТОК”[13], ее основная функция и ее возможности, на основе корпуса Чехова. Выяснилось, что был *создан грамматико-семантический словарь художественных произведений А.П.Чехова*. Совокупность словаря и электронного корпуса – это новый вид лексикографического продукта. Важность этого продукта определяется новым стилем, удобством словаря, возможностью пользователя выполнять различные независимые исследования по языковым признакам впоследствии[14]. Авторский корпус тесно связан с авторской лексикографией. Поэтому развитие лексикографии авторского корпуса в двусторонних отношениях с корпусной лингвистикой: невозможно рассмотреть прогресс одного без развития другого. К сожалению, в мировой лингвистике трудно отметить эти два направления лингвистики как развитая область; несмотря на то, что много работ по лексикологии авторского корпуса, по-прежнему остаются не изученными, с которыми мы сталкиваемся сегодня. Разработка авторских корпусов находится в стабильном состоянии. Поскольку расширение возможностей корпуса не влияет на развитие авторских корпусов. Существующие авторские корпуса (Корпус европейскиxx языков) очень просты в теггировании. Достижения русской лингвистики в этом отношении амбициозны. Корпуса великих писателей, поэтов, таких как А.П.Чехов, А.Пушкин, Ф.Достоевский,

[12] www.ruslang.ru/seminar_aut_lexocorg020413 сайтидаги “Теория и практика авторской лексикографии” семинар материаллари.

[13] “ИСТОК” – “Исследование Словаря, Текстовых Особенностей, Конкордансов”

[14] Кукушкина О.В., Поликарпов А.А., Суровцева Е.В. (Под ред. В.В.Дубичинский). Электронный корпус текстов художественных произведений А.П.Чехова: принципы организации и возможности лексикографического использования// Слово и словарь. Vocabulum et vocabularium. Сборник научных трудов по лексикографии. Вып. 12. Харьков- Клагенфурт, 2011. – С.216.

У.Шекспир в Британском национальном корпусе, Корпус Гете в Мюнхенском университетском зале и Корпус персидско-классической литературы, является моделью для дальнейшего развития. Исследования по авторскому корпусу привели к следующим выводам:

корпус является следственным и иллюстративным в соответствии с целью; авторский корпус также может быть создан для обеих целей. Исследователь создает свой собственный корпус для решения конкретной лингвистической проблемы и делает выводы для своих исследований; автор иллюстративного корпуса предпочитает, чтобы корпус широко использовался разными пользователями в разных видах деятельности (преподавание задач, использование тематических исследований, перевод, при создании текста использование в качестве лингвистической базы). Таким образом, авторский корпус представляет собой форму *исследования* и *иллюстративного* корпуса;

по виду изменения корпуса корпус делится на *устойчивый* и *динамичный*, авторский корпус является устойчивым корпусом, представляет собой не подвижный корпус по фрагменту, считается *полным текстовым корпусом*. Нет необходимости постоянно обновлять копию, как в случае с национальным корпусом; он создается один раз, его не нужно пополнять и редактировать. В качестве полнотекстового корпуса признается, что авторский корпус полностью включает в себя всё творчество автора, и всё его творческое наследие (полный текст работы) вводится. По виду речи письменная, по параллели единичный по специализации текста смешанный (поскольку включается творчество автора разного жанра), в зависимости от метода входа (по усмотрению автора корпуса) может быть разной: платной, свободной, закрытой;

по виду пояснения авторский корпус может быть морфологический, синтаксический, семантический, теггированный, по размеру текста полнотекстовый, диахронный в соответствии с хронологическим аспектом рассматривается как единый и общий авторский корпус в соответствии с совокупностью.

Третья глава **“Лингвистические основы авторского корпуса узбекского языка (на примере рассказа “Бемор” Абдуллы Каххара)”,** показывает особенности интерфейс корпуса Абдуллы Каххара, морфологическое и лингвистическое моделирование слов в узбекском языке, семантическую группу, и важность их роли в семантической разметки.

Раздел **“Особенности интерфейса корпуса Абдуллы Каххара”** основан на принципах создания авторских корпусов Абдуллы Каххара на основе общих принципов создания корпуса. При создании корпуса произведений А.Каххара особое внимание уделяется опыту мировых авторских корпусов, узбекскому языку и особенностям произведений А.Каххара. Правильное расположение основных компонентов авторского корпуса, содержание корпуса, материал доведённый до совершенства система тегов является основным фактором для пользователя/исследователя. В связи с этим в качестве создания авторского корпуса Абдуллы Каххара

основными этапами являются проектирование интерфейса (1), подбор материала к корпусу (2), переработка материала (3), разработка основных принципов теггирования корпуса (4), разработка морфологического, семантического, синтактического комплекса тегов (5), создание программного обеспечения корпуса (6). Дизайн интерфейса корпуса, состав, поисковые окна с различными возможностями, материал корпуса подготавливается специалистом корпусной лингвистики, а программное обеспечение корпуса подготавливается программистом. Основной признак корпуса это постоянное совершенствование, дополнение, это особенность является отличительной чертой корпуса от других электронных средств. Для этого основная база корпуса должна правильно спроектирована во время создания корпуса. Для корпуса произведений А.Каххара его пятитомник послужит очень хорошим источником. Отлично спроектированный корпус А.Каххара может использован как электронная библиотека и в качестве лингвистического корпуса.

В разделе **“Способы лингвистического моделирования и морфологической разметки слов в узбекском языке”** изучается лемматизация или лингвистическая модель тега. Рассказ “Бемор” выбранный в качестве объекта разметки может быть пособием для разработки принципов создания морфологической разметки текстов в узбекском языке. В систему морфологической разметки входят словоформа, лемма и тег. Словоформа – морфологическая единица в заданном тексте. Первая степень разметки словоформы это лемматизация, то есть подбор лексемы словоформы. При разметки флективных языков самый сложный процесс это лемматизация, то есть форму лексемы слова нужно соединить в качестве тега. Потому что во флективных языках грамматический смысл словоформы взаимосвязан с корнем слова. Но по сравнению с флективными языками лемматизация агглютинативных языков достаточно простой процесс. Бесформенная часть грамматической словоформы – корень равен лемме. В разметке лемма находится внутри знака <*>.

Во всех словосочетаниях лемматизация проводится таким образом, то есть про принципу “корень слова равен лемме”, тогда в категории глагола глагол-лемма представляется в форме 2-го лица повелительного наклонения. В словарях вошло в привычку предоставлять глагол в форме имени действия. Например: **<бормоқ>.**

Однако эта форма не подходит для корпуса, потому что текст в корпусе ищет форму **<бор>** вместо **<бормоқ>.** Исходя из этого, глагол-лемма задается в словах как ўқитди **<ўқи>**, бўлмади **<бўл>**, олди **<ол>**. Во “Фразеологическом словаре” глагол-лемма вставляется как фраза в виде лемматизации. При поиске стандартной фразы он использует результат семантической разметки, а не морфологической. Например, при использовании окончания **-моқ** образуется правильное предложение, потому-что, при лемматизации выражения *кўзи тиниб* мы не видим ни какого смысла в выражении <кўзи тин>. Следовательно, лучше всего преобразовать эту фразу как **<кўзи тинмоқ>, боши айланадиган <боши айланмоқ>.**

Все остальные словосочетания вводятся в лексеме как корневые в символе **<*>**: **саҳарга <саҳар>.** Также в случае слов есть проблема упрощения корня. Естественно, что в процессе оптимизации корень приводится в виде леммы[15]. Языковая форма категории подлежащего не включена в лемму, потому что эта форма не имеет лексического значения, и она представлена в разметке **қизча** в форме **<қиз>**. За исключением некоторых слов. Например, слова **боғча, шолча, кўрпача** равны лемме, поскольку использованное окончание **-ча** не является уменьшительной формой. В процессе разметки каждое слово требуется от 5 до 10, иногда даже больше морфологических тегов (пояснений). Лингвистическое моделирование тегов желательно, поскольку лингвистическая модель содержит морфологический тег с условной аббревиатурой. Разработаны специальные лингвистические формы моделей для каждого слова.

При разработке разметки категории глагола лингвистической модели для назначения анализа учитываются все грамматические категории глагола; в словесной лемме записываются слова, принадлежащие этим тегам. Вид разметки: **чўзилади** <чўз> [ф], [муст. ф.], [ҳар. ф.] , [ў-сиз ф.], [б-ли ф.], [ўзл. н.], [сод. ф.], [т.ф.], [х.м.], [к.з.], [III л.ед ч.]

На начальном этапе теггировании при лемматизации имени существительного основываются на его грамматическое значении[16]. **Имена существительные собственные** – это слова, которые представляют названия единичных предметов. Например*: Александр Сергеевич Пушкин* – имя, отчество и фамилия одного из писателей. **Имена существительные нарицательные** – это слова, которые являются названием большой группы однородных предметов (одушевленных или неодушевленных). Например: словом *писатели* называют большую группу людей, которые создают книги. Разметки имён существительных собственных имеет следующий вид: **Абдуғанибой** <Абдуғанибой> [сущ], **[имя сущ. соб.], [имя.],** [б.к.], [един. ч.], имя существительное нарицательное имеет следующий вид: **паркда <парк>** [сущ], **[нар. имя. сущ], [ан.о.], [ў.ж.н.],** [я.о.], [с.о.], [т.о.], [ў.п.к.], [един. ч.]

Определение в части речи формируется лемматизацией корня символом <*> и отделяет его от слово формы. Для лингвистической модели установлено значение определения “сифат = [сиф.]”: **йироқ <**йироқ> [сиф.]. В утверждении качества определяется общий грамматический смысл. Вид разметки: **йироқ <**йироқ> **[сиф.], [ас. с.],** [о.д.], [с.с.], [т.с.], [особ. ЛМГ].

Теггирование местоимения составляется по морфологическим значениям категории числительных, подлежащих, по падежу, по созданию, по строению и по смысловой группе: вид разметки: структура, формация, происхождение, принадлежность, конечная категория, добавляются как теги.

В части речи числительное теги представляются как лингвистические модели, которые состоят из группы родов характеризующих виды чисел. *Вид разметки*: **25** <йигирма беш> [с.], [сан. ЛМГ], [мур.с.].

---

[15] Поскольку в тексте рассказа “Бемор” не имеется такие слово, этот вопрос выходит за рамки анализа.

[16] Sayfullayeva R. va b. Hozirgi o`zbek adabiy tili. O`quv qo`llanma. – Тошкент: Fan va texnologiya, 2009. – Б.367.

В части речи наречие разметка наречия имеет категории такие как состав, деривация. Наречие слово форма присоединяется к следующим тегам: **ҳозир** <ҳозир> [рав.], [п. рав.], [с. рав.], [туб рав.].

Вспомогательные частицы речи представлены в виде лингвистических моделей, в виде **аммо <аммо>** [боғ.], [соф боғ.], [зид. боғ.]; **билан <билан>** [кўм.], [соф кўм.], [вос.м.]; **ҳатто <ҳатто>** [юк.], [куч. юк.], [соф юк.]

Была решена семантическая разметка "семантической группы, значение семантической группировки, значение группы и классификация текста в смысловом понимании". Семантическая разметка имена существительные собственные имеет следующий вид: **Сотиболдининг** <Сотиболди> [от], [ат. от.], [ш.н.]. Семантическая разметка имена существительные нарицательные имеет следующий вид: **осмон <**осмон> **[от],** [тур. от], [ан.о.], [ў.ж.н.]. Выделенный тег "Сфера. Состояние сферы. Форма" имеет связь с микро площадью **табибга <табиб>** [от], [тур. от], [ан.о.], [ш.н.], [касб ЛМГ]. Выделенный тег имеет прямую связь с микроплощадью "ремесло".

Семантическая разметка глагола – это добавление к тегам морфологической разметки. Семантическая разметка глагола имеет вид: **оғриб қолди** <оғриб қол> [ф], [муст. ф.], [ҳолат ф.], **[11.I.h], [14.2].** Выделенный тег имеет отношение к "микроплощади здоровья" и "связан с действиям и состоянию человека". Дополнение имеет особое место в семантической разметки тега.

В разделе **"Лингвистические основы синтаксической разметки"** синтаксическая разметка – совокупность тегов основанная на синтаксическом анализе текстов, является результатом парсинга на основе морфологического анализа, лексический вид формы разметки и другие синтаксические устройства (простое предложение, сложное предложение и др.) имеют связь между синтаксическими единицами[17]. При построении предложения совокупность информации – это пояснения самой полной информационной базы текстового синтаксиса. Когда речь идет о синтаксисе речи, важно охватить все эти символы. Первый тег, связанный с предложением это описание описывающее тип состава предложения. Вид разметки:

3. ***<СГ>*** *Табиб қон олди.****</СГ>***

4. ***<ҚГ>****Абдуғанибой унинг сўзини эшитиб кўп афсусланди, қўлидан келса ҳозир унинг хотинини оёққа бостириб беришга тайёр эканини билдирди, кейин сўради:* ***</ҚГ>.*** В поиске корпуса если задается команда "найти простое предложение" или "найти сложное предложение" то тогда будут продемонстрированы все простые и сложные предложения в тексте.

Один из тегов также поясняет предложения по виду цели: **"дарак гап" = <дг>, "сўроқ гап" = <сг>, "буйруқ гап"= <бг>** – **"повест-вовательное предложение" = <пов.п>, "вопросительное предложение" = <в.п>, "побудительные предложения" = <поб.п>.** Данные теги необязательно должы быть двупарными. Могут использоваться и отдельно то есть

[17] Захаров В.П., Богданова С.Ю. Корпусная лингвистика.– Иркутск: ИГЛУ, 2011. – С.93.

информация будет отображаться. Потому что каждое предложение имеет границу, и в конце этой границы можно дополнить информацию. Вид разметки: *<СГ>Девонаи Баҳоваддинга ҳеч нарса кўтардингми?* ***<сг>,*** *</СГ>*

Предложение по принципу построения бывает назывным и неопределенно-личным.

Назывное предложение – это односоставное предложение с главным членом-подлежащим. В назывных предложениях сообщается о существовании и наличии предмета. **Неопределенно-личные предложения** – тип **предложений**, главным членом которых является предикат в форме 3л. мн.ч. настоящего, будущего времени, в форме мн.ч. прошедшего времени и сослагательного наклонения, обозначающий действие или состояние неназванного личного субъекта. Вид разметки: *<СГ>* ***"****Кўнгилга армон бўлмасин" деб "чилёсин" ҳам қилдиришга тўғри келди.* <дг>, **<Е->, <ш.н.г>**, *</СГ>*.

Предложения, имеющие, наряду с главными, второстепенные члены, называются **распространенными** "распространённое предложение" = <рп>,**.** Нераспространенным "нераспространенное предложение" = <нп>, называется предложение, состоящее только из главных членов – подлежащего и сказуемого. Вид разметки:

1. *<ПП>Бемор оғирлашди. <дг>, <Е+>,* ***<йг>,*** *</СГ>*.

2. *<ПП>Шаҳарда битта докторхона бор. <дг>,* **<ёг>**, <Е+>, *</СГ>*.

Наличие неграмматических элементов также требует объяснения такие как например: "обращение" = <о>, "вставка" = <в>. Тогда этот тег используется в паре. Потому что нужно объяснить границу единиц. Вид разметки:

1. *<ПП> Буларнинг ҳаммаси,* ***<к>*** *албатта,* ***</к>*** *пул билан бўлади. </ПП>*

2. <ПП>, ***<у>*** *Худоё* ***</у>*** *аямди дайдига даво бейгин <бг>, <Е+>, <ёг>, </ПП>*.

Вид разметки сложного предложения:

4. <Б-сиз ҚГ> *Аллақандай бир хотин келиб толнинг хипчини билан савалади, товуқ сўйиб қонлади...* **</Б-сиз ҚГ>.**

5. <БҚГ> *Сотиболди хўжайинининг олдига арзга борди, аммо бу боришдан муддаоси нима эканини аниқ билмас эди.* **</БҚГ>**

6. <ЭҚГ >, </ЭҚГ>. Так как рассказ "Бемор" не имеет НСП, примеры не показаны.

7. Прямая речь – это дословное воспроизведение чужого высказывания. Для ее передачи используются специальные синтаксические конструкции, которые состоят из двух компонентов: слов **автора** и собственно **прямой речи**.

8. Косвенная речь – это пересказ чужого высказывания. Для ее оформления используется один из типов придаточного предложения – конструкция с придаточным изъяснительным. Главная часть таких предложений строится *от имени автора текста* и соответствует словам

автора при прямой речи, а придаточная часть передает содержание *высказывания и соответствует прямой речи.*

Косвенная речь имеет следующие теги:

3) косвенная речь = <КР>, </КР>;

4) прямая речь = <ПР>, </ПР>.

Вид разметки: **<КР>** Бегуноқ гўдакнинг саҳарда қилган дуоси ижобат бўлади, уйғотинг қизингизни! **</КР> - <ПР>** деди. **</ПР>.**

Следует отметить, что для корпуса А.Каххара важнейшая задача это отбор нужного материала: сбор писательской жизни, творчество, отбор всех работ коллекции шедевров А.Каххара, отбор статей и мемуаров жизни писателя. Если наличие такой информации в корпусе даёт возможность использования в качестве электронной библиотеки, то теггирование текстов в корпусе произведений А.Каххара послужит для многих многофункциональным источником, информации для различных лингвистических операций над текстом.

ЗАКЛЮЧЕНИЕ

1. Корпусная лингвистика – динамично развивающаяся отрасль лингвистики, так как корпус является основным инструментом лингвиста; информационный источник устного и письменного народного творчества национального культурного наследия. Содержание поисковой программы является надежной лингвистической базой для обеспечения эффективности лингвистических исследований корпусов с отличными разметками. Корпус – это система, основанная на естественном (электронном) языке, который поддерживает материал в контекстной форме, представляет собой всеобъемлющее, глубокое, объективное исследование языковых явлений, электронной библиотеки, словаря и грамматики. Исследование языка (изменение слова, историзм, неологизм, расширение и сужение смысла слова, наблюдение за появлением новых фразеологизмов) является широко запрограммированной системой при создании словарей в лингводидактике.

2. Самой сложной областью для языкового корпуса является автоматическая обработка текста (язык корпуса автоматически может выполнять орфографические, грамматические исправления), поисковые системы и программное обеспечение для перевода. Социальная значимость корпуса обширна. Это современный медиаинструмент для лингвистов, переводчиков, преподавателей, программистов, журналистов, редакторов и те кто работает в повседневной деятельности могут воспользоваться этим современным информационным средством.

3. Появление и развитие корпуса разделяется на два этапа – эпоха до компьютерного века и эпоха компьютерного века. Наконец в компьютерном веке он получил электронный вид. Корпусная лингвистика была основана в 1960-х годах с созданием корпуса Брауна; Корень русской корпусной лингвистики восходит к 1980-м годам в Упсальском университете в Швеции.

В это время мировая лингвистика достигла больших результатов; сегодня она также неуклонно развивается.

4. Корпус классифицируется по-разному: желательно классифицировать эти классы в соответствии со всем языком или определенным типом поведения и типом лингвистической разметки. Кроме того, имеются письменные, словесные, смешанные формы; мультимодальные, специальные текстовые корпуса. Корпус специальных текстов не так велик, он служит отдельным исследовательским заданиям и составляется в соответствии с планом автора. Специальный корпус отличается от Национального корпуса этой характеристикой.

5. Основная проблема принципов в построении корпуса – это проектирование корпуса, разметка, выбор подходящей поисковой системы – корпусного менеджера/конкорданса. Важнейшим этапом создания корпуса является разметка, широкая/узкая зависит от уровня и типа разметки: идеальная разметка – является гарантией универсального корпуса с широкими возможностями. В случае создании корпуса можно увидеть, что разметка первоначально была создана вручную, а затем создавались автоматические программы разметки такие как парсинг и таггинг. Наиболее важным инструментом в создании корпуса является программное обеспечение программ парсинга и таггинга.

6. Разработка искусственного интеллекта узбекского языка, способного “читать, понимать, обрабатывать, реагировать на него”, создавать программы, парсинг который может осуществлять разметку “узбекского языка”, также разработка таггинов является неотложной задачей для программистов. Без таких программ узбекская корпусная лингвистика не может развиваться, потому что корпусная лингвистика – это сфера, в которой сотрудничают лингвист и программист.

7. С точки зрения цели корпус делится на исследовательский и иллюстративный. Составитель создаёт свой исследовательский корпус для решения определенных лингвистических проблем, и тем самым делает выводы для своих исследований; автор иллюстративного корпуса создаёт корпус с широким кругом пользователей для выполнения различных действий. Можно сказать, что авторские корпуса могут быть созданы с двумя разными целями: для исследований и иллюстративных целей. Подход к корпусу с точки зрения разных аспектов даёт нам возможность характеризовать авторский корпус по следующим параметрам: авторский корпус по своей устойчивости и динамичности делится на такие виды: *устойчивый*, *полный* или по форме фрагмента *полнотекстовый*. Поскольку корпус не требует постоянного обновления, как в национальном корпусе: после его создания его не нужно дополнять и исправлять. Так как авторский корпус включает в себя полноценную работу автора (будут внесены полные тексты его творческого наследия) и он является полнотекстовым. Кроме того, по виду речи авторский корпус подразделяется на письменный, по параллели одноязычный, по специализации текста смешанный (потому что включены произведения автора различного жанра), по способу входа (в зависимости от

выбора автора) разный: *платный, свободный, закрытый*. По особенности пояснения *морфологический, синтаксический, семантический*, по хронологии *диахронный*, по обобщенности (по числу авторов) корпус может быть *единоавторным* и *многоавторным*.

8. Наиболее оптимальной версией авторского корпуса является появление превосходной разметки с отличным интерфейсом. Совершенно разработанный интерфейс упрощает использование и простоту использования. На основе авторского корпуса автоматически можно создать частотный, обратный и идеографический словарь.

9. Авторская лексикография и авторский корпус тесно связаны: невозможно увидеть прогресс первого без развития второго. Хотя авторская лексикография широко используется, развитие авторского корпуса остается стабильным. Идеально созданный, имеющий полную разметку авторский корпус можно использовать в таких сферах как лексикология, лексикография, история языка, диалектология, социолингвистика, психолингвистика, нейролингвистика, внести свой вклад в развитие этих областей также может создать новые возможности для проведения исследований. Можно создать конкорданс языка современных писателей или поэтов и также создать базу данных с использованием авторского корпуса.

10. В создании авторского корпуса помогут образцы таких корпусов как русскоязычный корпус А.П.Чехова, А.С.Пушкина, корпус Ф.Достоевского, английский корпус Шекспира, немецкий корпус Гёте, персидско-таджикский корпус Саади, Руми, Фирдоуси, Лойкали Шерали. Познакомившись с вышеупомянутыми авторскими корпусами и изучив их специфические, разные и аналогичные аспекты, методологию создания корпуса, концепцию корпуса, выбор источника корпуса, разметку корпуса, создание словаря, символ характеризующий лексему (заглавие, краткое содержание, грамматическая характеристика, частота); грамматические знаки слова, характер использования речи (контекст) вышеназванные задачи определили основные проблемы при создании авторского корпуса. Основываясь на общих принципах создания авторского корпуса, следует разработать принципы составления произведений Абдуллы Каххара. Необходимо обратить особое внимание на мировой опыт авторского корпуса, также на узбекский язык и сущность творчества А.Каххара. Проектирование интерфейса корпуса (1), выбор материала для корпуса (2), разработка принципов разметки корпуса (3), моделирование морфологических, семантических, синтаксических тегов разметки (4), разработка программного обеспечения является важнейшими этапами создания корпуса Абдуллы Каххара.

11. Необходимо разработать принципы морфологической, семантической, синтаксической разметки для создания корпуса произведений Абдуллы Каххара, поскольку нет программы, которая могла бы “понять узбекский язык” автоматически”. Дизайн интерфейса корпуса, структура, поисковые окна, их содержимое, материал корпуса разробатывается специалистом по корпусной лингвистике, а программное

обеспечение программистом. Важен выбор материала для корпуса А.Каххара: нужно собрать значимую информацию о жизни писателя, создание его произведений, подбор всех работ коллекции прекрасных произведений А.Каххара, статьи о жизни писателя, воспоминания. Наличие такой информации в корпусе предостовляет возможность использовать корпуса в качестве электронной библиотеки, а теггирование текстов корпуса произведения А.Каххара позволяет использовать корпуса как многофункциональный источник при исследованиях в литературоведении, истории, этнографии, лингвокультуролию и лингводуховности. По этой причине основная база корпуса должна быть надлежащим образом спланирована на момент ее создания. Пятитомное издание произведений А.Каххара надлежащим образом послужит основным источником для “Корпус произведений Абдуллы Каххара”.

12. При создании “Корпус произведений Абдуллы Каххара”, выбранный текст рассказа “Бемор” идеально подходит для образца при лингвистического и экстралингвистического теггирования для создания моделей. Выбранный объект – рассказ “Бемор” теггируется; то есть в тексте формируется лемма словоформы, лингвистическое пояснение словоформы – присоединяются теги; группируются морфологические, семантические, синтаксические теги и текст теггируется в такими группами тегов.

SCIENTIFIC COUNCIL PhD.30.05.2018. Fil.70.01 ON AWARD OF SCIENTIFIC DEGREE OF DOCTOR OF PHILOSOPHY AT KARSHI STATE UNIVERSITY

BUKHARA STATE UNIVERSITY

KHAMROEVA SHAHLO MIRDJONOVNA

LINGUISTIC FOUNDATIONS OF CREATING UZBEK LANGUAGE AUTHORSHIP CORPUS

10.00.01 – Uzbek language

DISSERTATION ABSTRACT FOR DOCTOR OF PHILOSOPHY (PhD) IN PHILOLOGICAL SCIENCES

Karshi − 2018

The theme of PhD dissertation is registered by Supreme Attestation Comission at the Cabinet of Ministry of the Republic of Uzbekistan under the number № B2018.3.PhD/Fil504

The dissertation has been prepared at Bukhara State University.

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

Scientific advisor:	**Mengliev Bakhtiyor Rajabovich** doctor of Philological sciences, professor
Official opponents:	**Murodova Nigora Kuliyevna** doctor of Philological sciences, professor
	Karimov Suyun Amirovich doctor of Philological sciences, professor
Leading organization:	**Urganch State University**

Defense of dissertation will be held on «___»____2018 at __ at the meeting of the Scientific Council number PhD.30.05.2018.Fil.70.01 at the Karshi State University. (Address: 180103, Karshi, Kochabog street, 17. (0 375) 225-34-13; fax: (0375) 221-00-56; e-mail: qarshidu@umail.uz). Karshi State University, the conference hall of faculty of Uzbek philology.

Doctoral dissertation can be found in the Information-Resource Center of the Karshi State University (registration number…). Address: 180103, Karshi, Kochabog street, 17. (0 375) 225-34-13; fax: (0375) 221-00-56; e-mail: qarshidu@umail.uz

Dissertation abstract sent out on «____» _____ 2018.
(Mailing report number ____ on «____» _____ 2018).

N.N.Shodmonov
chairman of Scientific Council awarding scientific degrees, Doctor of Philological sciences

G.N.Tojieva
secretary of Scientific Council awarding scientific degrees, Doctor of Philosophy (PhD)

D.Turaev
chairman of Scientific Seminar at the Scientific council awarding scientific degrees, Doctor of Philological sciences

INTRODUCTION (abstract of PhD thesis)

The aim of the research. In the research, corpus, its special peculiarity, its importance social, lexicological, educational and another fields, the history of corpus linguistic, types of corpus, learning linguistic value of authorship corpus, elaborating linguistic basis of creating the authorship corpus of Uzbek language were intended.

The assignments of the research.

To learn the corpus, corpus linguistic, its formation, development, the current state of corpus linguistic and general tendencies of creating corpus;

To investigate the linguistic basis of authorship corpus;

To determine general and different aspects of authorship corpus;

To learn experience of authorship corpus as the examples of the works by A.Chekhov, A.S.Pushkin, F.Dostoevcki, A.S.Griboedov, Shakespeare, Firdavsi, Rumi, Sadi Sherazi, Loyic Sherali;

To elaborate the tendencies of creating authorship corpus;

To characterize peculiarities of A.Kahhar's corpus, such as design, searching system, the meaning of database windows;

To base support the ways of linguistic modeling and tagging the words morphollogically as the examples of Kahhar's works;

To illuminate the importance of semantic group, huddle and square classification in the semantic tag;

To create the linguistic basis of syntactic tagging as examples of Kahhar's works.

As the object of the research, the corpus, its kinds, the authorship corpus are selected.

The subjects of the research consist of Kahhar's works, interface of corpus of his stories, identity of corpus and problem of tagging.

The scientific novelty of the research consists of the following:

It is founded that the significance of using from corpus, chance of recreating and centralizing information's in linguistic, corpus and lingvodidactics of it.

It is determined that there are three chapters of creating corpus, the first generations in as a form of electronic library, the second one the ability of recreating text, so in the new time, the originating of corpuses in simple tag although their large size, the history of linguistic corpuses in Russian and English, the present state of corpus linguistic in Russian, English, Turkish, Tajik languages among international corpuses, mass of them general satisfaction level, so different sates are in the majority and chance of linguistic analysis;

It is given that to project and to organize corpus are learnt, pointing to recreate text of tag in corpus and to show grammatical signs of word forms of tag are based, the kinds of corpus manager like searching, selecting, filtering assignments and kinds of BONITO, XAIRA, SARA, SQR, DDS;

It is opened that the scientific, practic, educational aim and tasks of authorship corpus, the difference from national, parallel, educational, multimodal corpuses, construction of authorship corpus, its structure, composition, general and different peculiarities of authorship corpus in Russian and English languages;

It is worked that to project organizing authorship corpus, to tag it and the tendency of selecting searching system (corpus manager).

Implementation of research results:

On the basis of scientific findings and conclusions drawn in the process investigating the grammatical form and word form as complex of opportunities and the occurrence of these opportunities in the syntax:

The study of issues related to the Uzbek linguistics system and its units determining the methodology of research in linguistics, the philosophical and linguistic description of substantial prices has been applied in the implementation of the fundamental project called "The Brief Description Vocabulary of System Linguistics" (Data № 89-03-3647 as 26th October 2018 given by the Ministry of higher and Secondary Specialized Education of the Republic of Uzbekistan). Based on the scientific results, an improvement in the description of the terms of corpus linguistics was achieved;

In the dissertation, the team used the conclusion regarding equalization of the word forms, forming the linguistic models of them and description of formed word in linguistic system to implement the implement 2008-2011 the strategies cited in "The derivational laws of linguistic development". The use of the presented result helped to rival the derivative laws of language (Data № 89-03-3647 as 26th October 2018 given by the Ministry of higher and Secondary Specialized Education of the Republic of Uzbekistan). The results of the study were used to explain the derivation phenomena;

Construction of corpus and technological processes of formulation, the signing the text reforming ability of the tag in corpus, to show grammatical sign of tag word formation, search, selection, filtration functions of corpus manager, the explanation of its types as: BONITO, XAIRA, SARA, SQR, DDS, an electronic description of authorship corpus, the distinctive comparison of electronic library, the construction of authorship formulation, tagging and the results of selection types of authority manager are used in the fundamental research project about the topic of ЁА1-ФҚ-0-07289 "The Aries English, Russian, Uzbek dictionary of national practical and science of art terms" that was accomplished in 2014-2016 related to the state scientific-technological programme National Painting and design institution named after Kamoliddin Bekhzod (Data № 89-03-3647 as 26th October 2018 given by the Ministry of higher and Secondary Specialized Education of the Republic of Uzbekistan). As a result of the linguistic modeling of lemmas, the scientific-popularity of a brief explanatory glossary of terms of applied and artistic art has been achieved and the vocabulary is filled with new sources.

The Interface of A.Kahhar was projected; The linguistic tag modules of A.Kahhar's works were used to creat off-line type of fragment belongs to authorship corpus of A.Kahhar's works (Data № 89-03-3647 as 26th October 2018 given by the Ministry of higher and Secondary Specialized Education of the Republic of Uzbekistan; certification № 000986 and № 000985 of IP-Center related to Intellectual Property agency of the Republic of Uzbekistan). As a result, a fragment of off-line version of the Corps of Abdullah Kahhar was created.

The structure and volume of the dissertation. The dissertation consists of introduction, three main chapters, conclusion and the list of used literature. The total length of the dissertation is 165 pages.

ЭЪЛОН ҚИЛИНГАН ИШЛАР РЎЙХАТИ
СПИСОК ОПУБЛИКОВАННЫХ РАБОТ
LIST OF PUBLISHED WORKS

I бўлим (I часть; I part)

1. Khamrayeva Sh. Morphological markup ang linguistic model // American Journal of Research. – USA, Michigan, 2018. – № 9-10. – P.187-198. (SJIF: 5,065. № 23)
2. Ҳамроева Ш. Корпус ва корпус лингвистикаси, унинг назарий асослари // Бухоро давлат университети илмий ахбороти. – Бухоро, 2017. – № 4. – Б. 57-62. (10.00.00. №1)
3. Ҳамроева Ш. Корпуснинг лингвистик ва бошқа соҳалардаги аҳамияти // ЎзМУ хабарлари. – Тошкент, 2017. – № 1/5. – Б. 471-474. (10.00.00. №15)
4. Ҳамроева Ш. Таълимда корпусдан фойдаланиш // Тил ва адабиёт таълими. – Тошкент, 2017. – № 9. – Б. 49-51. (10.00.00. № 9)
5. Ҳамроева Ш.М. Корпус тузиш тамойиллари // Бухоро давлат университети илмий ахбороти. – Бухоро, 2018. – № 3. – Б. 83-89. (10.00.00. №1)
6. Ҳамроева Ш. Корпус лингвистикасининг шаклланиши ва тараққиёти / Инновационные подходы в современной науке. Сборник статей по материалам XXV международной научно-практической конференции. – № 13 (25). – М., 2018. – 232 с. – С. 226-230.
7. Ҳамроева Ш. Рус корпус лингвистикаси тарихи / Культурология, искусствоведение и филология: современные взгляды и научные исследования. Сборник статей по материалам XII-XIII международной научно-практической конференции. – № 6-7 (11). – М., 2018. – 110 с. – С. 104-109.
8. Ҳамроева Ш. Тил корпусининг лексикографик аҳамияти / Ўзбек терминологияси: бугунги ҳолати ва истиқболи. Республика илмий-назарий анжуман материаллари. – Тошкент, 2017. – Б. 209-211.
9. Ҳамроева Ш. Корпус таълим воситаси сифатида / Тил ва таълим: муаммолар, истиқболдаги вазифалар. Республика илмий-амалий анжуман материаллари. – Тошкент, 2017. – Б. 43-45.

II бўлим (II часть; II part)

10. Ҳамроева Ш.М. Корпус лингвистикаси атамаларининг қисқача изоҳли луғати. – Тошкент: Камалак, 2018. – 96 б.
11. Khamrayeva Sh. Specific and prevalent peculiarities of the authorship corpus / IMPACT: International Journal of Research in Humanities, Arts and Literature. (IMPACT: IJRHAL) − Vol. 6, Issue 6, Jun 2018.– P.431-438. (Index Copernicus Impact Factor - 3,7784)
12. Хамроева Ш.М. ва б. Абдулла Қаҳҳорнинг муаллифлик корпуси. – Гувоҳнома № 000895. – Тошкент, 2018.
13. Хамроева Ш.М. ва б. Абдулла Қаҳҳорнинг “Бемор” ҳикояси корпуси. – Гувоҳнома № 000896. – Тошкент, 2018.

14. Ҳамроева Ш. Академик лицейларда “Луғат” мавзусини ўтишда ахборот технологияларидан фойдаланишнинг самарадорлиги / Филология масалалари. 8-жузв. Республика илмий-методик мақолалар тўплами. – Тошкент: Наврўз, 2015. – Б.69-71.
15. Менглиев Б., Ҳамроева Ш. Муаллифлик корпусининг мақсад ва вазифалари / Тил ва адабиёт таълимида замонавий ахборот ва педагогик технологиялар. Республика илмий-амалий анжумани материаллари. – Тошкент, 2018. – Б. 170-172.
16. Ҳамроева Ш. Идеографик разметка – идеографик луғат ёки тезаурус асоси / Ўзбек тилшунослиги ва туркий тиллар. Республика илмий-амалий анжуман материаллари. – Тошкент, 2018. – Б.109-113.

Автореферат Қарши давлат университетининг “ҚарДУ хабарлари” илмий-назарий, услубий журнали таҳририятида таҳрирдан ўтказилди (16.11.2018 йил)

Гувоҳнома № 14-061

19.11.2018. Босишга рухсат этилди.
Офсет босма қоғози. Қоғоз бичими 60x84 1/16.
“Times” гарнитураси. Офсет босма усули.
Ҳисоб-нашриёт т. 3.2 Шартли б.т. 3.7.
Адади 100 нусха. Буюртма № 34.

Қарши давлат университети
кичик босмахонасида чоп этилди.