АНДИЖОН ДАВЛАТ УНИВЕРСИТЕТИ
ҲУЗУРИДАГИ ИЛМИЙ ДАРАЖА БЕРУВЧИ
PhD.03/30.12.2019.Fil.60.02 РАҚАМЛИ ИЛМИЙ КЕНГАШ

МИРЗО УЛУҒБЕК НОМИДАГИ
ЎЗБЕКИСТОН МИЛЛИЙ УНИВЕРСИТЕТИ

РАХМАНОВА АЗИЗАХОН АБДУГАФУРОВНА

ЎЗБЕК ТИЛИ МИЛЛИЙ КОРПУСИНИ ЯРАТИШДА КОМПЬЮТЕР УСУЛЛАРИ

10.00.11 – Тил назарияси. Амалий ва компьютер лингвистикаси

ФИЛОЛОГИЯ ФАНЛАРИ БЎЙИЧА ФАЛСАФА ДОКТОРИ (PhD)
ДИССЕРТАЦИЯСИ АВТОРЕФЕРАТИ

Андижон – 2022

УДК: 811.512.133.93

Филология фанлари бўйича фалсафа доктори (PhD) диссертацияси автореферати мундарижаси
Оглавление автореферата диссертации доктора философии (PhD) по филологическим наукам
Contents of dissertation abstract of doctor of philosophy (PhD) in philological sciences

АНДИЖОН ДАВЛАТ УНИВЕРСИТЕТИ
ҲУЗУРИДАГИ ИЛМИЙ ДАРАЖА БЕРУВЧИ
PhD.03/30.12.2019.Fil.60.02 РАҚАМЛИ ИЛМИЙ КЕНГАШ

МИРЗО УЛУҒБЕК НОМИДАГИ
ЎЗБЕКИСТОН МИЛЛИЙ УНИВЕРСИТЕТИ

РАХМАНОВА АЗИЗАХОН АБДУГАФУРОВНА

ЎЗБЕК ТИЛИ МИЛЛИЙ КОРПУСИНИ ЯРАТИШДА КОМПЬЮТЕР УСУЛЛАРИ

10.00.11 – Тил назарияси. Амалий ва компьютер лингвистикаси

ФИЛОЛОГИЯ ФАНЛАРИ БЎЙИЧА ФАЛСАФА ДОКТОРИ (PhD)
ДИССЕРТАЦИЯСИ АВТОРЕФЕРАТИ

Андижон – 2022

Фалсафа доктори (PhD) диссертацияси мавзуси Ўзбекистон Республикаси Вазирлар Маҳкамаси ҳузуридаги Олий аттестация комиссиясида B2021.2.PhD/Fil1821 рақами билан рўйхатга олинган.

Диссертация Мирзо Улуғбек номидаги Ўзбекистон Миллий университетида бажарилган.

Диссертация автореферати уч тилда (ўзбек, рус, инглиз (резюме)) Андижон давлат университети веб-саҳифаси (www.adu.uz) ҳамда “Ziyonet” (www.ziyonet.uz) Ахборот таълим порталида жойлаштирилган.

Илмий раҳбар:	**Абдурахманова Муқаддас Турсуналиевна** филология фанлари номзоди, доцент
Расмий оппонентлар:	**Раупова Лайло Рахимовна** филология фанлари доктори, профессор
	Холиёров Ўрал Менглиевич филология фанлари бўйича фалсафа доктори(PhD)
Етакчи ташкилот:	**Самарқанд давлат университети**

Диссертация ҳимояси Андижон давлат университети ҳузуридаги PhD.03/30.12.2019.Fil.60.02 рақамли Илмий кенгашнинг 2022 йил “___” _____ соат _____ даги мажлисида бўлиб ўтади. (Манзил: 170100, Ўзбекистон Республикаси, Андижон шаҳар, Университет кўчаси 129-уй. Телефон/факс: 0 (374) 223 88 30, e-mail: agsu_info@edu.uz).

Диссертация билан Андижон давлат университетининг Ахборот-ресурс марказида танишиш мумкин (____ рақам билан рўйхатга олинган). (Ўзбекистон Республикаси, Андижон шаҳар, Университет кўчаси 129-уй. Тел.: 0 (374) 223 88 14).

Диссертация автореферати 2022 йил “___” _____ куни тарқатилди.
(2022 йил “____” _____ даги _____ рақамли реестр баённомаси)

Ш.Х. Шахабитдинова
Илмий даражалар берувчи илмий кенгаш раиси, ф.ф.д., профессор

Ф.Ф. Усманов
Илмий даражалар берувчи илмий кенгаш илмий котиби, ф.ф.ф.д.

М.Э. Умарходжаев
Илмий даражалар берувчи илмий кенгаш қошидаги илмий семинар раиси, ф.ф.д., профессор

Кириш (Фалсафа доктори (PhD) диссертацияси аннотацияси)

Тадқиқот мавзусининг долзарблиги ва зарурати. Бугунги кунда жаҳон тилшунослигида миллий тилнинг мавқеини ошириш, ижтимоий функцияларини кенгайтириш, коммуникациядаги ўрнини белгилаш, ахборот асри талабларига мувофиқ равишда такомиллаштиришга доир амалий ишлар олиб борилмоқда. Тил бирликларини тўплаш, умумлаштириш, луғавий қатламга кўра барча кўринишларини жамлаш, тарихий бирликларини акс эттириш, чегараланган лексик қатламга оид маълумотларни умумлаштириш давр эҳтиёжига айланди. Ҳозирги техник имкониятлардан фойдаланган ҳолда миллий тилларнинг умумий маълумотлар базасини яратиш, шу асосда тилнинг семантик имкониятларини, мазмун ифодалаш кўламини белгилаш глобал тараққиётнинг долзарб масалаларидан биридир.

Дунё тилшунослигида амалий тилшунослик, компьютер лингвистикаси, корпус лингвистикаси масалаларига бағишланган тадқиқотлар кўлами кенгаймоқда. Корпус лингвистикасининг тараққиёт тамойилларини, корпус яратишнинг компьютер усуллари, математик моделларини белгилаш, компьютер луғатларининг маълумотлар базаси сифатидаги аҳамиятини ёритиш, тезауруслар, конкордансларнинг лингвистик таъминотини изоҳлаш, корпус турларини таҳлил қилиш, миллий тил тараққиётидаги ўрнини кўрсатиш, ижтимоий соҳалар ривожидаги, таълим жараёнидаги самарадорлигини аниқлаш муҳим аҳамиятга эга бўлмоқда.

Ўзбекистонда мустақилликдан кейинги йилларда амалий тилшунослик соҳаларини ривожлантиришга эътибор қаратилди. Миллий тилдан амалий фойдаланиш кўникма ҳамда малакаларини шакллантириш, ўзбек тилининг замонавий ахборот-коммуникация тизимида кенг қўлланишини таъминлаш долзарб вазифага айланди. “Давлат тилининг ахборот ва коммуникация технологиялари, хусусан, Интернет жаҳон ахборот тармоғида муносиб ўрин эгаллашини таъминлаш, ўзбек тилининг компьютер дастурларини яратиш” муҳимлиги таъкидланди[1]. Ўзбек тилини интернет тилига айлантиришда, тил масалаларини компьютер ёрдамида ҳал қилишда, ўзбек тилининг функционал доирасини кенгайтиришда муҳим ўрин тутган миллий корпус яратиш технологияси, компьютер усулларининг монографик планда миллий тил негизида таҳлил қилингани тадқиқот мавзусинининг долзарблигини белгилайди. Миллий корпус яратишга доир графематик анализ ва разметкалашнинг ўзбек тилидаги тэг кўринишлари, коллокаций усули моделлари ишлаб чиқилгани мазкур тадқиқотнинг зарурийлигини асослайди.

Ўзбекистон Республикаси Биринчи Президентининг 2016 йил 13 майдаги ПФ-4997-сон “Алишер Навоий номидаги Тошкент давлат ўзбек тили ва адабиёти университетини ташкил этиш тўғрисида”ги Фармони, Ўзбекистон Республикаси Президентининг 2017 йил 7 февралдаги ПФ-4947-сон “Ўзбекистон Республикасини янада ривожлантириш бўйича Ҳаракатлар

[1] Ўзбекистон Республикаси Президентининг “Ўзбек тилининг давлат тили сифатидаги нуфузи ва мавқеини тубдан ошириш чора-тадбирлари тўғрисида”ги ПФ-5850-сон Фармони // Халқ сўзи, 2019 йил 22 октябрь. - № 218 (7448).

стратегияси тўғрисида”ги, 2019 йил 21 октябрдаги ПФ-5850-сон “Ўзбек тилининг давлат тили сифатидаги нуфузи ва мавқеини тубдан ошириш чора-тадбирлари тўғрисида”ги, 2020 йил 20 октябрдаги ПФ-6084-сон “Мамлакатимизда давлат тилини ривожлантириш ва тил сиёсатини такомиллаштириш чора-тадбирлари тўғрисида”ги Фармонлари; 2017 йил 17 февралдаги ПҚ-2789-сон “Фанлар академияси фаолияти, илмий тадқиқот ишларини ташкил этиш, бошқариш ва молиялаштиришни янада такомиллаштириш чора-тадбирлари тўғрисида”ги Қарори ҳамда бошқа меъёрий-ҳуқуқий ҳужжатларда белгиланган вазифаларни амалга оширишга тадқиқот натижалари муайян даражада хизмат қилади.

Тадқиқотнинг Ўзбекистон Республикаси фан ва технологиялар тараққиётининг устувор йўналишларига мослиги. Диссертация республика фан ва технологиялар ривожланишининг I. “Ахборотлашган жамият ва демократик давлатни ижтимоий, ҳуқуқий, иқтисодий, маданий, маънавий-маърифий ривожлантиришда инновацион ғоялар тизимини шакллантириш ва уларни амалга ошириш йўллари” йўналиши доирасида амалга оширилган.

Муаммонинг ўрганилганлик даражаси. Жаҳон тилшунослигида корпусга доир илк маълумотлар XX асрнинг 40-йилларида юзага келган[2] бўлса-да, корпус лингвистикасининг моҳияти, мақсади, назарий масалалари, корпус тузиш тамойиллари XX асрнинг 60-йилларига тўғри келади. Браун корпуси (1961-1964) корпус лингвистикасининг назарий ва амалий асослари акс этган дастлабки манба ҳисобланади[3]. Жон Синклернинг инглиз тили банкини шакллантиришдаги изланишлари корпус лингвистикаси асосларини такомиллаштиришга хизмат қилди[4]. “Компьютер лингвистикаси”га бағишланган мақолалар сериясида корпус масалалари ҳам таҳлил этилган[5]. Рус тилшунослигида А.Н.Баранов[6] В.П.Захаров[7], А.Б.Кутузов[8], Е.В.Недошивина[9], В.В.Рыков[10], В.Плунгян[11], К.Боярскийлар[12] корпус, унинг турлари, ўзига хос хусусияти, корпуснинг ижтимоий аҳамияти, корпус тузиш тамойиллари борасида тадқиқот олиб боришган. Муаллифлик корпуслари ҳақидаги махсус тадқиқотлар О.В.Кукушкина, А.А.Поликарпов,

[2]Курс “Корпусная лингвистика” (А.Б.Кутузов) Лицензия Creative commons Attribution Share-Alike 3.0 Unported (Электрон ресурс) - // lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf.

[3] Френсис Н., Кучера Г. Вычислительный анализ современного американского варианта английского языка. -М., 1967.

[4]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.

[5] Новое в зарубежной лингвистике. Вып: XXIV. Компьютерная лингвистика, 1989.

[6] Баранов А.Н. Введение в прикладную лингвистику. - М.: Эдиториал УРСС, 2001.

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

[8]Қаранг. Курс “Корпусная лингвистика” (А.Б.Кутузов) Лицензия Creative commons Attribution Share-Alike 3.0 Unported (Электрон ресурс) - // lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf.

[9]Недошивина Е.В. Программы для работы с корпусами текстов: обзор основных корпусных менеджеров. Учебно-методическое пособие. – Санкт-Петербург, − 2006. - 26 с.

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

[11]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

[12]Боярский К. Введение в компьютерную лингвистику. - Санкт-Петербург, 2013.

Е.В.Суровцевалар томонидан амалга оширилган[13].

Ўзбек тилшунослигида компьютер лингвистикаси, табиий тилни қайта ишлаш, статистик таҳлил масалаларига доир изланишларда корпус лингвистикасига ҳам тўхталиб ўтилган[14]. Компьютер лингвистикаси йўналишлари монографик тадқиқот объекти сифатида тадқиқ этилди[15]. Корпус лингвистикаси масалалари кейинги йилларда монографик планда ўрганила бошлади. Ш.Ҳамроева ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари ёритилган тадқиқотини амалга оширди. Тадқиқотда ўзбек тилшунослигида биринчи марта корпус, унинг ўзига хос хусусияти, назарий асослари ёритилган, лингвистик корпуснинг амалий ва таълимий аҳамияти очиб берилган[16]. А.Эшмўминов ўзбек тили миллий корпусининг синоним сўзлар базасини шакллантириш тамойилларини яратди. Маънодош сўзлар базасини тузишда семантик разметкадан фойдаланиш усулларини таҳлил қилди. Ўзбек тилининг лисоний синонимларини тэглаш алгоритмини ишлаб чиқди[17]. Бундан ташқари корпус лингвистикасининг маълумотлар базаси, лингвистик таъминотига доир тадқиқотлар амалга оширилмоқда.

Тадқиқотнинг диссертация бажарилган олий таълим ёки илмий-тадқиқот муассасасининг илмий тадқиқот ишлари режалари билан боғлиқлиги. Диссертация мавзуси Мирзо Улуғбек номидаги Ўзбекистон Миллий университети “Ўзбек тилининг ижтимоий, тарихий ва замонавий тараққиёти” мавзусида олиб борилаётган истиқболли илмий тадқиқот ишлари режасининг таркибий қисми ҳисобланади.

Тадқиқот мақсади миллий корпусларнинг лингвистик тадқиқот манбаси сифатидаги аҳамиятини очиб бериш, ўзбек тили миллий корпусини яратишнинг технологик тамойилларини белгилаш, компьютер усулларини тизимлаштириш, матнни разметкалаш, графематик анализ ва миллий разметка тэгларини ишлаб чиқишдан иборат.

Тадқиқотнинг вазифалари:

корпус, корпус лингвистикаси, унинг мазмун-моҳиятини, миллий корпуснинг тил тараққиёти ва ижтимоий соҳалардаги аҳамиятини ёритиш;

ўзбек тили миллий корпусини яратишнинг компьютер усулларини тавсифлаш, корпусларни шакллантиришда компьютер усулларидан

---

[13]Кукушкина О.В., Поликарпов А.А., Суровцева Е.В. 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. Под ред. В.В.Дубичинский.- Қаранг: Ҳамроева Ш.М. Корпус лингвистикаси атамаларининг қисқача изоҳли луғати. – Тошкент: «Камалак» нашриёти. – Б.6.

[14] A.Po‘latov, S.Muhamedova. Kompyuter lingvistikasi.-T.,2008; B.Yo‘ldoshev. Kompyuter lingvistikasi. –Toshkent,2009; A.Po‘latov. Kompyuter lingvistikasi. –Toshkent, 2011; A.Rahimov. Kompyuter lingvistikasi asoslari. –Toshkent,2011; F.Qurbonova. Kompyuter lug‘atlari: tezaurus. –Toshkent,2014; L.Abduhamidova. Tilshunoslikning yangi yo‘nalishi: kompyuter lingvistikasi. –Toshkent, 2015.

[15] Абдураҳмонова Н.З. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (Содда гаплар мисолида). Филол. фан. ... фалс. д.-ри (PhD)... дисс. – Тошкент, 2018; Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (расмий ва илмий услубдаги матнлар таҳрири дастури учун). Филол. фан. ... фалс. д.-ри (PhD)... дисс. – Тошкент, 2019.

[16] Ҳамроева Ш. М. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Филол. фан. ... фалс. д.-ри (PhD)... дисс. – Қарши, 2018.

[17] А.Эшмўминов. Ўзбек тили миллий корпусининг синоним сўзлар базаси. – Қарши, 2019.

фойдаланиш тамойилларини ишлаб чиқиш;

корпус яратишдаги матнни қайта ишлаш, конвертлаш, графематик анализ жараёнларини таҳлил қилиш;

корпусларни шакллантиришда разметкалаш тамойилларини белгилаш; морфологик, синтактик, семантик разметкалаш асосларини таҳлил қилиш;

ўзбек тили миллий корпусини яратишнинг компьютер усулларини тавсифлаш;

разметкалашнинг миллий асосларини ишлаб чиқиш.

Тадқиқотнинг объекти сифатида дунё тилшунослигида яратилган миллий корпуслар, уларнинг тузилиши, таркиби, халқаро корпус яратиш технологияси ва разметкалаш тизими олинган.

Тадқиқотнинг предметини миллий корпусларни яратишдаги конвертлаш, графематик анализ, разметкалаш, морфологик, синтактик, семантик разметка, тэглаш усулларининг йўналиши ва мазмуни ташкил этади.

Тадқиқотнинг усуллари. Тадқиқот мавзусини ёритишда моделлаштириш, тавсифлаш, қиёслаш, чоғиштириш, компонент таҳлил, статистик таҳлил методлари ва коллокаций усулидан фойдаланилган.

Тадқиқотнинг илмий янгилиги қуйидагилардан иборат:

ўзбек тили миллий корпусини тузишда фойдаланиладиган, ўзбек тилидаги киритилган матнни бўлакларга бўлиш, абзац, суперсинтактик бутунликлар талабига амал қилган ҳолда қисмларга ажратиш, матн мазмунини акс эттирувчи сарлавҳа қўйиш, матн бўлмаган элементларни олиб ташлаш, инициал билан ёзилган ўринларни ва чет сўзларни қайта кўриб чиқувчи графематик анализ усуллари ишлаб чиқилган;

автоматик таржима имкониятларини кенгайтириш учун хизмат қиладиган табиий тилни қайта ишлаш, лингвистик разметкалашнинг токенизация, лемматизация, стемминг, парсинг каби автоматик таҳрирлашнинг назарий-технологик асослари очиб берилган;

тезаурусларда лексеманинг антоними, пароними, омоними ёки контекстуал маъноларини фарқлаш учун зарур бўлган коллокаций моделлари ишлаб чиқилган;

автоматик таҳрир ва таржима дастурлари учун омонимларни фарқлашнинг ўзбек тили агглютинатив типологик ва туркий тилларга хос генеологик хусусиятларига асосланган ҳамда дастурда омонимлар автоматик тарзда фарқланадиган миллий тэглар тизими яратилган.

Тадқиқотнинг амалий натижалари қуйидагилардан иборат:

корпусларни шакллантиришнинг компьютер усуллари кўрсатиб берилган; интерфейсларни шакллантириш тамойиллари аниқланган; ўзбек тилидаги матнларни морфологик, семантик, синтактик разметкалаш усуллари ёритилган;

корпусларни шакллантиришда лингвистик моделлаштиришдан фойдаланиш усуллари очиб берилган; корпус лингвистикасининг компьютер усулларига доир терминологик луғат яратилган.

Тадқиқот натижаларининг ишончлилиги тўпланган материалларнинг ўзбек тилининг фонетик, лексик, грамматик табиатини акс эттириши, компьютер усулларига доир маълумотларнинг аниқ манбалар асосида келтирилгани билан изоҳланади.

Тадқиқот натижаларининг илмий ва амалий аҳамияти. Тадқиқот ўзбек тилининг миллий корпусини яратишнинг компьютер усулларини тизимлаштиришда, назарий асосларини белгилашда, миллий тэглаш тамойилларини аниқлашда ўзига хос ўрин тутади.

Ишнинг амалий аҳамияти компьютер лингвистикаси, корпус лингвистикасининг фан сифатида ўқитилиши жараёнида дастур ва режалар тузишда, ўқув адабиётларини яратишда, миллий корпус, параллел матнлар корпуси, муаллифлик корпусларини шакллантиришда илмий-назарий ҳамда амалий манба бўла олиши билан изоҳланади.

Тадқиқот натижаларининг жорий қилиниши.

Ўзбек тили миллий корпусини яратиш технологияси, усуллари ва матн разметкаларини ишлаб чиқиш бўйича олинган илмий натижалар асосида:

ўзбек тили миллий корпусини яратиш имконини берувчи матнларни саралаш ва қайта ишлаш, конвертлаш, графематик анализ усулларига доир хулосалардан 2017-2020 йилларда бажарилган “Development of the interdisciplinary master program on Computational Linguistics at Central Asian Universities” номли ERASMUS CLASS лойиҳасида кенг фойдаланилган, жумладан, лойиҳа доирасида бажарилган силлабусларни тузишда, дарслик ёзишда диссертацияда тадқиқот предмети сифатида белгиланган морфемаларнинг кетма-кетлиги ҳамда тадқиқот натижасида корпус лингвистикасига доир тадқиқотлар кўламини кенгайтириш, семантик, синтактик разметкалар асосидаги кенг маълумотлар базасини шакллантириш, компьютер усулларини ишлаб чиқишда, разметкалашда миллий тилнинг типологик ва генеологик хусусиятларига оид илмий хулосаларидан мазкур лойиҳанинг материалларини ишлаб чиқишда фойдаланилган (Erasmus + National Erasmus+Office-Uzbekistan ташкилотининг маълумотномаси). Натижада миллий корпусларни шакллантиришнинг компьютер усулларини тизимлаштиришга эришилган;

жаҳон тилшунослигида миллий тилнинг мавқеини ошириш, ижтимоий функцияларини кенгайтириш, корпус яратишнинг компьютер усуллари, математик моделларини белгилаш, компьютер луғатлари, миллий корпус, муаллиф корпусларининг маълумотлар базаси сифатидаги аҳамиятини ёритиш борасидаги илмий-амалий таклиф ва тавсиялардан Республика Маънавият ва маърифат марказининг 2020-2021 йилларда ўтказилган тарғибот фаолияти, марказнинг 2017-2020 йилларда амалга оширилган “Шарқ ренессанси даври алломалари ва мутафаккирлари асарларининг илмий-фалсафий ва бадиий-лингвистик тадқиқи” амалий тадқиқотлар давлат гранти лойиҳасида, шунингдек, Республика Маънавият ва маърифат маркази мажлисининг 2021 йил январдаги Ўзбекистон Республикасида 2021 йилда маънавий-маърифий ишлар самарадорлигини ошириш ва соҳани ривожлантиришни янги босқичга

кўтаришга доир қўшимча чора-тадбирлар дастурининг VI йўналишига оид “Халқимизнинг тарихий мероси, урф-одатлари ва миллий тарбия анъаналарини асраб-авайлаш, кенг аҳоли қатламлари, айниқса, ёшларимиз ўртасида динлараро бағрикенглик, миллатлараро тотувлик ва ўзаро меҳр-оқибат муҳитини мустаҳкамлаш”, 50-банди “Ёшларни миллий ва умуминсоний қадриятларга ҳурмат руҳида тарбиялашда “Алпомиш”, “Гўрўғли”, “Рустамхон” каби халқ оғзаки ижоди намуналарини кенг истифода этган ҳолда “Миллий қадриятлар” фестивалини ўтказиш” каби мавзулардаги мақсадли ва манзилли тарғибот тадбирларини ташкил этишда фойдаланилган (Республика маънавият ва маърифат кенгаши Республика маънавият ва маърифат марказининг 2021 йил 11 октябрдаги 02/08/1261-сон маълумотномаси). Натижада маънавий-маърифий ишлар самарадорлигини оширишга қаратилган маълумотлар базаси шакллантирилган;

ўзбек тили миллий корпусини яратиш, унинг давлат тили сифатидаги нуфузи ва мавқеини оширишга доир илмий-назарий хулосалардан “Ўзбекистон тарихи” телеканалининг “Мавзу” ва “Тақдимот” кўрсатувлари сценарийларини ёзишда фойдаланилган (Ўзбекистон Миллий телерадиокомпанияси “Ўзбекистон” телерадиоканали” ДУКнинг 2021 йил 19 апрелдаги 02-13-598-сон маълумотномаси)**.** Натижада компьютер лингвистикаси имкониятлари, корпусларнинг амалий аҳамияти ҳақидаги назарий билимлар доираси кенгайтирилган.

Тадқиқот натижаларнинг апробацияси. Мазкур тадқиқот натижалари бўйича 3 та халқаро ва 5 та республика илмий-амалий анжуманларида маърузалар қилинган.

Тадқиқот натижаларнинг эълон қилинганлиги. Диссертация мавзуси бўйича жами 17 та илмий иш чоп этилган, жумладан, 2 та луғат, Ўзбекистон Республикаси Вазирлар Маҳкамаси ҳузуридаги Олий аттестация комиссиясининг докторлик диссертациялари асосий илмий натижаларини чоп этиш тавсия этилган илмий нашрларда 7 та мақола, улардан 5 таси хорижий журналда эълон қилинган.

Диссертациянинг тузилиши ва ҳажми. Диссертация кириш, уч асосий боб, хулоса, фойдаланилган адабиётлар рўйхати ва иловалардан иборат. Диссертациянинг умумий ҳажми 158 саҳифани ташкил этади.

ДИССЕРТАЦИЯНИНГ АСОСИЙ МАЗМУНИ

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

Диссертациянинг **“Корпус лингвистикаси тараққиётининг назарий**

асослари” деб номланган I бобнинг биринчи параграфида “*Корпус лингвистикасининг мустақил соҳа сифатидаги тараққиёти”* ҳақида маълумот берилган.

Ҳар бир тилнинг мавқеи ахборот-коммуникациядаги ўрни билан белгиланади. Тилнинг ахборот алмашув функциясини таъминлашда компьютер лингвистикасининг, корпус лингвистикасининг аҳамияти катта. Корпуслар муайян тилнинг хусусиятларини ёритишда, имкониятларини акс эттиришда, тилшунослик соҳаларини, хусусан, компьютер лексикографиясини такомиллаштиришда, ижтимоий соҳа тушунчаларини оммалаштиришда амалий аҳамият касб этади. Параграфда корпус турларининг илмий тадқиқот материали сифатидаги аҳамияти кўрсатиб берилган.

Бобнинг иккинчи параграфида *“Ўзбек тили миллий корпусини яратишнинг назарий асослари”* ёритилган. Матнлар корпуси муайян тилнинг луғат бойлигини акс эттиради. Матнлар корпуси – матн ёки суперсинтактик бутунликлар ёрдамида ифодаланган йирик ҳажмли маълумотлар базасидир. Миллий корпусларнинг луғат фонди фақат синхроник характерда бўлмай, диахрон тараққиёт асосидаги лексик бирликларни ҳам қамраб олади. Бу эса тилнинг умумтараққиёт босқичларига хос фонетик, лексик, грамматик хусусиятларни таҳлил қилиш, луғат фонди ҳажмини аниқлаш, ривожланиш тамойилларини белгилаш имконини яратади.

Ўзбек тилшунослигида компьютер лингвистикаси, табиий тилни қайта ишлаш, статистик таҳлил масалаларига доир изланишларда корпус лингвистикасига ҳам тўхтаб ўтилган[18]. Ўзбек корпус лингвистикасининг шаклланишида компьютер лингвистикасига оид тадқиқотларнинг ҳам алоҳида ўрни бор[19]. Корпус лингвистикаси масалалари кейинги йилларда монографик планда тадқиқ этила бошлади[20].

Корпус лингвистикасига бағишланган тадқиқотларда миллий корпусларнинг лингвистик хусусиятлари ёритилган. Хусусан, рус адабий тилининг миллий корпусини шакллантириш натижалари таҳлил қилинган, иловалар характери ҳамда вазифалар белгиланган[21]. Рус тили миллий корпусининг нутқ кўникмаларини эгаллашга таъсири масалалари

[18] Po‘latov A., Muhamedova S. Kompyuter lingvistikasi. – Toshkent, 2009; B.Yo‘ldoshev. Kompyuter lingvistikasi. - Toshkent, 2009; Po‘latov A. Kompyuter lingvistikasi. – Toshkent, 2011; Rahimov A. Kompyuter lingvistikasi asoslari. – Toshkent,2011; F.Qurbonova. Kompyuter lug‘atlari: tezaurus. – Toshkennt, 2014; Abduhamidova L. Tilshunoslikning yangi yo‘nalishi: kompyuter lingvistikasi. – Toshkent, 2015.

[19] Абдураҳмонова Н. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти. (Содда гаплар мисолида). Филол. фан... (PhD) diss. – Toshkent, 2018; Хакимов М.Х. Расширяемый входной язык математического моделирования естественного языка для многоязычной ситуации машинного перевода // ЎзМУ хабарлари. – Тошкент, 2009. № 1. – Б. 75-80; Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (расмий ва илмий услубдаги матнлар таҳрири дастури учун). Филол. фан. фалс.(PhD)... дисс. – Тошкент, 2019.

[20] Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филол. ф. бўйича ф. д. (PhD) дис. автореф. – Қарши, 2018; Эшмуминов А.Ўзбек тили миллий корпусининг синоним сўзлар базаси. Филол. ф. бўйича ф. д. (PhD) дис. автореф. – Қарши, 2019.

[21] Венцов А.В., Грудева Е.В., Касевич В.Б., Ягунова Е.В. Национальный корпус русского литературного языка: некоторые результаты, приложения и задачи / Научно-техническая информация. - Сер. 2. - № 6. - С. 35-36; Венцов А.В., Грудева Е.В. О корпусе русского литературного языка / Russian Linguistics. – № 2. - С. 195 - 209.

ўрганилган[22]. Рус тили миллий корпусининг жаҳон тажрибаси контекстидаги хусусиятлари таҳлил этилган[23].

Ўзбек тилшунослигида яратилган конкорданс луғатлар миллий корпусни шакллантиришдаги муҳим босқичдир[24]. Айниқса, мумтоз манбалар конкордансларининг яратилаётгани компьютер лексикографиясини такомиллаштириш билан бир қаторда аждодларимиз маънавияти, маърифий-ахлоқий қарашларининг бугунги авлодга етказилишида, ўзбек тили тарихий асосларини, тарихий сўзлар, архаизмлар фондини шакллантиришда муҳим аҳамият касб этади[25]. Бобда ўзбек миллий корпусини яратишнинг амалий аҳамияти, ўзбек тили функционал доирасини кенгайтиришдаги ўрни асослаб берилган.

Диссертациянинг II боби **“Ўзбек тили миллий корпусини шакллантиришда графематик, морфологик усуллар”** таҳлилига бағишланган. Биринчи параграфда “*Миллий корпус матнини танлаш ва қайта ишлаш”* масалалари ёритилган. Корпусни шакллантиришнинг ўзига хос босқичлари, тамойиллари мавжуд. Ўзбек тили миллий корпуси маълумотлар базасини яратишнинг компьютер усулларини ишлаб чиқишни талаб қилади.

Корпус лингвистикасига доир тадқиқотларда корпус яратиш технологияси, воситалари, компьютер усуллари борасида сўз юритилган.

А.Баранов амалий тилшуносликка бағишланган тадқиқотида корпус лингвистикаси масалаларига кенг тўхталган[26].

В.Захаров корпус лингвистикасининг тараққиёт масалалари ёритилган ҳамкорликдаги тадқиқотида корпус яратиш усуллари ҳақида сўз юритган, корпусларни шакллантиришнинг технологик жараёни сифатида 9 та босқични қайд этган[27].

Корпусларни шакллантириш методлари дунё корпус лингвистикаси тажрибасини ўрганиш жараёнида ҳам такомиллашиб борди. Мавжуд корпусларнинг тузилиши, хусусиятлари ўзбек тили миллий корпусини яратишда асос бўлади.

Тадқиқотчи Н.Атабоев 450+ миллион сўзли COCA (Corpus of contemporary American English) “Замонавий америка инглиз тили” корпуси тузилиши, хусусиятлари, имкониятларини тадқиқ қилар экан, корпус тамойиллари асосида корпуслар яратиш ҳамда улардан амалий фойдаланишга қаратилган корпус лингвистикаси методологиясини олтита методдан иборат эмпирик тизимга ажратган[28].

Юқорида келтирилган маълумотларни ўрганиш асосида ва

---

[22] Венцов А.В., Касевич В.Б., Ягунова Е.В. Корпус русского языка и восприятие речи / Научно-техническая информация. - Сер. 2. - № 6. - С. 25 - 27.

[23] Шаров С.А. Представительный корпус русского языка в контексте мирового опыта / С.А.Шаров // Научно-техническая информация. - Сер. 2. - № 6. - С. 12-16.

[24] Алишер Навоийнинг “Ҳайрат ул-аброр” достони конкорданси. – Тошкент: ТДШИ, 2012.

[25] Умаров Э. Алишер Навоийнинг “Хамса” достонлари конкорданслари. – Тошкент, 2011. - Б. 4.

[26] Баранов А.Н. Введение в прикладную лингвистику. - М.: Эдиториал УРСС, 2001.

[27] Захаров В.П., Богданова С.Ю. Корпусная лингвистика. - Иркутск: ИГЛУ, 2011. - С.35-36.

[28] Атабоев Н. Инглиз тили корпусининг функционал хусусиятлари. Ф. ф. ф. д. (PhD) ... дисс. автореф. - Тошкент, 2020. - Б. 17.

кузатувларимиз натижасида ўзбек тили миллий корпусини яратишнинг қуйидаги тамойилларини ажратдик:

1. Корпус матнларини танлаш ва тизимлаштириш тамойили.
2. Матнларни график жиҳатдан бир хиллаштириш тамойили.
3. Матнларни морфологик жиҳатдан қайта ишлаш тамойили.
4. Матнларни синтактик жиҳатдан қайта ишлаш тамойили.
5. Матнларни семантик жиҳатдан қайта ишлаш тамойили.
6. Матнларни разметкалаш ва қидирув тизимига мослаш тамойили.
7. Корпус дизайнини амалга ошириш.

Ўзбек тили миллий корпусини шакллантиришда графематик, морфологик усуллар матнларни форматлаш, қайта ишлаш, конвертлаш, графематик ва морфологик анализ, морфологик разметкани ўз ичига олган босқич ҳисобланади. Ўзбек тили миллий корпусини шакллантиришда семантик ва синтактик усуллар, матнларни семантик, синтактик анализ ва синтактик разметкалаш босқичини ташкил қилади. Ривожланган тилларда компьютер усуллари тизимлаштирилган ҳолда жамланган[29].

Миллий корпус яратишнинг компьютер усуллари жамланган фондлар ўзбек тили миллий корпусини яратишда назарий асос ва амалий кўрсатма вазифасини бажаради.

Тадқиқотда корпусни шакллантиришнинг асосий компьютер усуллари сифатида матн киритиш, машина ўқийдиган формада қайта шакллантириш, анализ ва матнларни дастлабки қайта ишлаш, конвертлаш ва графематик анализ ҳамда морфологик, семантик, синтактик разметкалаш жараёнлари таҳлил қилинди.

Диссертацияда ўзбек тили миллий корпусига киритиладиган матнлар турлари, жанрлари тавсифи келтирилган. Миллий корпуснинг лингвистик тадқиқотлардаги аҳамияти таъкидланган.

Репрезентативлик корпус лингвистикасининг асосий тушунчаларидан бири бўлиб, матнларнинг турли даврларни, жанрларни, муаллифларни қамраб олиш хусусиятидир. Тадқиқотларда корпус лингвистикаси минимум икки хил типдаги матн корпусини таҳлил қилишини кўрсатди.

1. Универсал корпуслар – нутқий фаолиятнинг барча қиррасини намоён этадиган ёзма ва оғзаки матнлар корпуси.

2. Ижтимоий нутқнинг айрим лингвистик ёки маданий феноменларини ўз ичига оладиган корпуслар, масалан, мақоллар корпуси, газета нутқидаги сиёсий метафоралар[30].

Ўзбек тилининг миллий корпуси универсал характерга эга бўлиб, нутқий фаолиятнинг барча кўринишларини ўз ичига олади.

Бобнинг иккинчи параграфи "*Миллий корпусни шакллантиришда конвертлаш усули ва графематик анализ"* деб номланиб, унда ўзбек тилининг миллий корпусини шакллантиришда матнлар конвертлаш усули асосида

[29] Амиева А.М.,Филимонов В.В., Сергеев А.П., Тарасов Д.А. Инструменты корпусной лингвистики. - С-257.
[30] Богданова С.Ю. Исследование слова и предложения компьютерными методами // Слово в предложении: кол. монография / Под ред. Л.М.Ковалевой (отв. ред), С.Ю.Богдановой, Т.И.Семеновой. - Иркутск: ИГЛУ, 2010.

мувофиқлаштирилади. Бунда дунё миқёсидаги корпус яратиш тажрибасида қўлланаётган конвертлаш усулларидан фойдаланилади.

Корпус муайян мақсадда махсус тизим орқали шакллантирилган, матн маълум бир сақлаш бирликлари орқали конвертланган маълумотлар базасидир. Маълумотларни конвертлаш маълумотларни бир форматдан бошқасига ўтказиш, қайта ҳосил қилишдир. Масалан, “plain text” форматидаги матнни “Open Document Text” форматидаги матнга ўзгартириш, мультимедиали (график, мусиқали) файлларни қайта шакллантириш, MSI Windows installer пакетидаги EXE файлига ўзгартириш конвертлаш ҳисобланади. Маълумотларни ҳар бир компьютер дастури ўзига хос тарзда қайта ишлайди. Конвертлаш жараёнида маълумотлар тўлиқ сақланиб қолиши ёки айрим маълумотлар йўқотилиши мумкин. Баъзи ҳолларда ортиқча маълумотларнинг қўшилиб қолиш ҳолатлари ҳам кузатилади. Конвертлаш алоҳида дастурлар орқали амалга оширилади.

Миллий корпусни шакллантириш қуйидаги компонентлардан иборат:

1. Графематик анализ. Сўзларни, рақамлар комплексини, формулаларни ажратиш.

2. Морфологик анализ. Киритилган матндаги сўзларнинг морфологик талқини.

3. Синтактик анализ. Гапларни тобелик дарахти асосида тузиш.

4. Семантик анализ. Матннинг семантик графасини тузиш.

Корпусларни шакллантиришнинг компьютер усулларидан бири графематик анализдир.

Графематик анализ (ГрафАН) - табиий матнни занжир кўринишида бошланғич анализ қилиш дастури. Графематик анализ жараёни тадқиқотларда қайд этилган[31].

Графематик анализ корпусга мўлжалланган матнларни қайта ишлаш, уларни қидирув тизимига мослаштириш учун муҳим бўлган жараён ҳисобланади.

Графематик анализнинг таркибий қисми бўлган халқаро графематик дескрипторларнинг қуйидаги ўзбекча муқобилларини таклиф қиламиз:

Халқаро номланиши	**Ўзбекча номланиши**	**Графематик дескрипторлар**
RLE	LE	ўзбекча лексема
ILE	ALE	арабча лексема
ILE	FLE	форсча-тожикча лексема
ILE	RBLE	рус-байналмилал лексема
DEL	AYR	айирувчи
PUN	TB	тиниш белгиси
DC	RK	рақамлар комплекси
DSC	RHK	рақамли-ҳарфлар коплекси
CRAUNK		Юқоридаги белгиларга эга бўлмаган мураккаб боғланишлар

[31]АОТ (автоматическая обработка текста): Технологии: Графематика.

Графематик анализ матнни шунчаки белгилар кесимида таҳлил қилиш усули эмас, балки тадқиқотчидан муайян лексикологик, этимологик билимни ҳам талаб қилади. Масалан, ўз ва ўзлашган қатлам лексемаларини ажратиш учун сўзлар этимологиясини билиш талаб этилади.

Матнни саралаш уни белгилаш жараёнини ифода этади. Корпус лингвистикасида разметка тушунчаси матнни махсус кўрсаткичлар асосида белгилашни ифода этади. “Разметка корпуснинг асосий тавсифидир; у корпусни интернет тизимидаги оддий матнлар коллекциясидан ёки кутубхонасидан фарқлаб туради[32].

Разметкалар матн хусусиятларини ёритишда, керакли маълумотни тез топишда амалий аҳамият касб этади.

Разметкалар матнга берилган қўшимча маълумот, илова сифатида ҳам қайд этилган.

Разметка матнга ва унинг компонентларига махсус тэгларни ёзишда ифодасини топади. В.Захаров лингвистик, матннинг лексик, грамматик характеристикасини ва бошқа хусусиятларини ёритувчи ҳамда экстралингвистик (автор ва матн ҳақида маълумот, нашр вақти ва жойи, жанри, мавзуси) разметкаларни фарқлайди[33]. Лингвистик бўлмаган разметкаларга матнни форматловчи разметкалар (сарлавҳа, абзац), муаллиф ва матн ҳақидаги маълумотларга оид разметкалар (автор исми, ёши, жинси, умр саналари; матн сарлавҳаси, тили, нашр этилган жойи ва йили ҳақидаги маълумотлар) киради[34].

Масалан, “Қисаси Рабғузий” асарининг экстралингвистик разметкаси қуйидагича белгиланади:

<head></head> Носириддин Бурхониддин ўғли Рабғузий. 47 ёш. XIV аср. “Қиссаси Рабғузий”. Насрий асар. Қисса. Ўзбек тилида (кирилл алифбосида). Тошкент, 1991.

Ҳужжатнинг структур разметкаси (абзацлар, гаплар, сўзларни ажратиш) ва хусусий лингвистик разметка, одатда, автоматик тарзда амалга оширилади.

Автоматик разметка натижаларини таҳрир қилиш: хатоларни тузатиш ва ҳар хилликни бартараф этиш билан (қўлда ёки ярим автомат шаклда) амалга оширилади[35].

Разметкалар ахборот қидирув тизими учун ишлаб чиқилган махсус белгилар тизими бўлиб, маълумотларнинг тез ва осон топилишини таъминлайди. Лингвистик разметка типлари орасида морфологик, синтактик, семантик, анафорик, просодик, дискурсив каби разметкалар ажратилади[36]. Бундай разметкалар корпусларда матн юзасидан қидирув олиб бориш, у ёки бу матннинг ўзига хос белгиларини ажратишда муҳимдир.

Бобнинг учинчи параграфи *“Морфологик анализ ва морфологик*

[32] http://rusorpora.ru; http://rusorpora.ru

[33] Захаров В.П., Богданова С.Ю. Корпусная лингвистика. – Иркутск:ИГЛУ, 2011. - С. 45.

[34] Амиева А.М., Филимонов В.В., Сергеев А.П., Тарасов Д.А. Инструменты корпусной лингвистики. - С. 253.

[35] Захаров В.П., Богданова С.Ю. Корпусная лингвистика. - Иркутск:ИГЛУ, 2011. - С. 35-36.

[36] Амиева А.М., Филимонов В.В., Сергеев А.П., Тарасов Д.А. Инструменты корпусной лингвистики. - С. 253.

разметка” тавсифига бағишланган. Корпуслар тизимидаги махсус морфологик белгилар таҳлил жараёни учун хизмат қилади. “...морфологик таҳлил модули – сўзшаклдан унинг лемма (лексеманинг луғатдаги шакли)га қадар таҳлил қилинишидир[37]”. Жаҳон тилшунослигида корпуслар махсус халқаро тэглар оқали разметкаланган. Бу эса маълумотларнинг қидирув тизимига конвертланишида, матнларни тез ва осон топишда муҳимдир.

Корпус маълумотларнинг компьютер базаси экан, уни яратишда махсус процедура ва дастурлардан фойдаланиш талаб этилади. Табиий тилни қайта ишлашнинг асосий процедураси сифатида: токенизация, лемматизация, стемминг, парсинглар ажратилади.

Токенлар (сўз шакллари) чапда <> белгиси билан ажратилади. Масалан,

[<Олий> <таълим> <муассасалари> <орасида> <соғлом> <рақобат> <муҳитини> <шакллантириш> <рейтингни> <аниқлашнинг> <мақсади> <ҳисобланади>].

Ҳар бир токен таркибида леммалар мавжуд ва бу морфологик таҳрирнинг кейинги жараёнларида аниқлаштирилади:

“муассасалар” токени “муассаса” леммасига, “рағбатлантириш” токени “рағбатлан” леммасига мансубдир: ‘<муассасалар>’

‘муассаса’ **<ot> <t.ot> <Joti> <b.>** (“муассаса” леммасининг от туркумига хос категориал белгилари: от>турдош от>жой оти> бирликда).

‘<рағбатлантириш>’

‘рағбатлан’**<f> <ff> <shø> <zø>** (“рағбатлантириш” леммасининг феъл туркумига хос категориал белгилари: феъл>фаолият-жараён феъли>шахссиз>замон кўрсаткичисиз шакл).

Токенизацияда `*` белгиси сўзнинг бош ҳарф билан ёзилишини билдиради.

Юқоридаги матн токенизациясида токенларга ажратиш дастурининг чекланган жиҳатлари ҳам кўзга ташланади: - (чизиқча)нинг алоҳида олиниши ёки жуфт сўзларни иккита токен ҳолида ажратиш семантик жиҳатдан муаммоларни юзага келтириши мумкин: **<илмий> <-> <педагогик>, <профессор> <-> <ўқитувчиларнинг>.**

Морфологик разметкалаш морфологик анализга асосланади. Сўзни таркибий қисмларга ажратиш стем ва леммаларга ажратиш орқали амалга оширилади.

“Стемминг мақсади – семантика бўйича ўхшаш сўз шаклларининг мосларини аниқлаш”[38]. *Лемматизация* – морфологик анализнинг специфик масаласи сўзнинг бошқа сўзшаклларидан келиб чиқадиган дастлабки шаклланиш жараёни. Лемматизация таҳлил давомида бир сўз сифатида қараладиган бир сўзнинг турли флектив шакллари гуруҳини намоён этади.

[37] Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (расмий ва илмий услубдаги матнлар таҳрири дастури учун). Филол. фан. фалс.(PhD) ... дисс. – Тошкент, 2019.-Б. 14-15.

[38] Турсунов М., Қаршиев А. Ўзбекча матнларни морфологик ва лексик таҳлил қилиш алгоритмлари ва дастурлари. / “Амалий математика ва информацион технологияларнинг долзарб муаммолари” Халқаро анжуман тезислари тўплами. Of the international scientific conference “Actual problems of applied mathematics and information technologies”. – Тошкент, 2019 йил 14-15 ноябрь. - Б. 290-291

Асос шакл *лемма* ҳисобланади[39].

Разметкалар қўйишда тилнинг ўзига хос хусусиятлари ҳам эътиборга олинади. Масалан, ўзбек тилидаги омонимларни фарқлаш учун махсус тэглар талаб қилинади. Ўзбек тилидаги матнларда омонимларни фарқлашда кодлардан, символлардан, туркумни билдирувчи тэглардан фойдаланиш талаб этилади. Бунинг учун изоҳли луғатларда қўлланадиган рим рақамларидан тэг сифатида фойдаланиш мумкин. Фақат рим рақамларини қатъий тарзда белгилаб олиш керак бўлади: I – от, II – феъл, III – сифат, IV – равиш.

Ўзбек тилида от-феъл омонимлиги кўплиги даражасига кўра биринчи ўринда туради (тут I – дарахт, тут II – ҳаракат). От-сифат омонимлиги (чанқоқ I "ташналик" – чанқоқ III "ташна") ёки феъл-сифат омонимлиги (ёт II – ҳаракат, ёт III – "бегона") кейинги ўринда туради. От-равиш омонимлиги учрамайди. Сифат-равиш-феъл омонимлиги кам учрайди: тик II ҳаракат, тик III "қия", тик IV "тўппа-тўғри".

Бир туркумга оид сўзларни фарқлаш зарурати омонимлар учун танланган тэгларнинг чекланган томонини кўрсатади. Омонимик қатори аъзоларини фарқлаш учун яна махсус белги талаб қилинади: том I – "уйнинг, бинонинг устини бекитиб турувчи тепа қисми"; том I – "бўлим", "бўлак", "қисм", "жилд". Бир хил тэгланган омонимларни таржима дастури яна фарқламайди. Бу омонимик шаклларга қўшимча равишда махсус белгилар қўйиб чиқиш керак бўлади: <том **Iu**> - "уйнинг тепа қисми"; <том **Ij**>. Шундагина юқоридаги муаммоли ҳолат бартараф этилади. Компьютер дастури омонимик шакллардан тэг асосида белгиланган тегишли сўзни ажратган ҳолда таржима қилади

Тадқиқотимиз давомида ўзбек тилидаги омонимлик билан боғлиқ муаммони бартараф этиш мақсадида махсус тэглар тизими ишлаб чиқилди[40].

Диссертациянинг **"Ўзбек тили миллий корпусини шакллантиришда синтактик ва семантик усуллар"** деб номланган учинчи бобида маълумотлар базасини шакллантиришнинг синтактик ва семантик усуллари ҳақида сўз юритилган. Бобнинг биринчи параграфида "*Синтактик анализ ва синтактик разметка"* тавсифланган. Матнлар корпусида маълумотлар синтактик жиҳатдан ҳам таҳлил қилинади. Бунда мазмуннинг бирикмалар, қисқа ифодалар, бирикмани таъминловчи грамматик шакллар нуқтаи назаридан ёндашилади. Разметкалар тилнинг грамматик қурилиши, тартибини ифода этади: **S + O+ V** моделида **S** = эга, **O**= тўлдирувчи, **V**=кесим.

Синтактик разметкалар синтактик таҳлилларга асосланади. "Синтактик таҳлил – матндаги сўзларнинг грамматик боғланишини ифода этади"[41].

Синтактик анализда парсингларга ажратиш алоҳида ўрин тутади. ***Парсинг*** лексемаларнинг (сўз, токенларнинг) чизиқли кетма-кетлигини унинг формал грамматикаси билан чоғиштиришдир. Натижада тобелик дарахти

[39] Захаров В.П., Богданова С.Ю. Корпусная лингвистика. – Иркутск: ИГЛУ, 2011. - С. 39.
[40] Омоним тэглари ишга илова қилинади.
[41] Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (расмий ва илмий услубдаги матнлар таҳрири дастури учун). Филол. фан. фалс. (PhD) ... дисс. – Тошкент, 2019. - Б.14-15.

(синтактик дарахт) юзага келади. Катта ҳажмли корпуслар учун автоматик синтактик анализаторларнинг қурилиши компьютер лингвистикасининг муҳим соҳаларидан бири ҳисобланади[42].

Синтактик разметкалар тобелик дарахти моҳияти орқали намоён бўлади.

Синтактик разметкалар тилларнинг морфологик ва синтактик хусусиятларидан келиб чиқиб белгиланади.

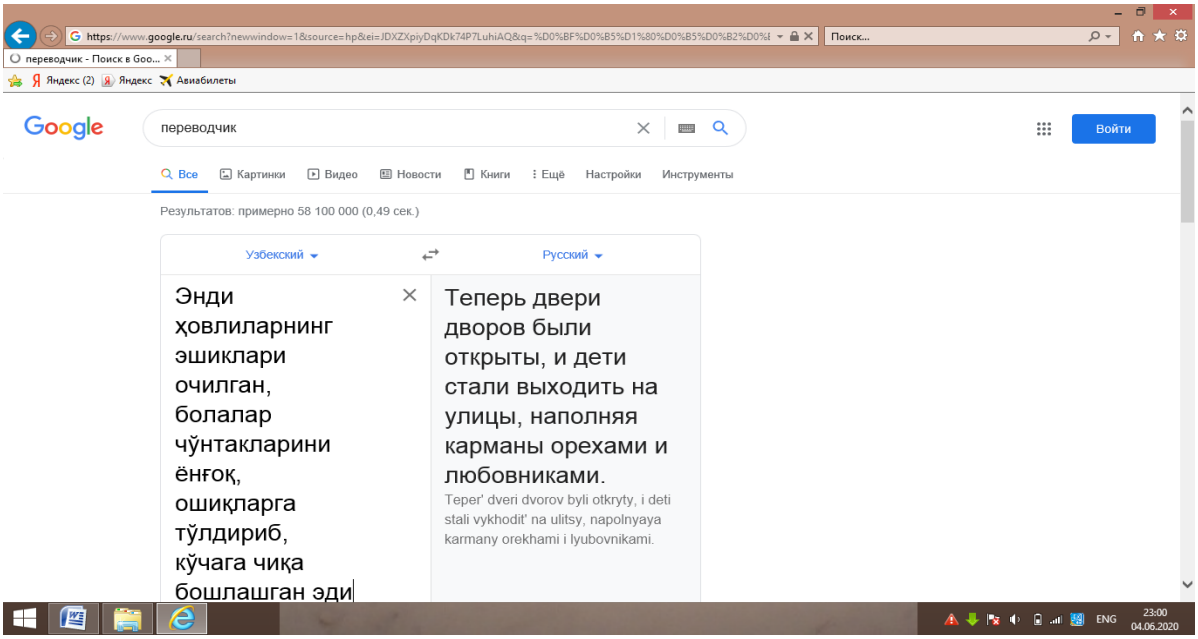
М.В.Копотев, А.Мустайокилар ҳозирги тилшуносликда корпус асосидаги ёндашувнинг аҳамиятини бир қатор мисоллар орқали кўрсатиб беришар экан, шундай ёндашувларнинг бир усули сифатида коллокацийлар (яъни лексемалар бирикмаси) тадқиқи ҳозирда корпус тадқиқотларининг энг оммабоп мавзуларидан бири бўлиб қолаётганлигини таъкидлашган[43].

Миллий корпусни шакллантиришда ва ундан фойдаланишда омонимларнинг ўзига хос мураккабликларга эгалигини *ошиқ* омоними мисолида кузатамиз:

ошиқ I – “ишқ соҳиби”; ошиқ I – “тана қисми”, ошиқ I – “ўйин номи”;

ошиқ II – “шошмоқ”; ошиқ III – “ортиқ”.

Ошиқ омонимлари иштирок этган матнларни компьютер дастури (Google translator)да таржима қилиб кўрамиз:



Ушбу мисолдаги ўйин ашёсини билдирувчи ***ошиқ*** сўзи “маҳбуба”, “севгили” маъносидаги омонимидан фарқланмаган.

Диссертацияда ошиқ омоними иштирокидаги 9та матн таржимаси таҳлилга тортилган. Таржима мазмуни, моҳияти изоҳланган. Келтирилган тўққизта мисолдан 3 тасида “ўйин ашёси” маъносидаги ***ошиқ*** ўрнида “ишқ соҳиби” маъносидаги омоними таржима қилинган (2, 3, 4). Биттаси тўғри таржима қилинган (5). Биттасида шахснинг нарса-буюмга бўлган муҳаббати

[42] Захаров В.П., Богданова С.Ю. Корпусная лингвистика. - Иркутск: ИГЛУ, 2011. - С. 41.

[43] Копотев М.В., Мустайоки А. Современная корпусная русистика // Slavica Helsingiyensia 34. Инструментарий русистики: корпусные подходы. – Хельсинки, 2008. - С. 10-13.

икки жинсдаги шахс муҳаббати маъносида таржима қилинган (1). Иккитасида *ошиқ* сўзи таржима қилинмаган, тушириб қолдирилган (7, 8). Биттасида семантик жиҳатдан ёндош сўз билан алмаштирилган (9). Бу мисолда ҳам маъно тўғри талқин қилинмаган.

Ошиқ омоним лексемаларининг таржимада, аксарият ўринларда, "ишқ соҳиби" маъносидаги сўз сифатида талқин қилиниши лексик бирликнинг қўлланиш частотаси билан белгиланади.

Ҳозирги инглиз тилида туркум омонимиясини морфологик разметкада бартараф этиш учун қуршовдан (олдин ёки кейин келган сўздан: правые и левые коллокаты) фойдаланилади. Бирикма сўзнинг қайси туркумга мансублигини аниқ кўрсатиб туради[44].

Коллокаций (правые и левые коллокаты – биз бу усулни "синтактик қуршов" деб номладик) усулидан ўзбек тилидаги омоним сўзларни ажратишда фойдаланиш мумкин. Бу усулни ***ошиқ*** омонимик қатори мисолида таҳлил қиламиз.

	Ошиқ омоними иштирокидаги контекст	**Ошиқ сўзидан олдинги ёки кейинги қисм**	**Ошиқ лексемаси туркуми ва маъноси**
1	*Ойқиз бу қизга узоқ тикилгач, уни ашула ва дутор* ***ошиғи*** *бўлса керак, деб ўйлади.*	*дутор* ø → ***ошиғи****+бўлса;* ***ошиғи****+бўлса*+г.в.→ 0	шахс оти (ш.о), "меҳр соҳиби"
2	*Энди ҳовлиларнинг эшиклари очилган, болалар чўнтакларини ёнғоқ,* ***ошиқ****ларга тўлдириб, кўчага чиқа бошлашган эди.*	0 →***ошиқларга***; ***ошиқ***+г.ш.+г.ш. → *тўлдириб*	турдош от (т.о), "ўйин ашёси"

Бу ўринда **0** -қисмнинг мавжуд эмаслиги, **ø**-грамматик шаклнинг белгисиз эканлигини, **г.ш** - грамматик шаклни, **г.в**. -грамматик воситани билдиради.

1-мисолда лексемадан олдин келган қисмда қаратқич келишигининг белгисиз қўлланган унинг от туркумига мансублигини кўрсатади. Чунки қаратқич келишиги ҳар доим **от+от** типидаги сўзларни боғлайди. 3-мисолда тушум келишигининг белгисиз қўлланган ҳам шу сўзнинг турдош отга мансублигини билдиради. Чунки тушум келишигининг белгисиз қўлланиши фақат турдош отларга хосдир. 5-, 8-мисоллардаги чиқиш келишиги кўрсаткичи лексеманинг равиш туркумига мансублигини кўрсатишга хизмат қилади.

Синтактик қуршов ошиқ омоним лексемасига доир битта янгиликни билишимизга асос бўлди. Лексема келтирилган 11та мисолдан 7 тасида (5, 6, 7, 8, 9) равиш туркумига хос хусусиятни намоён этмоқда. Ўзбек тилининг изоҳли луғатида равиш туркумига хос ***ошиқ*** лексемаси қайд этилмаган.

[44] Богданова С.Ю. Исследование слова и предложения компьютерными методами // Слово в предложении: кол. монография / Под ред. Л.М.Ковалевой (отв. ред), С.Ю.Богдановой, Т.И.Семеновой. - Иркутск: ИГЛУ, 2010.

“Керагидан ортиқ, кўп; ортиқча”, “кераксиз, беҳуда” маъноларидаги ошиқча сўзи қайд этилган. Қайси туркумга оидлиги айтилмаган[45].

Функционал омонимияни ечиш учун рус тилида статистик усул, қоида асосидаги усул ва контекстуал усул қўлланади. Бундай дастурларни яратиш жуда кўп вақт талаб қилади[46].

Синтактик қуршов усулини полисемантик сўзларни таҳлил қилишда ҳам қўллаш мумкин. Тадқиқотда *бош* полисемантик лексемасининг коллакацийлар асосидаги таҳлили келтирилган.

Синтактик қуршов методи кўп маъноли сўзнинг туркум хусусиятларини намоён этмайди. Бу ҳолат кўп маъноли сўзнинг бир сўз эканлиги ва барча ўринларда ундан олдин ёки кейин келган қисм бир хил грамматик маънонинг реаллашувига хизмат қилиши маълум бўлди. Синтактик қуршов кўп маъноли лексемадан олдинги ва кейинги қисмларнинг иштирокини, бу қисмларнинг марказий лексемага боғланиш усулларини, грамматик маъно ифодаловчи воситалар иштирокини кўрсатишга хизмат қилади.

Иккинчи параграфда “*Семантик анализ ва семантик разметка”* хусусиятлари изоҳланган. Семантик разметкалар матн таркибидаги семантик жиҳатдан алоқадор сўзлар, ассоциатив муносабатлар асосида шакллантирилади.

Корпус лингвистикасини ривожлантиришда, информацион қидирув тизимида, семантик таҳлилларда ўзига хос ўрин тутадиган, электрон база сифатида қийматга эга бўлган компьютер луғатларидан бири тезауруслардир. Тезауруслар семантик разметканинг мукаммал кўринишини ифода этади.

Тезаурус (юнонча θησαυρός “хазина” деган маънони англатади) – матннинг асосий мазмун-мундарижасини акс эттирадиган калит сўзлар, терминлар, таянч тушунчалар базаси асосидаги компьютер луғатидир. Терминлар тезаурусга қатъий семантик принциплар асосида киритилади, бунда бирликларнинг гипо-геперонимик (тур-жинс), ҳоло-мероник (бутун-бўлак), иерархик (поғонали) муносабатлари ҳамда ассоциатив мазмуний боғланишлари эътиборга олиниши лозим. Шу ўринда таъкидлаш зарурки, кейинги пайтларда қидирув тизимлари базасидаги тезауруслар гиперҳаволалар билан ҳам таъминланмоқда, бу эса (user)га бирмунча қулайликлар яратади. Яъни қидирилаётган терминларга алоқадор бўлган бошқа яқин тушунчалар билан боғлиқ маълумотларни ҳам осон топиш имконини беради[47].

Тезаурусларнинг маълумотлар базаси сифатидаги қийматлари ёритилган тадқиқотлар ҳам корпуслар ҳақида муайян тасаввур беришга хизмат қилади[48]. Тезаурусларнинг тузилиши, ишлаш тамойиллари, компьютер базаси сифатидаги имкониятлари, WordNet тезаурус базаси ҳақидаги маълумотлар ҳам илмий-амалий аҳамиятга эга[49].

---

[45] Ўзбек тилининг изоҳли луғати. III. –Т.: Ўзбекистон миллий энциклопедияси, 2007. - Б. 172.

[46]Богданова С.Ю. Исследование слова и предложения компьютерными методами // Слово в предложении: кол. монография / Под ред. Л.М.Ковалевой (отв. ред), С.Ю.Богдановой, Т.И.Семеновой. – Иркутск: ИГЛУ, 2010.

[47] Rahimov A. Kompyuter lingvistikasi asoslari. – Toshkent: Akademnashr, 2011. - B. 117.

[48] Rahimov A. Kompyuter lingvistikasi asoslari. – Toshkent: Akademnashr, 2011. - B. 117.

[49]Thye Global WordNet Assoiation htpp:// global wordnet.org/ wordnets-in-thye –world.

Тезаурусларда лексеманинг синонимлари, матндаги маънолари, гипо-гиперонимик қаторлари, антонимлари, паронимлари, контекстуал маънолари акс этади. Масалан,

Туз нарса-буюм оти тезауруси:

1. Сўзнинг ёки тил бирлигининг морфология ва синтаксисга алоқадорлиги: *от (турдош от)*

2. Талаффуз қилиниши: *туз, тус.*

3. Семантик таркиби: а) сўзнинг асл маъноси: Молекулалари таркибига кислота қолдиғи билан боғланган металл атомлари кирган мураккаб моддалар[50].

Шу бирикманинг овқат учун ишлатиладиган бир тури (натрий хлорид). *Ош тузи. Қоратикан туз. Майда туз. Йирик туз. Ошга туз солмоқ.*

Кўчма маъноси: Умуман, овқат, егулик-ичгулик нарса.

Кўчма маъноси: Кимсанинг бошқа одамга берган, едирган-ичирган (еган-ичган) овқати. *Туз бермоқ. Тузини емоқ (ичмоқ*);

б) синоними: *намак;*

кўчма маънода синоними: *таъм, маза.*

в) омоними: **Туз I** 1 Текис ер, кенг текислик; **туз II** феълнинг иккинчи шахс, буйруқ майли**:** жадвал тузмоқ, уюшма тузмоқ; **туз III** \р\ *карт.* 1 Ўйин турига қараб энг катта ёки бир очко қийматга эга бўладиган карта;

г) парониими: *тус* – кўриниш, ранг.

д) гипероними: *кимёвий элемент, ишқорлар гуруҳи;*

е) гипоними: *ош тузи, ишқор.*

4. Яқин маънодаги сўзлар: *туз, тузли, тузсиз, тузлама, туз-намак, нон-туз.*

5. Этимологияси: туркий; “чанг, гард, кукун ҳолатидаги модда .

6. Фразеологизмларда қўлланиши: *ярасига туз сепмоқ, туз(им) кўр қилгур (қарғиш;* нонкўрлик қилган кишига қарата айтиладиган ибора), *туз ялаб* (ҳеч нарса емай-ичмай, оч ҳолда), тузини ичиб, тузлуғига тупламоқ *(*ёки *тупурмоқ*; нонкўрлик қилмоқ, нонтепкилик қилмоқ, яхшиликка ёмонлик қайтармоқ); *тузини оқламоқ (*ёки *ҳалолламоқ;* кимсанинг берган ош-тузи, боққани, парвариш қилгани зое кетмаганини исбот қилмоқ, яхшиликка яхшилик қайтармоқ).

Бирикмалар таркибида қўлланиши: *тузини кўрмоқ, тузини тотмоқ* (тузини кўрмоқ; мазасини кўрмоқ, таъмини тотмоқ)*, ош тузи, лах туз, қизил туз, оқ туз, туз ҳақи, туз таъми, туз мазаси.*

7. Бошқа тилларда қўлланиши: рус тилида: *соль,* инглиз тилида: *salt*

“Ўзбек тилининг WordNet – тезаурусини ишлаб чиқиш анъанавий ўзбек тили лексикографияси ва замонавий ахборот технологияларини уйғунлаштиришни тақозо қилади. Корпус технологияларининг қўлланиши ўзбек тили сўзларини ажратилган ифодасини ва уларнинг реал контекстуал қуршовдаги лексик-семантик вариантларини акс эттирадиган ресурс яратишга

[50] *Туз* от-лексемасининг маъноларини изоҳлашда “Ўзбек тилининг изоҳли луғати”дан фойдаланилди. IV жилд. - Тошкент, 2008.- Б. 179.

имкон беради”[51].

М.В.Копотев, А.Мустайокилар ўз маърузаларидан бирида матнни автоматик таҳлил қилишнинг ривожланиш босқичларини ифода этгани ҳақида ёзишган: лемматизация (яъни сўзшаклининг бошланғич шакли ҳақидаги автоматик маълумот-лемма)→сўз туркуми разметкаси (яъни сўзшаклига морфологик белгиларни яримавтоматик ёзиш) →парсинг (яъни синтактик бирликка маълум белгиларни автоматик ёзиш) →тезауруслар тузиш (“семантик разметка”) →семантик тўрларни яратиш[52].

Разметкаларни иккига ажратиш мумкин. 1. Халқаро разметкалар. 2. Миллий тилда қабул қилинган разметкалар. Халқаро миқёсда қабул қилинган махсус белгилар тизими тушунчаларни умумий тарзда кодлашга хизмат қилади.

Миллий тилда яратилган корпуслар шу миллат ижтимоий доирасида фойдаланилади. Бундай корпусларнинг мақсади миллий тил соҳибига шу тилга оид матн мазмунини, тушунчалар ва терминларни етказишни мақсад қилади. Шу боис миллий корпусларда миллий тил асосидаги разметкаларни қўллаш ижобий натижа беради.

Миллий тилдаги разметкалардан фойдаланиш хориж тажрибасида ҳам кузатилади. Хитой тилида маълумотларни осон етказиш мақсадида миллий тилдаги разметкалардан фойдаланилади. Маълумотни тез етказиш мақсадида иероглифларнинг бош ҳарфларидан олинган қисқартмалардан фойдаланилади. Кўп ҳолларда фойдаланувчилар яқин-атрофдаги бегоналарга нисбатан ҳам қўлланувчи қариндошлик терминларини қисқартирадилар:

г – 哥 哥 gēge (сўзма-сўз: ъакаъ) “йигит”, “ака”

(суҳбатдошга ҳурматни ифодалайди)[53].

Миллий тэглар базаси лингвистик тадқиқотлар учун аҳамиятли бўлган корпусларни шакллантириш учун зарурдир. Маълумотларнинг тушунарли бўлиши, моделлаштириш, қидирув жараёни қулайлигини таъминлаш мақсадига хизмат қилади.

ХУЛОСА

1. Ҳар бир тилда яратилган миллий корпуслар маълумотлар базаси, биринчи навбатда, шу тилнинг луғат бойлигини, грамматик хусусиятларини акс эттиришга, лексик-семантик имкониятларини кўрсатишга хизмат қилади. Ўзбек тилининг миллий корпуси шу тилнинг дунё коммуникацион тизимида ўз ўрнига эга бўлишини таъминлайди. Ўзбек тилининг халқаро информацион-

[51]Матлатипов Г.Р., Мадатов Х.А. Методология автоматизации постройки некоторых онтологий типа WordNet для узбекского языка через существующих WordNet. “Амалий математика ва информацион технологияларнинг долзарб муаммолари”. Халқаро анжуман тезислари тўплами. Of the international scientific conference “Actual problems of applied mathematics and information technologies”. - Тошкент, 2019 йил 14-15 ноябрь. – Б. 286-287.

[52] Копотев М.В., Мустайоки А. Современная корпусная русистика // Slavica Helsingiyensia 34. Инструментарий русистики: корпусные подходы. Хельсинки, 2008. - С. 13

[53]Кислов А.В., Колпачкова У.Н. Влияние интернета на современный китайский язык / Компьютерная лингвистика и вычислительные онтологии. Выпуск 1. 2017. - С. 72-86.

коммуникацион тизимда фаол иштирок этишига асос бўлади.

2. Ҳозирда дунё тилларида бир қатор корпуслар яратилган бўлиб, ҳажми, тилларни қамраб олиш даражаси, структураси, мақсади, контенти жиҳатидан ўзаро фарқланади. Корпуслар маълумот бериш билан бир қаторда илмий тадқиқотлар учун манба вазифасини ҳам бажаради, лингвистик тадқиқотлар доирасини кенгайтиришга хизмат қилади. Ўзбек тилининг миллий корпуси ўзбек тили тарихини, лексемалар структурасида, семантикасида содир бўлган ўзгаришларни акс эттириш имкониятини яратади. Миллий корпуслар орқали историзмлар, архаизм, неологизмларнинг қўлланиш даври ҳақида ҳам маълумот олиш мумкин.

3. Корпусларни яратиш матнларни тўплаш, жамлаш, маълумотлар базасини шакллантиришни ифода этадиган конвертлаш, графематик анализ, матнларни аннотациялаш, махсус разметка тэглари остида бирлаштириш жараёнларини ўз ичига олади. Корпусларни шакллантиришда матнларга ишлов бериш, қайта кўриб чиқиш, таҳрирлаш талаб этилади.

4. Маълумотларни конвертлаш – маълумотларни бир форматдан бошқасига ўтказиш, қайта шакллантириш жараёнини ифода этади. Корпус матнларининг шакл хусусиятлари ва саводхонлик жиҳатидан талаб даражасида бўлиши конвертлаш жараёни ва бўлакларга бўлиш, абзац, суперсинтактик бутунликлар талабига амал қилган ҳолда қисмларга ажратиш, матн мазмунини акс эттирувчи сарлавҳа қўйиш, матн бўлмаган элементларни олиб ташлаш, инициал билан ёзилган ўринларни, чет сўзларни қайта кўриб чиқишни ўз ичига олган графематик анализ (ГрафАН)нинг сифатли ташкил этилиши билан боғлиқ.

5. Матн ва унинг қисмларига қўшимча маълумот ёзиш, илова қилиш вазифасини бажарадиган разметкалаш экстралингвистик ва лингвистик характерда бўлади. Экстралингвистик разметкалар матннинг яратилиши ва шакл жиҳатларига оид библиографик маълумотлар, матннинг жанр ва услубий хусусиятларини характерловчи белгилар, муаллиф ҳақида маълумот, файл номи, кодлаш параметрлари, разметка тили версияси, иш босқичлари бажарувчилари) каби формал белгиларни ўз ичига олади, матн паспорти характерига эга бўлади.

6. Корпусларни шакллантиришда табиий тилни қайта ишлаш, лингвистик разметкалашнинг токенизация, лемматизация, стемминг, парсинг жараёнларини мақсадли ташкил қилиш талаб этилади. *Токенизация* корпусни шакллантиришдаги асосий технологиялардан бири сифатида аниқликка асосланиши лозим. Лексема ва сўзшаклларини ажратиш жараёни бўлган лемматизация, сўз асосларини топишга қаратилган стемминг жараёнлари, сўз ёки лексемаларнинг (токенларнинг) чизиқли кетма-кетлигини унинг формал грамматикаси билан мувофиқлаштирувчи парсинглар матнларга ишлов беришнинг компьютер усулларини такомиллаштиришни талаб этади.

Морфологик разметка кенг тарқалган белгилаш тизими бўлиб, нафақат сўз туркуми ёки гап бўлаги, балки унинг қисмларини ҳам инобатга олади, сўз туркумини кўрсатиш билан бирга шу туркумга хос грамматик категорияларни

ифодалай олиши жиҳатидан эътиборга моликдир.

7. Синтактик разметкалар синтактик таҳлилларга асосланади. Синтактик таҳлил матндаги сўзларнинг грамматик боғланиши, грамматик шакллар, морфологик белгилар ва хусусиятларини, парсингларни аниқлашга хизмат қилади.Синтактик разметка лексик бирликлар ва турли синтактик конструкциялар ўртасидаги синтактик муносабатни акс эттиради.

8. Сўзларни ажратишнинг компьютер усули бўлган коллакацийлар семантик валентликка асосланади. Коллакацийлар омоним сўзларни фарқлаш ва кўп маънолиликни ёритиш усули сифатида ҳам қаралиши мумкин. Бунда матннинг сўз семантикасини ёритиш хусусиятига асосланади. Контекст, сўзнинг ўзидан олдин ва кейинги бирликлар билан бирикиши омоним ва кўп маъноли сўзларнинг маъносини аниқлаштиришга хизмат қилади.

Разметкалар “сақлаш бирлиги” номи билан ҳам юритилади. “Сақлаш бирлиги” матннинг семантик жиҳатдан муайян сўз ёки тушунчалар, гаплар остига бирлаштирилишини ифода этади.

9. Тезаурусларда сўз семантикасига асосланилгани учун семантик разметкалаш имконияти юзага келади. Тезаурусларда лексеманинг синонимлари, матндаги маънолари, гипо-гиперонимик қаторлари, антонимлари, паронимлари, контекстуал маъноларининг аниқланишида семантик разметкаларга асосланилади.

Семантик разметкалар тезаурус, ассоциатив луғатлар яратишда, корпуснинг семантик гуруҳларини шакллантиришда аҳамиятлидир.

10. Ўзбек тилидаги матнларни саралаш, графематик таҳрирлаш, разметкалаш, тэглаш, тил ўқитиш жараёнлари халқаро дастурлар ва моделлар ёрдамида амалга оширилади. Ўзбек тили миллий корпусини яратишда, компьютер усулларини ишлаб чиқишда, разметкалашда миллий тилнинг типологик ва генеологик хусусиятларини эътиборга олиш муҳимдир. Разметкалашда халқаро тэглар билан бир қаторда миллий тушунчаларни акс эттиришга қаратилган тэглардан фойдаланиш лозим.

Миллий разметкалар ўзбек тилининг ахборот услубини шакллантиришда, уни компьютер тилига айлантиришда, ўзбек тилининг мавқеи ва нуфузини оширишда муҳим ўрин тутади.

НАУЧНЫЙ СОВЕТ № PhD.03/30.12.2019.Fil
НА СОИСКАНИЕ УЧЁННОЙ СТЕПЕНИ ПРИ
АНДИЖАНСКОМ ГОСУДАРСТВЕННОМ УНИВЕРСИТЕТЕ

НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ УЗБЕКИСТАНА
ИМЕНИ МИРЗО УЛУГБЕКА

РАХМАНОВА АЗИЗАХОН АБДУГАФУРОВНА

КОМПЬЮТЕРНЫЕ МЕТОДЫ СОЗДАНИЯ НАЦИОНАЛЬНОГО КОРПУСА УЗБЕКСКОГО ЯЗЫКА

10.00.11 – Теория языка. Прикладная и компьютерная лингвистика

АВТОРЕФЕРАТ ДИССЕРТАЦИИ ДОКТОРА ФИЛОСОФИИ (PhD)
ПО ФИЛОЛОГИЧЕСКИМ НАУКАМ

Андижан – 2022

Тема диссертации доктора философии (PhD) зарегистрирована под номером B2021.2.PhD/Fil1821 в Высшей аттестационной комиссии при Кабинете Министров Республики Узбекистан.

Диссертация выполнена в Национальном Университете Узбекистана имени Мирзо Улугбека.

Автореферат диссертации опубликован на трёх языках (узбекском, русском, английском (резюме)), веб-страница Андижанского государственного университета (www.adu.uz) также в информационно-образовательном портале “Ziyonet” (www.ziyonet.uz).

Научный руководитель:	**Абдурахманова Мукаддас Турсуналиевна** кандидат филологических наук, доцент
Официальные оппоненты:	**Раупова Лайло Рахимовна** доктор филологических наук, прфессор
	Холиёров Урал Менглиевич доктор философии (PhD) по филологическим наукам
Ведущая организация:	**Самаркандский государственный университет**

Защита диссертации состоится «____» _____ 2022 года в ____ часов на заседании Научного совета под номером PhD.03/30.12.2019.Fil.60.02 при Андижанском государственном университете (Адрес: 170100, Республика Узбекистан, город Андижан, улица Университетская дом 129. Телефон/факс: 0 (374) 223 88 30, e-mail: agsu_info@edu.uz).

С диссертацией можно ознакомиться в информационно – ресурсном центре Андижанского государственного университета (зарегистрирована под номером _____). (Республика Узбекистан, город Андижан., улица Университетская дом-129. Тел.: 0 (374) 223 88 14).

Автореферат диссертации разослан «___»____ 2022 года.
Реестр протокола рассылки «____» _____ _____ 2022 года.

Ш.Х. Шахабитдинова
Председатель Научного совета по присуждению учёных степеней, доктор филологических наук, профессор.

Ф.Ф. Усманов
Научный секретарь Научного совета по присуждению ученых степеней, доктор филологических наук.

М.Э. Умарходжаев
Председатель научного семинара при Научном совете по присуждению учёных степеней, доктор филологических наук, профессор

Введение (аннотация диссертации доктора философии (PhD))

Актуальность и востребованность рассматриваемой проблемы в выбранной научной отрасли. В мировой лингвистике проводится практическая работа по повышению статуса национального языка, расширению его социальных функций, установлению его роли в коммуникации, совершенствованию его в соответствии с требованиями информационной эпохи. Совокупность языковых единиц, обобщение, сбор всех проявлений в соответствии с лексическим определением, отражение исторических единиц, обобщение соответствующей информации стало насущной необходимостью нашего времени. Создание общей базы данных национальных языков, использующей современные технические возможности, на этой основе определение семантических возможностей языка, масштабов выражения контента, является одним из актуальных вопросов глобального развития.

В мировой лингвистике существует сфера исследований по вопросам прикладной лингвистики, компьютерной лингвистики, корпусной лингвистики. Важно уточнить принципы развития корпусной лингвистики, компьютерные методы создания корпусов, определение математических моделей, важность компьютерных словарей в качестве баз данных, тезаурусы, пояснения лингвистического обеспечения конкордансов, анализ видов корпусов, обозначение их места в развитии национального языка, определить его эффективность в развитии социальных сфер, в образовательном процессе.

В годы независимости в Узбекистане большое внимание уделялось развитию прикладной лингвистики. Формирование навыков и умений практического использования национального языка, широкого применения узбекского языка в современной информационно-коммуникационной системе стало актуальной задачей. В проекте Закона Республики Узбекистан «О государственном языке» в новой редакции, усовершенствованный с учетом требований сегодняшнего дня была подчеркнута важность[1] «Обеспечение занятия государственным языком достойного места в области информационных и коммуникационных технологий, в частности во всемирной информационной сети Интернет, создание компьютерных программ узбекского языка». Технология создания национального корпуса, которая играет важную роль в преобразовании узбекского языка в язык Интернета, решение языковых вопросов с помощью компьютера, функциональный охват узбекского языка, анализ компьютерных методов в монографическом плане на основе национального языка определяет актуальность темы исследования. Разработка графематического анализа и разметки узбекского языка в качестве тэга, а также моделей методов коллокации для создания национального корпуса основана на необходимости данного исследования.

В определенной степени результаты данного исследования служат реализации задач, определенных в нормативно-правовых актах Республики

Узбекистан. В особенности Указ первого Президента Республики Узбекистан от 13 мая 2016 г. № УП-4997 «О создании 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», Указ Президента Республики Узбекистан, от 21.10.2019 г. № УП-5850 «О мерах по кардинальному повышению роли и авторитета узбекского языка в качестве государственного языка», Указ Президента Республики Узбекистан, от 20.10.2020 г. № УП-6084 «О мерах по дальнейшему развитию узбекского языка и совершенствованию языковой политики в стране»; Постановление Президента Республики Узбекистан, от 17.02.2017 г. № ПП-2789 «О мерах по дальнейшему совершенствованию деятельности академии наук, организации, управления и финансирования научно-исследовательской деятельности».

Соответствие диссертации приоритетам развития науки и технологий Республики Узбекистан. Исследование проводилось в соответствии с приоритетным направлением развития науки и технологий республики Узбекистан. I. «Духовно-нравственное и культурное развитие демократического и правового общества, формирование инновационной экономики».

Степень изученности проблемы. Первые данные о корпусе в мировой лингвистике появились[2] в 40-х годах XX века. Сущность, назначение, теоретические вопросы корпусной лингвистики, принципы построения корпуса совпадают с 60-ми годами XX века. Первым источником описывающий теоретические и практические основы корпусной лингвистики является корпус Брауна (1961-1964 гг.)[3]. Развитию основ корпусной лингвистики послужили труды Жона Синклера в исследованиях формирования банка английского языка[4]. В статьях посвещённых серии «Компьютерная лингвистика» анализированы задачи корпуса[5]. В лингвистике русского языка были проведены ряд исследований о корпусе, его видам, особенностям, социальной значимости и принципам построения корпуса такими исследователями, как А.Н.Баранов[6], В.П.Захаров[7], А.Б.Кутузов[8], Е.В.Недошивина[9], В.В.Рыков[10], В.Плунгян[11], К.Боярский[12]. Специальные исследования по авторскому корпусу[13] были проведены О.В.Кукушкиной, А.А.Поликарповым, Е.В.Суровцевой.

В лингвистике узбекского языка исследования в области компьютерной лингвистики, обработки естественного языка, статистического анализа также затрагивали корпусную лингвистику[14]. Направления компьютерной лингвистики исследовались в качестве монографического исследовательского объекта[15]. Проблемы корпусной лингвистики в последующие годы изучались в монографическом плане. Ш. Хамраева провела исследование, в котором были продемонстрированы лингвистические основы создания авторского корпуса узбекского языка. В исследовании, впервые в лингвистике узбекского языка рассматривался корпус, его отличительная особенность, теоретические основы, изложены

прикладные и образовательные значения лингвистического корпуса[16]. А. Эшмуминов создал принципы формирования базы синонимов слов национального корпуса узбекского языка. Он проанализировал методы использования семантической разметки при создании базы синонимичных слов. Разработал алгоритм тэгирования лингвистических синонимов узбекского языка[17]. В базе данных корпусной лингвистики проводятся исследования лингвистического обеспечения.

Соответствие исследования планам научно-исследовательской работы высшего учебного заведения или научно-исследовательского учреждения, в котором выполнялась диссертация. Тема диссертации является частью перспективного плана научно-исследовательских работ, проводимых в Национальном университете Узбекистана имени Мирзо Улугбека по теме “Социальное, историческое и современное развитие узбекского языка”.

Цель исследования подчеркнуть важность национального корпуса в качестве источника лингвистического исследования, определить технологические принципы создания национального корпуса узбекского языка, систематизировать компьютерные методы, разметка текста, анализ тэгирования задач, разработка графематического анализа и национальной разметки тэгов.

Задачи исследования:

разъяснить такие понятия как корпус, корпусная лингвистика, дать определение содержанию и сущности корпуса, раскрыть значение национального корпуса в развитии языка, также в социальной сфере;

описать компьютерные методы создания национального корпуса узбекского языка, используя компьютерные методы разработать принципы формирования корпуса;

анализировать процессы графематического анализа, конвертирования, обработки текста при создании корпуса;

определить принципы разметки формировании корпусов, анализировать основы морфологической, синтактической, семантической разметки;

описать компьютерные методы создания национального корпуса узбекского языка;

разработать основы национальной разметки.

Объектом исследования являются национальные корпуса в мировой лингвистике, их строение, структура, технология создания национального корпуса и система разметки.

Предметом исследования служит основное содержание и направление метода тэгирования, разметка, морфологическая, синтактическая и семантическая разметка, графематический анализ, конвертирование при создании национальных корпусов.

Методы исследования. При решении поставленных задач в данной работе применялись различные лингвистические методы исследования,

включая методы моделирования, описания, сравнения, сопоставления, компонентного анализа, статистического анализа и методы коллокации.

Научная новизна исследования заключается в том, что:

при создании национального корпуса узбекского языка разработаны методы графематического анализа, такие как, разделение на части текста представленного на узбекском языке, абзац, соблюдая стандарты суперсинтактической целостности деление на части, вставка заголовка отражающего смысл текста, удаление нетекстовых элементов, повторное просматривание наличия инициалов и зарубежных слов;

раскрыты научно-технологические основы автоматического редактирования, таких как парсинг, стемминг, лемматизация, токенизация лингвистической разметки, которые служат для расширения возможностей автоматического перевода при обработке естественного языка;

разработаны модели коллокаций, необходимые для различения контекстуальных значений или омонимов, паронимов, антонимов лексем в тезаурусе;

на основе агглютинативных типологических и генеалогических особенностей узбекского языка для программ автоматического редактирования и перевода создана система национальных тегов, в которых омонимы автоматически дифференцируются.

Практическая значимость исследования заключается в том, что:

предоставлены компьютерные методы формирования корпусов;

определены принципы формирования интерфейсов;

выявлены методы морфологической, семантической, синтаксической разметки текстов в узбекском языке;

определены требования к использованию лингвистического моделирования при формировании корпусов;

создан терминологический словарь по компьютерным методам корпусной лингвистики.

Достоверность результатов исследования подтверждается тем, что собранные материалы отражают фонетический, лексический, грамматический характер узбекского языка, информацию по компьютерным методам основанных на достоверных источниках.

Научная и практическая значимость результатов исследования. Исследование занимает особое место в систематизации компьютерных методов создания национального корпуса узбекского языка, определении теоретических основ, также принципов национального тэга.

Практическая значимость работы объясняется тем, что в процессе преподавания компьютерной лингвистики, корпусной лингвистики как науки, исследование может стать научно-теоретическим и практическим ресурсом при создании программ и планов, учебной литературы, при формировании национального корпуса, корпуса параллельных текстов, авторского корпуса.

Внедрение результатов исследования.

Заключения о роли компьютерных методов в создании корпуса были широко использованы в проекте ERASMUS CLASS под названием «Development of the interdisciplinary master program on Computational Linguistics at Central Asian Universities», который была проведён в 2017-2020 годах. В частности, в рамках данного проекта использованы материалы при разработке силлабуса, написании учебника в определении последовательности морфем в качестве предмета исследования в диссертации, также в расширении сфер научных исследований корпусной лингвистики, формировании обширной базы данных на основе семантической и синтактической разметки, при разработке компьютерных методов, научных заключений разметки о типологических и генеалогических особенностях национального языка (справочник организации Erasmus + National Erasmus+Office-Uzbekistan). В результате была достигнута систематизация компьютерных методов формирования национальных корпусов.

В диссертации выдвинуты предложения и рекомендации по повышению статуса национального языка, расширению социальных функций, компьютерных методов создания корпусов, определению математических моделей, использованию компьютерных словарей национальному корпусу, а также авторскому корпусу в качестве базы данных. Существенный вклад в изучение проблемы внёс Республиканский Центр Духовности и просвещения, где 2020-2021 году в агитационно-разъяснительном направлении деятельности центра были освещены предложения и рекомендации в государственном гранте прикладных исследований под названием «Научно-философское и художественно-лингвистическое исследование трудов ученых и мыслителей периода Восточного ренессанса», который был проведён в 2017-2020 гг.; кроме того, на заседании Республиканского Центра Духовности и просвещения от января 2021 году по теме «Сохранение исторического наследия, обычаев и традиций национального воспитания нашего народа, укрепление атмосферы терпимости, согласия и взаимной привязанности среди широких слоев населения, особенно среди нашей молодежи» по VI направлению программы дополнительных мер по повышению эффективности духовно-просветительской работы в Республике Узбекистан в 2021 году и ожиданию развития сферы на новый уровень от 50-пункта «Воспитание молодёжи в духе уважения к национальным и общечеловеческим ценностям с широким использованием образцов устного народного творчества, таких как, «Алпамиш», «Гуругли», «Рустамхан» был проведён фестиваль под названием «Национальные ценности»», где при организации целенаправленных и целенаправленно-агитационных мероприятий были использованы материалы диссертации. (Справка №02/08/1261 Республиканского совета духовности и просвещения Республиканского Центра духовности и просвещения от 11 октября 2021 года); В результате была сформирована база данных, направленная на повышение эффективности духовно-просветительской работы.

Материалы исследования, в том числе научно-теоретические выводы о создании национального корпуса узбекского языка, его влияние и позиции в качестве государственного языка, были использованы для написания сценарий передач «Мавзу» и «Такдимот» телеканала «История Узбекистана» (Справка № 02-13-598 редакции государственного унитарного предприятия Национальной Телерадиокомпании Республики Узбекистан Телерадиоканал «Узбекистан» от 19.04.2021 года). В результате расширился объем теоретических знаний о возможностях компьютерной лингвистики, практической значимости корпусов.

Апробация результатов исследования. По результатам данного исследования опубликованы доклады на 3 международных и 5 республиканских научно-практических конференциях.

Объявление результатов исследования. Всего по теме диссертации опубликовано 17 научных работ, в том числе 2 словаря, 7 статей в научных изданиях, рекомендованных к публикации основными научными результатами докторских диссертаций Высшей аттестационной комиссией при Кабинете Министров Республики Узбекистан, 5 из них опубликованы в зарубежных журналах.

Структура и объем диссертации. Диссертация состоит из введения, трех основных глав, заключения, списка использованной литературы и приложений. Общий объем диссертации составляет 158 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении обосновывается актуальность темы диссертации, описываются цель и задачи, объект и предметы исследования, указывается её соответствие приоритетным направлениям развития науки и техники в республике, изложена научная новизна и практические результаты исследования, выявлена научная и практическая значимость полученных результатов, представлены материалы о результатах исследования с точки зрения внедрения в практику, об опубликованных работах и о структуре диссертации.

В первом разделе I главы диссертации, озаглавленной **«Теоретические основы развития корпусной лингвистики»**, приводится информация о *«Развитии корпусной лингвистики как самостоятельной науки»*.

Статус каждого языка определяется ролью в информационной коммуникации. Большое значение имеет компьютерная лингвистика, а также корпусная лингвистика в обеспечении информационно-обменной функции языка. Корпуса имеют практическое значение для изложения особенностей того или иного языка, отражения его возможностей, совершенствования отраслей языкознания, в частности компьютерной лексикографии, популяризации понятий социальной сферы.Раздел характеризует всю важность классификаций корпусов в качестве научно исследовательского материала.

Второй раздел посвящён «*Теоретическим основам создания национального корпуса узбекского языка*». Корпус текстов отражает словарный запас конкретного языка. Корпус текстов — крупномасштабная база данных, с помощью которой выражаются тексты и суперсинтактические единицы. Словарный фонд национального корпуса охватывает не только синхронический характер, но и лексические единицы, основанные на диахронном развитии. Это даёт возможность проанализировать фонетические, лексические, грамматические особенности, характерные для общих этапов развития языка, определить размер словарного фонда, определить принципы развития.

В лингвистике узбекского языка при изучении вопросов касающихся компьютерной лингвистики, обработки естественного языка, статистического анализа, особое внимание уделялось также и корпусной лингвистике[20]. Отмечая формирование корпусной лингвистики узбекского языка, подчёркивают особую роль исследований, в частности, исследований в области компьютерной лингвистики[21]. В последующие годы вопросы корпусной лингвистики начали изучаться в монографическом плане[22];

В исследованиях по корпусной лингвистике рассматриваются лингвистические особенности национального корпуса. В частности, были проанализированы результаты формирования национального корпуса русского литературного языка, определены задачи и характер приложений[23]. Изучены задачи восприятия речи национального корпуса русского языка[24]. Проанализированы особенности национального корпуса русского языка в контексте мирового опыта[25].

Конкорданс словари, созданные в лингвистике узбекского языка, являются важным этапом при формировании национального корпуса[26]. В частности, создание конкордансов классических источников, играет важную роль в совершенствовании компьютерной лексикографии, а также в передаче духовных, образовательных и нравственных взглядов наших предков нынешнему поколению, в формировании исторических основ узбекского языка, исторических слов, фонда архаизмов[27]. В разделе обосновывается практическое значение создания национального корпуса узбекского языка, а также роль расширения функциональных возможностей узбекского языка.

Глава II **«Морфологические, графематические методы формирования национального корпуса узбекского языка»** посвящена анализу. В первом разделе рассмотрены вопросы *«Отбора и обработки текста национального корпуса»*. Существуют определенные этапы, принципы формирования корпуса. Национальный корпус узбекского языка требует разработки компьютерных методов для создания базы данных.

Исследования, где изучается корпусная лингвистика, рассматривают технологию создания корпуса, средства и компьютерные методы.

В своих трудах посвящённых прикладной лингвистике А. Баранов большое значение придал изучению вопросов корпусной лингвистики[28].

В совместном исследовании, где рассматривались вопросы развития корпусной лингвистики, В. Захаров знакомит нас с методами создания корпуса, также отмечает 9 этапов формирования корпусов в качестве технологических процессов[29].

Методы формирования корпусов совершенствовались и в процессе изучения опыта мировой корпусной лингвистики. Структура, особенности существующих корпусов составляют основу создания национального корпуса узбекского языка.

Исследователь Н.Атабаев, изучая структуру, особенности, возможности COCA (Corpus of Contemporary American English) - «Корпус современного американского английского — электронный корпус текстов» 450+ миллионов слов, разделил методологию корпусной лингвистики, основанную на принципах корпуса, направленную на создание и практическое использование корпусов на эмпирическую систему, состоящую из шести методов[30].

На основе изучения вышеприведенных данных и в результате наших наблюдений мы выделили следующие принципы создания национального корпуса узбекского языка:

1. Принцип отбора и систематизации корпусных текстов.
2. Принцип графической сортировки текстов.
3. Принцип морфологической обработки текстов.
4. Принцип синтактической обработки текстов.
5. Принцип семантической обработки текстов.
6. Принцип разметки текстов и настройки под поисковую систему.
7. Реализация дизайна корпуса.

При формировании национального корпуса узбекского языка графематические, морфологические методы являются этапами форматирования, обработки, конвертирования текстов, графематического и морфологического анализа, а также морфологической разметки.

Семантические и синтаксические методы при формировании национального корпуса узбекского языка составляют этап семантического, синтаксического анализа, также семантической, синтаксической разметки.

В современных языках компьютерные методы сосредоточены в систематизированном виде[31].

Компьютерные методы создания национального корпуса собирательных фондов служат теоретической основой и практическим руководством при создании национального корпуса узбекского языка.

В исследовании при формировании корпуса в качестве основных компьютерных методов проанализированы процессы ввода текстов, преобразование на машинный язык, анализ и первичная обработка текстов, конвертирование и графематический анализ, а также морфологическая, семантическая и синтактическая разметка.

В исследовании представлены характеристика жанров и виды текстов, включённые в национальный корпус узбекского языка. Была подчеркнута важность Национального корпуса в лингвистических исследованиях.

Важнейшее понятие корпусной лингвистики – репрезентативность. Под репрезентативностью понимается необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т.п. Исследования показали, что корпусная лингвистика анализирует как минимум два вида текстовых корпуса.

1. Универсальный корпус — корпус письменных и устных текстов, отражающих все аспекты речевой деятельности.

2. Корпус, включающий некоторые лингвистические или культурные феномены социальной речи, например, корпус пословиц, политических метафор газетной речи[32].

Национальный корпус узбекского языка имеет универсальный характер, который включает в себя все проявления речевой деятельности.

Второй раздел главы озаглавленный *«Графематический анализ и метод конвертирования при формировании национального корпуса»*, применяет методы конвертирования текстов при формировании национального корпуса узбекского языка. Данные методы конвертирования, используются в мировой практике при создании корпусов.

Корпус формируется посредством специальной системы для определенной цели, а текст, представляет собой базу данных, которая конвертируется с помощью сохранённых единиц. Конвертирование данных — преобразование данных из одного формата в другой. Например: преобразование (конвертация) текста из формата «plain text» в текст формата «OpenDocument», конвертация мультимедийных файлов (графических, музыкальных и т. п.), преобразование (конвертирование) исполняемого файла .EXE в установочный пакет MSI Windows Installer. Каждая компьютерная программа обрабатывает данные по-своему. Конвертирование может быть с потерей информации или без потери информации. В некоторых случаях могут также добавляться дополнительные данные. Конвертирование выполняется посредством отдельных программ.

Формирование национального корпуса включает в себя некоторые компоненты:

Выделяются следующие компоненты:

1. графематический анализ — выделение слов, цифровых комплексов, формул и т.д.;

2. морфологический анализ — построение морфологической интерпретации слов входного текста;

3. синтаксический анализ — построение дерева зависимостей всего предложения;

4. семантический анализ — построение семантического графа текста.

Графематический анализ один из компьютерных методов формирования корпусов.

Графематический анализ (ГрафАн) - это программа начального анализа естественного текста, представленного в виде цепочки символов. Процесс графематического анализа встречается в исследованиях[33].

Графематический анализ — важный этап в процессе обработки текстов и их приспособления к поисковой системе.

Приведём перечень главных дескрипторов, один из которых обязательно должен присутствовать на каждой строке графематической таблицы.

Название	Узбекское название	Графематические дескрипторы
RLE	LE	узбекская лексема
ILE	ALE	арабская лексема
ILE	FLE	персидско-таджикская лексема
ILE	RBLE	русская иностранная лексема
DEL	AYR	разделитель
PUN	TB	знак препинания
DC	RK	цифровой комплекс
DSC	RHK	цифро-буквенный комплекс
CRAUNK		сложный узел, присваивается последовательностям, не обладающим вышеперечисленными признаками

Графематический анализ — это не просто метод анализа текста в качестве символов. Но и также требует от исследователя определенных лексикологических, этимологических знаний. Например, чтобы различать собственные и освоенные лексемы слоя, требуется знание этимологии слов.

Сортировка текста определяет процесс разметки. В корпусной лингвистике понятие разметки заключается в приписывании текстам и их компонентам специальных меток. «Разметка - это основное описание корпуса; она отличает корпус от обычной коллекции или же библиотек текстов в системе интернет[34].

Разметка имеет практическое значение для выделения особенностей текста, быстрого поиска необходимой информации.

Разметка фиксируется в качестве дополнительной информации текста, также в качестве приложения.

Разметка (tagging, annotation) заключается в приписывании текстам и их компонентам специальных меток тэгов (tag, tags). В. Захаров условно разделяет разметки на лингвистические, внешне лингвистические и экстралингвистические (сведения об авторе и сведения о тексте: автор, название, год и место издания, жанр, тематика)[35]. К внешне лингвистическим относятся: разметки, отражающие особенности форматирования текста (заголовки, абзацы), разметки, относящиеся к сведениям об авторе и тексте. У автора могут быть указаны его имя, возраст, пол, годы жизни и другие, а у текста указывается название, язык, год и место)[36].

Например, экстралингвистические разметки в произведении “Қиссаси Рабғузий” отмечаются следующим образом:

<head></head> Носириддин Бурхониддин ўғли Рабғузий. 47 ёш. XIV аср. “Қиссаси Рабғузий”. Насрий асар. Қисса. Ўзбек тилида (кирилл алифбосида). Тошкент, 1991.

Структурная разметка документа (выделение абзацев, предложений, слов) и собственно лингвистическая разметка обычно осуществляются автоматически.

На следующем этапе осуществляется корректировка результатов автоматической разметки: исправление ошибок и снятие неоднозначности (вручную или полуавтоматически)[37].

Разметка — это специальная система символов, разработанная для информационно-поисковой системы, которая позволяет быстро и легко находить информацию. «К лингвистическим типам разметки относят морфологические, семантические, анафорические и просодические разметки»[38]. В корпусах такие разметки важны для проведения поиска по поверхности текста, для различения своеобразных признаков того или иного текста.

Третий раздел главы повящен *«Морфологическому анализу и морфологической разметке»*. Специальные морфологические признаки в системе корпуса служат для процесса анализа. «...модуль морфологического анализа — это анализ словоформы до его леммы (форма лексемы в словаре)»[39]. В мировой лингвистике корпуса имеют разметку в виде специальных международных тэгов. Это важно при конвертировании информации в поисковой системе, быстром и легком поиске текстов.

Поскольку корпус представляет собой информационно-компьютерную базу, при его создании требуется использование специальных процедур и программ. В качестве основной процедуры обработки естественного языка выделяют: токенизацию, лемматизацию, стемминг и парсинг.

Токены (словоформы) слева отмечается знаком <>. Например,

[<Олий> <таълим> <муассасалари> <орасида> <соғлом> <рақобат> <муҳитини> <шакллантириш><рейтингни> <аниқлашнинг> <мақсади> <ҳисобланади>].

В составе каждого токена имеется лемма, которая определяется в последующем морфологическом анализе:

Токен “муассасалар” относится к лемме “муассаса”, токен “рағбатлантириш” относится к лемме “рағбатлан”:

‘<муассасалар>’

‘муассаса’ **<ot> <t.ot> <Joti> <b.>** (“муассаса” категориальные знаки свойственные лемме имени существительного: имя существительное>имя существительное нарицательное>название места> в ед.ч.).

‘<рағбатлантириш>’

‘рағбатлан’**<f> <ff> <shø> <zø>** (“рағбатлантириш” категориальные знаки свойственные лемме глагола: глагол>глагол действия>без лич> форма глагола).

В токенизации знак `*` означает, что слово пишется с заглавной буквы.

В приведенном выше тексте токенизации также видны ограниченные аспекты программы разделения токенов: - отдельно взятая (черта) или разделение двух пар слов в качестве токена, может вызвать проблемы

семантического характера: **<илмий> <-> <педагогик>, <профессор> <-> <ўқитувчиларнинг>.**

Морфологическая разметка основана на морфологическом анализе. Разделение слова на составляющие части осуществляется путем деления его на стем и леммы.

«Сте́мминг (англ. stemming — находить происхождение) — это процесс нахождения основы слова для заданного исходного слова»[40]. *Лемматизация* — специфическая задача морфологического анализа, т.е. процесс образования первоначальной формы слова, исходя из других его словоформ. Лемматизация проявляется в группе различных флективных форм слова, рассматриваемое при анализе как единое слово. Основа формы является *лемма*[41].

Разметка учитывает специфические особенности языка. Например, для различения омонимов в узбекском языке требуются специальные тэги. Для различения омонимов в текстах узбекского языка пользуются тэгами, которые означают код, символ, части речи. В этом случае в качестве тэга, можно воспользоваться римскими цифрами, которые применяются в толковых словарях. Только при этом нужно будет твёрдо отметить римские цифры:

I –имя существительное, II–глагол, III–имя прилагательное, IV–наречие.

В узбекском языке омоним имя существительное – глагол по степени множественности стоит на первом месте (тут I – «дерево», тут II – «действие»). Омоним имя существительное – имя прилагательное (чанқоқ I «жажда» – чанқоқ III «жаждущий») или омоним глагол – имя прилагательное (ёт II – «действие», ёт III – «чужой»), стоит на втором месте. Омоним имя существительное – наречие не встречается. Омоним имя прилагательное – наречие – глагол встречается очень редко: тик II «действие», тик III «склон», тик IV «ровно».

Необходимость различать слова по частям речи указывает на ограниченную сторону тэгов, выбранных для омонимов. Для различения членов омонимического ряда требуется более специальный символ: том I – «крыша, кровля»; том II – «глава», «часть», «кусок», «том». Одинаково тэгированных омонимов программа переводчик не различает. В дополнение к этим омонимическим формам необходимо будет выставить специальные знаки: <том **Iu**>–«крыша, кровля»; <том **Ij**>. Только тогда вышеуказанная проблемная ситуация будет устранена. Компьютерная программа переводит с омонимических форм с выделением соответствующего слова, которое определяется на основе тэга.

В ходе нашего исследования была разработана специальная система тэгов, чтобы устранить проблему связанную с омонимами в узбекском языке[42].

Третья глава диссертации **«Синтактические и семантические методы формирования национального корпуса узбекского языка»** посвящена синтактическим и семантическим методам формирования базы данных. Первый раздел главы характеризует *«Синтактический анализ и синтактическую разметку»*. В корпусе текстов данные также анализируются синтаксически. Имеется ввиду с точки зрения сочетаний содержания, кратких

выражений, грамматических форм, которые обеспечивают сочетание. Разметки выражают грамматический порядок строение языка: в модели **S + O+ V**, **S** = подлежащее, **O**= дополнение, **V**=сказуемое.

Синтаксическая разметка, является результатом синтаксического анализа. «Синтактический анализ — выражает грамматическую связь слов в тексте».

В синтактическом анализе особое место занимает парсинг (англ. parsing). ***Парсинг*** сравнивает линейный алгоритм лексем (слово, токены) с его формальной грамматикой. В итоге образуется дерево зависимости (синтактическое дерево). Создание автоматических синтаксических анализаторов для объемных корпусов является одним из важных направлений компьютерной лингвистики[44] .

Синтаксическая разметка представляет собой дерево зависимостей.

Синтаксические разметки определяются морфологическими и синтаксическими особенностями языков.

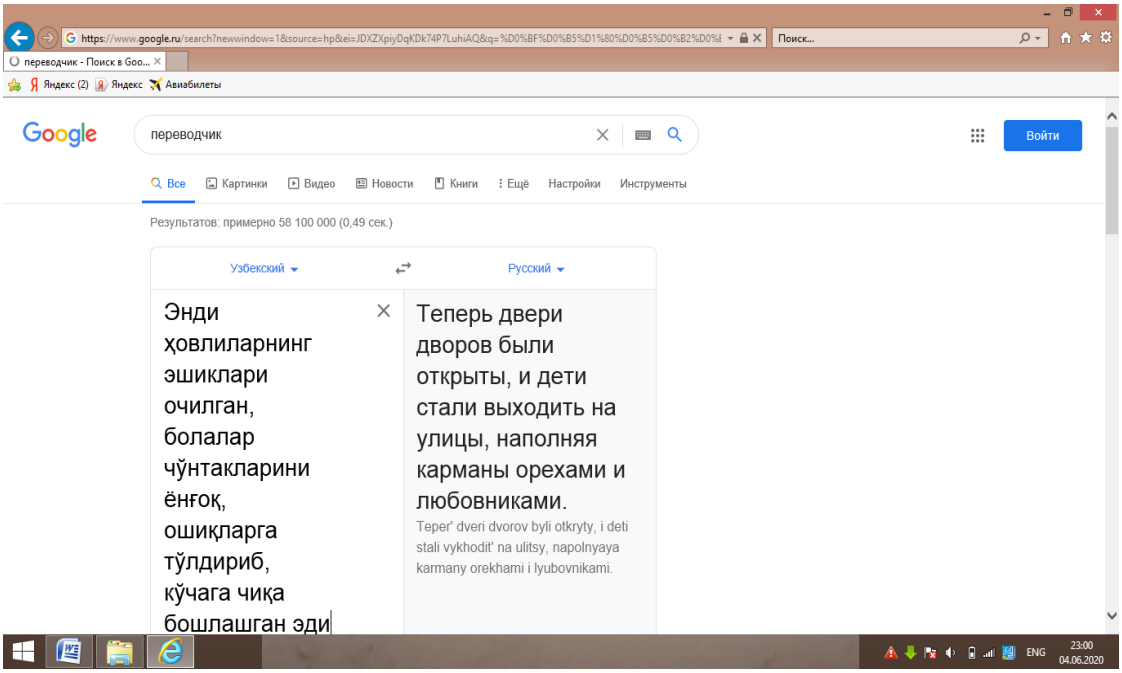
М.В.Копотев и А.Мустайоки на ряде примеров показывая важность корпусного подхода в современной лингвистике, отметили, что в настоящее время одной из самых популярных тем корпусных исследований является исследование коллокаций (то есть сочетаний лексем)[45].

При формировании национального корпуса и его использовании, можно заметить, что омонимы имеют свою специфическую сложность. Это мы можем наглядно увидеть на примере слова *ошиқ:*

ошиқ I – «влюблённый»; ошиқ I – «часть тела»; ошиқ I – «игровой предмет» «название игры»;

ошиқ II– «торопиться»; ошиқ III– «лишний».

Тексты, где встречается омоним *ошиқ* попробуем перевести с помощью компьютерной программы (Google translator):



В этом примере омоним ***ошиқ*** обозначает игровой предмет, но переводится по иному. Со смыслом «влюблённый», «любимая».

В диссертации анализу были подвергнуты переводы 9 текстов, с участием омонима ***ошиқ***. Разъясняется содержание и суть перевода. Из 9 примеров в 3 омоним ***ошиқ*** вместо «игрового предмета», переводится как «влюблённый» (2,3,4). Правильным был только один перевод (5). В одном переводе описывается смысл любви, к какому либо предмету или же любовь между двумя лицами (1). В двух переводах слово ***ошиқ*** не переведено (7, 8). В одном тексте в семантическом отношении омоним был заменён на вспомогательное слово (9). Даже в данном примере значение не было правильно истолковано.

Лексемы омонима ***ошиқ*** в переводах, во многом толкуется словом «влюблённый» и определяется частотой употребляемой лексической единицы.

Современный английский язык в морфологической разметке использует устранение омонимов частей речи (правые и левые коллокаты). Сочетание чётко определяет, к какой части речи принадлежит слово[46].

В узбекском языке при выделении омонимов используют метод коллокации (правые и левые коллокаты — этот метод мы назвали «синтактическим окружением»). Данным методом мы проанализируем омоним ***ошиқ***.

Контекст с участием омонима *ошиқ*	**Часть перед и после слова *ошиқ***	**Лексема *ошиқ* и её смысл**
Ойқиз бу қизга узоқ тикилгач, уни ашула ва дутор ***ошиғи*** *бўлса керак, деб ўйлади.*	*дутор* ø → ***ошиғи****+бўлса;* ***ошиғи****+бўлса*+г.в.→ 0	шахс оти (ш.о), “меҳр соҳиби”
Энди ҳовлиларнинг эшиклари очилган, болалар чўнтакларини ёнғоқ, ***ошиқ****ларга тўлдириб, кўчага чиқа бошлашган эди.*	0 →***ошиқларга***; ***ошиқ***+г.ш.+г.ш. → *тўлдириб*	турдош от (т.о), “ўйин ашёси”

Здесь мы видим, что **0**-часть не существует, **ø**-грамматическая форма без знака, **г.ш** – обозначает грамматическую форму, **г.в –** грамматическое средство.

В 1-примере, в части, где встречается перед лексемой, использование родительного падежа без обозначения, определяет, что омоним принадлежит к части речи существительного. Потому что родительный падеж связывает слова типа **имя существительное + имя существительное**. В 3-примере использование винительного падежа без обозначения, уведомляет, что данное слово относится к имени существительному нарицательному. Потому

что использование винительного падежа без обозначения свойственно к существительным нарицательным. В 5, 8-примере указатель исходного падежа, указывает, что лексема принадлежит к части речи наречие.

Синтактическое окружение стало основой того, что мы узнали ещё одну новость, связанную с лексемой омонима ***ошиқ***. Из приведённых 11-примеров в 7 (5, 6, 7, 8, 9) лексема показывает особенности свойственные части речи наречие. В толковом словаре узбекского языка лексема ***ошиқ,*** принадлежащая к части речи наречие не отмечается. Смысловые значения слова ***ошиқ*** отмечаются в словах «слишком много, много; лишний», «ненужный, бесполезный». Но не указывается принадлежность, к той или иной части речи[47].

В русском языке для решения функциональной омонимии используются статистические, контекстуальные методы и методы на основе правил. Создание таких программ занимает очень много времени[48].

При анализе полисемантических слов используют метод синтактического окружения.

В диссертации приведён анализ *главных* полисемантических лексем на основе коллокаций.

Метод синтактического окружения не демонстрирует особенности многозначных слов. Стало очевидным, что многозначные слова по принципу это однозначное слово, и во многом часть, которая идёт впереди и после представляет одинаковый грамматический смысл. Синтактическое окружение представляет участия частей до и после многозначных лексем, привязку данных частей к методам центральных лексем, а также участие средств обозначающих грамматический смысл.

Во втором разделе поясняются особенности *«Семантического анализа и семантической разметки».* Семантические разметки формируются на основе ассоциативных отношений, в качестве семантических слов в составе текста.

Тезаурусы являются одним из разновидностей компьютерных словарей, особое место занимают в развитии корпусной лингвистики, информационно-поисковой системе, семантических анализах, функционируют в качестве электронной базы данных. Тезаурусы выражают завершённый вид семантической разметки.

Тезаурус (от греч. θησαυρός «сокровище») — ключевые слова, представляющие собой основное содержание текста, термины, компьютерный словарь основанный на базе опорных понятий. Термины вводятся в тезаурус на основе твердых семантических принципов, учитывая семантические отношения между гипо-гиперонимическими (вид-пол), холо-мероническими (целый-часть), иерархическими (ступенчатые) единицами, также ассоциативных содержаний. Вдобавок нужно отметить, что в последнее время тезаурусы, находящиеся в базе поисковой системы обеспечиваются гиперссылками, что создают множество преимуществ

пользователю (user). То есть это облегчает поиск информации, связанной с другими близкими понятиями, которые связаны с искомыми терминами[50].

Исследования, раскрывающие суть тезаурусов в качестве базы данных, предоставляют определённое представление о корпусах. Также научно-практическое значение имеет информация о структуре тезауруса, принципы работы, возможности в качестве компьютерной базы, также база тезауруса WordNet[52].

В тезаурусах отражаются синонимы лексем, смысловые значения текстов, графы гиперонимов, антонимы, паронимы, контекстуальные значения. Например:

Туз — тезаурус вещи-предмета:

1. Отношение слова или языковой единицы к морфологии и синтаксису: *имя существительное (имя существительное нарицательное)*
2. Призношение: *туз, тус*
3. Семантический состав: а) истинное значение слова: Сложные вещества, молекулы которых содержат атомы металлов, связанные с кислотным остатком[53].

Это тип того же соединения, используемого в пищу (хлорид натрия). *Ош тузи. Қоратикан туз. Майда туз. Йирик туз. Ошга туз солмоқ* (Поваренная соль, Столовая соль. Мелкая соль. Крупная соль. Солить пищу).

Переносное значение: пища, хлеб-соль.

Переносное значение: пища, которую кто-то дает другому человеку, которую он ест и пьет (ест и пьет). *Туз бермоқ. Тузини емоқ (ичмоқ)* (Накормить хлебом-солью. Оправдать хлеб-соль);

б) синоним: *намак;*

в переносном смысле *синоним: таъм, маза* (вкус).

в) омоним: **Туз I** 1 Голое место, степь, залежь; **туз II** глагола второго лица, повелительное наклонение**:** составить таблицу, создать организацию; **туз III** \р\ *карт.* 1 Карта, которая имеет наибольшее или наименьшее значение (одно очко) в зависимости от типа игры;

г) пароним: *тус* – вид, цвет.

д) гипероним: *кимёвий элемент, ишқорлар гуруҳи* (химический элемент, щёлочная группа);

е) гипоним: *ош тузи, ишқор* (поваренная соль, щёлочь).

4. Слова, близкие по смысловому значению: *туз, тузли, тузсиз,тузлама, туз-намак, нон-туз* (соль, соляной, без соли, соленья, хлеб-соль).

5. Этимология: тюрк; «пыль, грязь, порошкообразное вещество».

6. Фразеологизм, использование в разговорах: *ярасига туз сепмоқ* (сыпать соль на чью-либо рану; бередить, растравлять чьи-либо раны); *туз(им) кўр қилгур, қарғиш, тузини ичиб, тузлуғига тупламоқ (ёки тупурмоқ; нонкўрлик қилмоқ, нонтепкилик қилмоқ, яхшиликка ёмонлик қайтармоқ)*

(букв. отведав соли плюнут в солонку) соотв. ответить на добро неблагодарностью; *туз ялаб (ҳеч нарса емай-ичмай, оч ҳолда)* (лизать соль (абсолютно ничего не есть)); *тузини оқламоқ* (ёки ҳалолламоқ; кимсанинг берган ош-тузи, боққани, парвариш қилгани зое кетмаганини исбот қилмоқ, яхшиликка яхшилик қайтармоқ) оправдать хлеб-соль (оправдать оказанное ему доверие).

Использование в словосочетании: *тузини кўрмоқ, тузини тотмоқ* (попробовать на вкус, узнать вкус), *ош тузи, лах туз, қизил туз, оқ туз, туз ҳақи, туз таъми, туз мазаси* (словосочетания где встречается слово туз).

7. Использование в других языках: в русском языке: *соль*, на английском языке: *salt.*

«Разработка WordNet — тезауруса узбекского языка требует сочетания традиционной лексикографии узбекского языка и современных информационных технологий. Применение технологий корпуса позволяет создать ресурс, отражающий выделенное выражение слов узбекского языка и их лексико-семантические варианты в реальном контекстуальном окружении»[54].

М.В.Копотев, А.Мустайоки, в своих лекциях отметили этапы развития автоматического анализа текста: лемматизация (то есть на автоматическом сведении словоформ к начальной форме), → частеречная разметка (определение части речи каждого слова) → парсинг (приписывание определенных синтаксических признаков слову или сочетанию слов) → создание тезаурусов («семантическая разметка») → создание семантических сетей[55].

Разметку можно разделить на два: 1. Международная разметка. 2. Разметка национального корпуса. Система специальных символов, принятая на международном уровне, служит для кодирования понятий в общем виде.

Корпус, созданный на национальном языке, используется в социальных рамках одной и той же нации. Цель таких корпусов состоит в том, чтобы донести до носителя национального языка текстовое содержание, понятия и термины, относящиеся к этому языку. Таким образом, использование разметки на основе национального языка в национальных корпусах даёт положительный результат.

Использование разметки на национальном языке также наблюдается в зарубежном опыте. Для того чтобы легко передавать информацию на китайском языке, используется разметка на национальном языке. Для быстрой передачи информации используются аббревиатуры от инициалов иероглифов. В большинстве случаев пользователи сокращают условия родства, которые также распространяются на близких незнакомых людей:

г – 哥 哥 гēге (сўзма-сўз: ъакаъ) «юноша», «брат»

(означает уважение к своему собеседнику)[56].

Национальная база тэгов необходима для формирования корпусов, важных для лингвистических исследований. С целью обеспечить удобство при поиске, моделировании и усвоению информации.

ЗАКЛЮЧЕНИЕ

1. База данных национальных корпусов, созданная на национальном языке, служит, прежде всего, для отражения словарного запаса, грамматических особенностей данного языка, для демонстрации его лексико-семантических возможностей. Национальный корпус узбекского языка гарантирует, что этот язык займет свое место в мировой системе коммуникации. Узбекский язык является основой для активного участия в Международной информационно-коммуникационной системе.

2. В настоящее время на мировых языках созданы ряд корпусов, которые различаются между собой по размеру, уровню языкового охвата, структуре, назначению, содержанию. Помимо предоставления информации, корпус также служит источником научных исследований, расширяя круг лингвистических исследований. Национальный корпус узбекского языка предоставляет возможность отразить историю узбекского языка, изменения в структуре лексем, семантике. Также можно получить информацию о периоде применения историзмов, архаизмов, неологизмов с помощью национального корпуса.

3. Создание корпусов включает в себя процессы объединения под специальной разметкой тэгов, аннотации текстов, графематического анализа, конвертирования, которые выражают формирование базы данных, сбор и хранение текстов. При формировании корпусов требуется обработка, повторный просмотр, редактирование текстов.

4. Конвертирование текстов — представляет собой процесс повторного формирования, преобразование данных из одного формата в другой. Наличие особенностей форм корпусных текстов и требование уровня грамотности, процесс конвертирования, абзац, соблюдая стандарты суперсинтактической целостности деление на части, заголовок отражающий смысл текста, удаление нетекстовых элементов, наличие инициалов, повторный просмотр зарубежных слов связаны с качественно организованным графематическим анализом (ГрафАН).

5. Разметка заключается в приписывании текстам и их компонентам специальных тэгов: лингвистических и внешних (экстралингвистических). Экстралингвистическая разметка включает как содержательные элементы данных (библиографические данные, признаки, характеризующие жанровые и стилевые особенности текста, сведения об авторе), так и формальные (имя файла, параметры кодирования, версия языка разметки, исполнители этапов работ).

6. При формировании корпуса требуется целенаправленная организация процессов обработки естественного языка, лингвистической разметки, токенизации, лемматизации, стемминга, парсинга. Токенизация должна основываться на конкретности как одной из основных технологий при формировании корпуса. Лемматизация, т.е. процесс образования первоначальной формы слова, исходя из других его словоформ Процесс, несколько отличный от лемматизации, называется стеммингом, он состоит в нахождении стема (основы) слова. Парсинг, процесс сопоставления линейной последовательности лексем (слов, токенов) языка с его формальной грамматикой.

Морфологическая разметка, широко используемая система меток, включает не только признак части речи, но и признаки грамматических категорий, свойственных данной части речи.

7. Синтаксическая разметка, является результатом синтаксического анализа. Синтаксический анализ служит для определения грамматической связи слов в тексте, грамматических форм, морфологических признаков и особенностей, синтаксического анализа.Синтактическая разметка описывает синтаксические связи между лексическими единицами и различные синтаксические конструкции.

8. Основывающиеся на семантической валентности коллокации, являются одним из компьютерных методов, предназначеные для выделения слов. Также коллокации рассматриваются с точки зрения метода, который различает слова омонимы и многозначные слова. В этом случае основываются на раскрытии семантики слов. Контекст, служит для определения смысла многозначных слов, для объединения омонимов и последующих единиц, идущих до и после данного слова.

Разметка также называется «единицей хранения». «Единица хранения» объединяет слова и термины, под единым предложением исходя с точки зрения семантической особенности.

9. В тезаурусе появляется возможность семантической разметки основанной на семантике слова. При определении контекстуальных значений, паронимов, антонимов, гипо-гиперонимических рядов, синонимов лексем в тезаурусе основываются на семантической разметке.

Семантическая разметка важна при формировании семантической группы в корпусе, создании ассоциативных словарей, тезауруса.

10. В узбекском языке сортировка текста, графематическое редактирование, тэгирование, процессы обучения языку осуществляются с помощью международных программ и моделей. При создании национального корпуса узбекского языка, разработке компьютерных методов при разметке важно учитывать типологические и генеалогические особенности национального языка. При разметке наряду с международными

тэгами необходимо использовать тэги, которые отражают национальные понятия.

Национальная разметка выполняет важную роль в формировании информационного стиля узбекского языка, преобразовании его в компьютерный язык, повышении статуса и авторитета узбекского языка.

SCIENTIFIC COUNCIL AWARDING SCIENTIFIC DEGREES
PhD.03/30.12.2019.FIL.60.02
IN ANDIJAN STATE UNIVERSITY

NATIONAL UNIVERSITY OF UZBEKISTAN NAMED AFTER MIRZO ULUGBEK

RAKHMANOVA AZIZAXON ABDUGAFUROVNA

"COMPUTER TECHNIQUES OF CREATING THE NATIONAL CORPORAOF UZBEK LANGUAGE"

10. 00.11 – Language theory, applied and computational linguistics

DISSERTATION ABSTRACT
FOR THE DOCTOR OF PHILOSOPHY (PhD) ON PHILOLOGICAL SCIENCES

Andijan – 2022

The theme of the doctoral (Ph.D.) thesis is registered by Supreme Attestation Commission under the Cabinet of Ministers of the Republic of Uzbekistan in number B2021.2.PhD/Fil1821

The doctoral thesis has been carried out at the National University of Uzbekistan named after MirzoUlugbek.

The abstract of the thesis is in three languages (Uzbek, Russian, English (resume)) is placed on the website (www.adu.uz) of the Scientific Council and Information and Educational portal "ZiyoNET" (www.ziyonet.uz).

Scientific adviser:	**AbdurakhmonovaMukaddasTursunalievna** Candidate of Philological Sciences, Associate professor
Official opponents:	**Raupova Laylo** Doctor of Philology, Professor
	Xoliyorov Oral Menglievich Doctor of Philological sciences (PhD)
Leading organization:	**Samarkand State University**

The defense of the dissertation will take place on "____" _____2022 at _____at the meeting of the Scientific Council awarding Scientific Degrees PhD.03/30.12.2019.Fil.60.02 at the Andijan State university (Address: 170100, Andijan region, University str., 129. Ph. number: 0(374) 223 88 14; Fax: 0(374) 223 88 30 e-mail: agsu_info@edu.uz).

The dissertation could be reviewed in the Information Resource Centre of the Andijan State University. (Registration number_____). Address: 170100, Andijan region, University str., 129. Ph. number: 0(374) 223 88 14.

Abstract of dissertation sent out on "_____" _____2022
(Protocol at the register №____of "____" _____2022)

ShakhabitdinovaSh.Kh
Chairman of the Scientific
Council awarding Scientific degrees,
Doctor of Philology, Professor

Usmanov F.F.
Scientific Secretary of the
Scientific Council awarding
Scientific degrees, PhD

Umarkhudjaev M.E.
Chairman of the Scientific
Seminar at the Scientific Council awarding
Scientific degrees,
Doctor of Philology, Professor

INTRODUCTION (abstract of Ph.D. thesis)

The research aims are to highlight the importance of national corpora as a source of linguistic research, to define the technological principles of creating the Uzbek national corpus, to systematize computer methods, text marking, analysis of tagging issues, graphical analysis and to develop national markup tags.

The object of research is the national corpus which is created in world linguistics, its structure, composition, the technology of international corpus creation, and marking system.

The subjectof the research is the direction and content of the methods of conversion, graphemic analysis, marking, morphological, syntactic, semantic marking, tagging in the creation of national corpora.

The scientific novelty of the research is:

To create graphical analysis in the Uzbek national corpus, to divide the Uzbek text into parts, paragraphs, supersyntactic integrity, to put a title that reflects the content of the text, to remove non-text elements, and to reconsider the initials and foreign words methods have been developed.

To reveal the theoretical and technological foundations of natural language processing, such as tokenization, lemmatization, stemming, analyzing linguistic marking, and to expand the possibilities of automatic translation.

To develop collocation models in thesaurus that is necessary to distinguish the antonym, paronym, homonym, or contextual meaning of a lexeme

The system of national tags for automatic editing and translation programs which is based on the agglutinative typological and genealogical features of the Uzbek language of homonyms, and the program automatically distinguishes homonyms is created.

The implementation of research results. Based on scientific results obtained on the development of technology, methods in the creation of the corps were widely used in the project ERASMUS CLASS “Development of the interdisciplinary master program on Computational Linguistics at Central Asian Universities”in 2017-2020.In particular, morphemes which are identified as the subject of the research used as a source in creating the syllabus and writing textbooks, to form an extensive database based on semantic, syntactic markings, and to develop materials of this project computer methods, scientific conclusions on the typological and genealogical features of the national language were used. (Reference of Erasmus+National Erasmus+Office-Uzbekistan). As a result, a systematization of computer methods of forming national corps has been achieved.

The research work served as a database in raising the status of the national language in world linguistics, enlarging social functions, computer methods of building a corpus, defining mathematical models, and acted as a database from scientific-practical proposals and recommendations on the importance of computer

dictionaries, national corps and author’s corpsin Advocacy activities of the Republican Center for Spirituality and Enlightenment in 2020-2021, in the project of the state grant of scientific research of the center “Scientific-philosophical and artistic-linguistic research of works of scientists and researchers of the Eastern Renaissance” carried out in 2017-2020.Although, dissertation materials were used in VI direction of the program of additional measures of the meeting of the Republican Center for Spirituality and Enlightenment in January 2021 to increase the effectiveness of spiritual and educational work in the Republic of Uzbekistan in 2021 and to wait for a new stage of development of the industry and “Preservation of the historical heritage, customs, and traditions of national education of people, strengthening the atmosphere of inter-religious tolerance, interethnic harmony and mutual love among the general population, and to hold the festival “National Values” with a wide use of examples of folklore, such as “Alpomish”, “Gorugli”, “Rustamkhan” in educating young people in the spirit of respect for national and universal values. (Republican Council of Spirituality and Enlightenment of the Republican Center for Spirituality and Enlightenment. Reference No. 02/08/1261 dated October 11, 2021) As a result, a database aimed at increasing the effectiveness of spiritual and educational work has been formed.

Research materials, including scientific and theoretical conclusions on the creation of the national corpus of the Uzbek language, its prestige, and status as the state language, were used in writing the scripts for the program “Mavzu” and “Takdimot” on the TV channel “History of Uzbekistan”.(Reference of the National Television and Radio Company of Uzbekistan No. 02-13-598 dated April 4, 2021). As a result, the scope of theoretical knowledge about the possibilities of computer linguistics, the practical significance of corps has expanded.

Structure and extent of the dissertation. The dissertation consists of an introduction, three chapters, a conclusion, a reference, an appendix and a total volume of the dissertation is 158 pages.

ЭЪЛОН ҚИЛИНГАН ИШЛАР РЎЙХАТИ
СПИСОК ОПУБЛИКОВАННЫХ РАБОТ
LIST OF PUBLISHED WORKS

I бўлим (I часть; I part)

1. Rakhmanova A.A., The role of parallel text in corpus linguistics // International Scientific Journal ISJ Theoretical and applied science Philadelphia, USA issue 11, volume 91, 2020. – P. 66-70.(Impact Factor SJIF 5,6)
2. Рахманова А.А. Тилшуносликда Миллий корпус яратишнинг муҳим аҳамияти // Илм сарчашмалари. – Урганч, 2019. – №11. – Б.93-95. (10.00.00.№3)
3. Рахманова А.А. Применение лингвистических корпусов в методике преподавания иностранных языков // ЎзМУ хабарлари. – Тошкент, 2019. –№1/1. –Б. 164–167. (10.00.00.№15)
4. Рахманова А.А. Корпус лингвистикаси тараққиёти ва корпус турлари // Хорижий филология: тил, адабиёт, таьлим. СамДЧТИ илмий-услубий журнали. – Самарқанд, 2019. – №4(74). – Б. 88-93. (10.00.00.№10)
5. Рахманова А.А. Параллел матнлар корпусининг миллий тил хусусиятларини ёритишдаги аҳамияти // “Тил ва адабиёт таълими” илмий-методик журнали. – Тошкент, 2020.– №3. – Б. 43-44. (10.00.00.№9)
6. Рахманова А.А. Корпусларнинг лингвистик тадқиқотлардаги ўрни // ЎзМУ хабарлари. – Тошкент, 2021. №1/1. – Б. 230-234. (10.00.00.№15)
7. Рахманова А.А. Конкордансы в практике письменного перевода в английском языке / “Ташқи сиёсий ва ташқи иқтисодий алоқалар учун кадрлар тайёрлашда тил таълими ва таржима масалалари” мавзусидаги XII анъанавий илмий-амалий конференция тўплами. – Тошкент, 2019. – Б. 131–133.
8. Рахманова А.А. Ўзбек тилидаги омонимлар учун тэглар луғати / Ўзбек тили тараққиёти ва халқаро ҳамкорлик масалалари. Халқаро конференция материаллари. – Тошкент, 2021. – Б.171–173.
9. Raxmanova A.A. Korpus va uning turlari / ЎзМУ Инглиз филология кафедраси “Замонавий тилшунослик ва таржимашуносликнинг долзарб муаммолари” илмий мақолалар тўплами. – Тошкент, 2020. – Б. 58–63.
10. Рахманова А.А. Амалий тилшунослик соҳасининг ривожида миллий тил имкониятларининг кенгайиши / Компьютер лингвистикаси: муаммо ва ечимлар. Халқаро онлайн илмий-амалий конференция материаллари. –Тошкент, 2021. – Б. 223-225.
11. Рахманова А.А. Ўзбек тилидаги омонимлар билан боғлиқ дастурий муаммоларни бартараф этиш усуллари / Ўзбек миллий ва таълимий корпусларини яратишнинг назарий ҳамда амалий масалалари мавзусидаги халқаро илмий-амалий конференция. – Тошкент, 2021. – Б. 300-305.

12. Rakhmanova A.A., Abdurakhmanova M.T., Xolmanova Z.T. Representation in linguistic issues in corpus // Journal of Critical reviews. Volume 7, Issue 2, may, 2020. – P. 120-127. (Scopus)

13. Abdurakhmanova M.T., Rakhmanova A.A., Роль корпусной лингвистики в развитии Национального языка / The 3rd International Scientific Conference Science and Education in the 21 st century: Theory and Practice. Kars, Turkey February 20-21, 2020. – P. 96–101.

14. Рахманова А.А., Акрамжанова М.И. Корпус лингвистикаси: мазмуни ва моҳияти / ЎзМУ "Филологик таълимни такомиллаштириш муаммолари" Республика илмий-амалий конференцияси илмий мақолалар тўплами. – Тошкент, 2020. – Б. 82-84.

15. Абдурахманова М.Т., Рахманова А.А. Корпус лингвистикасида полисемия / Компьютер лингвистикаси: муаммолар, ечим, истиқболлар. Республика I илмий-техникавий конференция материаллари. – Тошкент, 2021. – Б. 18-21.

16. Abdurahmonova M.T., Raxmanova A.A. Parallel matnlar korpusinng milliy til xususiyatini yoritishdagi ahamiyati / ЎзМУ "Ўзбек филологиясининг тараққиёт тамойиллари" илмий мақолалар тўплами. – Тошкент, 2019. – Б.38-40.

17. Рахманова А.А., Акрамджанова М.И. Параллел матнлар корпусининг маданий муносабатларини ёритишдаги ўрни / Ўзбек миллий ва таълимий корпусларини яратишнинг назарий ва амалий масалалари. Халқаро илмий-амалий конференция материаллари. – Тошкент, 2021. – Б. 118-120.

Автореферат «ЎзМУ хабарлари» илмий журнали таҳририятида таҳрирдан ўтказилиб, ўзбек, рус ва инглиз тилларидаги матнлар ўзаро мувофиқлаштирилди.

Босмахона лицензияси:

9338

Бичими: 84x60 $^{1}/_{16}$. «Times New Roman» гарнитураси.
Рақамли босма усулда босилди.
Шартли босма табоғи: 3,5. Адади 100. Буюртма № 20/22.

Гувоҳнома № 851684.
«Tipograff» МЧЖ босмахонасида чоп этилган.
Босмахона манзили: 100011, Тошкент ш., Беруний кўчаси, 83-уй.