МАТНЛАР КОРПУСИНИ ЯРАТИШ

Хурсанов Нурислом Искандарович,
Алишер Навоий номидаги Тошкент давлат ўзбек тили ва адабиёти университитети, таржима назарияси ва амалиёти кафедраси ўқитувчиси

Аннотация. *Мақолада компютер лингвистикасининг корпус яратишдаги ўрни, бунда матнлар корпусини тузиш, унинг таркиби шунингдек турлари ҳақида фикр юритилган. Матнлар корпусининг аҳамияти, корпус тушунчасига жаҳон ва маҳаллий олимларнинг берган тавсифлар, корпуслардан фойдаланиш шарт-шароитлари, бадиий, илмий, публицистик, расмий матнлар корпусларини тузиш заруратлари ҳақида баён қилинади.*

Калит сўзлар: *матнлар корпуси, матн турлари, корпус, бадиий, илмий, публицистик, расмий матнлар корпуслар, компютер технологиялари, компютер лингвистикаси*

СОЗДАТ КОРПУС ТЕКСТОВ

Хурсанов Нурислом Искандарович,
Преподаватель кафедры теории и практики перевода Ташкентский государственный университет узбекского языка и литературы имени АлишераНавои

Аннотация. *В статье дается представление о роли компьютерной лингвистики в создании корпуса, структуре корпуса текстов в тексте, а также его составе. Объясняется важность корпуса текстов, описания, данные мировыми и местными учеными понятию Корпуса, условия использования Корпуса, необходимость составления корпуса художественных, научных, публицистических, официальных текстов.*

Ключевые слова: *корпус текстов, типы текстов, корпус, художественный, научный, публицистический, корпус официальных текстов, компьютерные технологии, компьютерная лингвистика*

CREATE A CORPUS OF TEXTS

Khursanov Nurislom Iskandarovich,
Lecturer of the Department of Theory and Practice of Translation
Tashkent State University of Uzbek Language and Literature named after Alisheranavoi

Abstract. *The article gives an idea of the role of computational linguistics in the creation of the corpus, the structure of the corpus of texts in the text, as well as its composition. It explains the importance of the corpus of texts, the descriptions given by world and local scientists to the concept of the Corpus, the conditions for using the Corpus, the need to compile a corpus of artistic, scientific, journalistic, official texts.*

Keywords: *corpus of texts, types of texts, corpus, artistic, scientific, journalistic, corpus of official texts, computer technologies, computational linguistics*

Корпус лингвистикаси – компьютер технологиялари ёрдамида лингвистик корпора қуриш ва фойдаланишнинг умумий тамойилларини ишлаб чиқиш билан шуғулланадиган ҳисоблаш лингвистикасининг бир бўлимидир. Матнларнинг лингвистик ёки лингвистик корпус атамаси деганда муайян лингвистик муаммоларни ҳал қилиш учун мўлжалланган тил маълумотларининг катта, электрон шаклда тақдим этилган, бирлаштирилган, тизими тушунилади. “Матнлар корпуси” тушунчаси матн ва лингвистик маълумотларни бошқариш тизимини ҳам ўз ичига олиб, у сўнгги пайтларда кўпинча корпус менежери деб аталади. Бу корпусдаги маълумотларни қидириш, статистик маълумотларни олиш ва фойдаланувчига қулай шаклда натижаларни тақдим этиш учун дастурий воситаларни ўз ичига олувчи ихтисослаштирилган қидирув тизимидир.

Яратиш мақсадга мувофиқлиги ва корпуслар фойдаланиш маъноси қуйидаги шарт-шароитлар билан белгиланади:

1) корпуснинг етарлича катта (вакиллик) ҳажми маълумотларнинг типиклигини кафолатлайди ва тил ҳодисаларининг бутун спектрини тўлиқ тасаввур қилишни таъминлайди;

2) ҳар хил турдаги маълумотлар корпусда табиий контекстуал шаклда бўлиб, уларни ҳар томонлама ва холисона ўрганиш имконини беради;

3) бир марта яратилган ва тайёрланган, маълумотлардан қайта-қайта фойдаланиш мумкин, тадқиқотчилар томонидан турли мақсадларда фойдаланишга мщлжалланган.

Айтиш мумкинки, барча замонавий лингвистик тадқиқотлар баъзи луғатлар, грамматикани тузиш бўйича ишлар матнлардан

фойдаланишга қаратилган. Матнларни табиий тилда қайта ишлашга мўлжалланган замонавий ақлли дастурий тизимларни ишлаб чиқиш ҳам катта экспериментал лингвистик базани талаб қилади. Корпус маълумотларига бўлган талаб тегишли техник имкониятларнинг пайдо бўлишига тўғри келди.

Компьютер лексикографиясини электрон матнлар корпуси ёки параллель матнлар корпусларисиз тазаввур қилиш мумкин эмас. Матнлар корпуси («corpus» лотинча «тана» деган маънони англатади) - бу электрон ҳолда сақланадиган муайян тил бирликлари бўлиб, улар тилшунослар томонидан турли хил муаммоларни ҳал этиш ҳамда турли йўналишдаги тадқиқотлар учун заруриятга қараб турли шаклларда тузилади. Булар фонема, графема, морфемалардан тортиб ундан каттароқ бирликлар - лексема, гап ва матнлардан (бадиий ёки илмий асар, газета ва журнал матнлари) ташкил топиши мумкин. Уларнинг қай тарзда сақланишига қараб махсус дастурлар ёрдамида ҳар бир керакли сўз ёки сўз бирикмаси учун унинг қўлланиши бўйича дарҳол мисоллар топилиши, имло бўйича вариантлари, синонимик қаторлари топилиши мумкин. Матнлар корпусига оид илмий тадқиқотлар салмоғининг кўпайиши натижасида тилшуносликда корпус лингвистикаси йўналиши шаклланди .

Тадқиқотчи Б. Данияров тил корпуслари – тил бўйича тадқиқот ва амалий топшириқлар ечими учун инкор этиб бўлмас иш қуроли деб таърифлайди. У оддий электрон кутубхонадан фарқланади. Электрон кутубхонанинг мақсади - халқнинг ижтимоий-сиёсий, маънавий, иқтисодий ҳаётини акс эттирувчи бадиий ва публицистик асарларни нисбатан тўлиқ қамраб олишга эришишдир. Электрон кутубхона матнлари тил нуқтаи назаридан ишлов берилмаганлиги сабабли тадқиқотлар учун ноқулайлик туғдиради .

Профессор Б.Менглиев таҳлилларига кўра, мавжуд корпуслар таркибидаги матнларнинг нисбатига қарайдиган бўлсак, бадиий адабиёт ҳиссаси 40% ни ташкил этишига гувоҳ бўламиз. Бунинг таркибига мемуар асарлар ҳам кириб кетадики, бу жанр тил хусусияти бадиий ва публицистик услуб оралиғида бўлиб, жонли тилни ўрганиш учун анча қулай. Европа тиллари корпусларида бадиий адабиёт материали 20% ни ташкил этади. Масалан, замонавий ёзувчилар тил хусусиятини ўрганишга бағишланган 20 дан ортиқ тақдиқот мавжуд бўлса-да, улар ҳали тўлалигича бу муаммони ўрганиб бўлди, дейиш қийин . Чунки алоҳида ёзувчи асарининг тил хусусиятидаги ўзгаришга ҳали тилдаги янги ҳодиса деб қараб бўлмайди.

Биринчи лингвистик матн корпора ўтган асрнинг 60-йилларида пайдо бўлган. 1963 йилда Браун университетида (АҚШ) биринчи марта машина воситасида (Brown Corpus) матнларнинг катта корпуси яратилган эди. Корпус муаллифлари В. Френсис ва Ҳ. Кусералар уни

инглиз тилининг (Америка версиясининг) беш юз икки минг сўзли насрий босма матнлари тўпламини яратдилар. Бу матнлар Америка Қўшма Штатларида инглиз тилидаги босма нашрнинг ўн бешта энг машҳур жанрларига тегишли эди ва 1961 да босилди. Корпусга уни бирламчи статистик қайта ишлаш учун кўплаб материаллар – частотали ва алфавитли-частотали луғат, турли статистик тарқатмалар қўшилди. Браун корпусининг пайдо бўлиши умумий қизиқиш ва жонли муҳокамаларни уйғотди. Аввало, улар матн танлаш тамойиллари ва потенциал бундай корпуснинг вазифалар таркиби ҳақида тўхталиб ўтди. Сўнгра Lancaster инглиз тили корпуси, Uppsala рус тили корпуси тузилди. Замонавий инглиз тили корпуслари орасида энг машҳури Британия Миллий корпуси, инглиз тилининг халқаро корпуси, инглиз тилининг лингвистик банки ва бошқалардир. Ҳозирги кунда корпора дунёнинг кўплаб тиллари учун яратилган. Шунингдек, ўзбек тилининг Миллий корпусини яратиш бўйича ҳам ишлар олиб борилмоқда.

90-йилларнинг биринчи ярмида корпус тилшунослиги ниҳоят тил фанининг алоҳида тармоғи сифатида шаклланди. Шу билан бирга, у ҳисоблаш лингвистикаси билан чамбарчас алоқада бўлиб, унинг ютуқларидан фойдаланади ва ўз навбатида уни бойитади.

Маълумотлар корпусидаги қидирув ҳар қандай сўз учун конкорданс қуриш имконини беради. Манбага ҳаволалар билан контекстда ушбу сўзнинг барча ишлатилишлари рўйхати шаклланади. Корпусда тил ва нутқ бирликлари ҳақида турли хил маълумотномалар ва статистик маълумотларни олиш учун ишлатилиши мумкин. Хусусан, корпус асосида сўз шакллари, лексемалар, грамматик категориялар частотаси, турли вақт оралиғидаги частоталар ва контекстларнинг ўзгаришини кузатиш, лексик бирликларнинг биргаликда юзага келиши ҳақида маълумотлар олиш ва бошқалар. Муайян давр учун тил маълумотларининг вакиллик қатори тилнинг лексик таркибини ўзгартириш жараёнлари динамикасини ўрганиш, турли жанрларда ва турли муаллифлардан лексик ва грамматик хусусиятларни таҳлил қилиш ва бошқаларга имкон беради. Корпора турли тарихий ва замонавий луғатлар тайёрлаш бўйича кўп ўлчовли лексикографик ишлар учун манба ва восита бўлиб хизмат қилиши ҳам кўзда тутилган. Корпусдан маълумотлар қуриш ва грамматикани такомиллаштириш, тил ўқитиш мақсадлари учун фойдаланиш мумкин.

Айтиш мумкинки, корпус лингвистикаси ўз предмети сифатида кенг фойдаланувчилар манфаатлари йўлида лингвистик тадқиқотлар учун мўлжалланган тил маълумотларининг ишончли массивларини яратиш ва улардан фойдаланишнинг назарий асослари ва амалий механизмларига эга.

Корпус яратувчиларнинг вазифаси корпус яратилаётган тилнинг қуйи қисмига тегишли иложи борича кўпроқ матнларни тўплашдан иборат.

Лекин энг асосийси тил материали миқдоридагина эмас, балки унинг мутаносиблигидадир. Биз корпусни бир тил ёки бир нечта тил бирлашган модел, деб айтишимиз ҳам мумкин. Корпус тилшунослигининг энг муҳим тушунчаси репрезентативликдир. Репрезентативлик деганда турли даврлар, жанрлар, услублар, муаллифлар ва бошқа матнлар корпусида зарурий-етарли ва мутаносиб вакиллик тушунилади. Репрезентативлик таърифига турлича ёндашувлар мавжуд бўлиб, умумий тил) корпусга нисбатан бу тушунчани қатъий математик ҳисоблаб, таърифлаб бўлмайди, лекин буни корпуснинг лойиҳалаш босқичида ҳам, унинг ишлаш босқичида ҳам излаб топиш мумкин.

“Корпус” атамаси одатда чекли ўзгармас катталикдаги матнлар тўпламини англатади. Вақт ўтиши билан корпуснинг ҳажми ва таркиби ўзгариши мумкин, лекин бу ўзгаришлар унинг репрезентативлигини ўзгартирмаслиги ёки уни керагича ўзгартириши мумкин.

Турли лингвистик муаммоларни ҳал қилиш учун фақат бир қатор матнларга эга бўлиш етарли эмас. Бундан ташқари, матнларда турли хил қўшимча лингвистик ва экстралингвистик маълумотлар мавжуд бўлиши талаб этилади. Корпус тилшунослигида белгили корпус ғояси ана шу тарзда юзага келган. Аннотация матнлар ва уларнинг таркибий қисмларига махсус теглар белгилаш: ташқи, экстралингвистик (муаллиф ҳақида маълумот ва матн ҳақида маълумот: муаллиф, сарлавҳа, йил ва нашр жойи, жанр, мавзу; муаллиф ҳақидаги маълумотларга нафақат унинг исми, балки ёши, жинси, ҳаёт йиллари ва бошқалар ҳам кириши мумкин.

Ахборотни бу кодлаш мета-маркуп), структура (боб, параграф, жумла, сўз шакли) ва матн элементларининг лексик, грамматик ва бошқа хусусиятларини тавсифловчи лингвистик маълумотлар, дейиш тўғри бўлади. Ушбу метадата мажмуи, асосан, тадқиқотчиларга корпус томонидан тақдим этилган имкониятларни белгилайди. Бу маълумотларни танлашда тадқиқот мақсадларига ва тилшуносларнинг эҳтиёжларига, шунингдек, матнга маълум қўшимча хусусиятларни киритиш имкониятларига асосланиш лозим. Маркупнинг лингвистик турлари орасида қуйидагилар фарқланади: морфологик маркап, чет ел терминологиясида том маънода – қисман маркап ишлатилади. Аслида, морфологик теглар нутқнинг бир қисмининг белгисини эмас, балки нутқнинг бу қисмига хос грамматик категорияларнинг белгиларини ҳам ўз ичига олади. Бу белгилашнинг асосий тури: биринчидан, энг катта корпора морфологик жиҳатдан маркировка қилинади, иккинчидан, морфологик таҳлил таҳлил таҳлилнинг кейинги шакллари учун асос бўлиб ҳисобланади – синтактик ва семантик, учинчидан, компютер морфологиясидаги ютуқлар автоматик равишда катта корпорани белгилашга имкон беради; синтактик таҳлил натижаси бўлган синтактик белгилар; морфологик таҳлил маълумотлари асосида амалга оширилади.

Белгилашнинг бу тури лексик бирликлар ва турли синтактик конструкциялар (масалан, тобе гап, феъл ибора ва бошқалар) ўртасидаги синтактик муносабатларни тасвирлайди.); семантик маркап семантика учун ягона семантик назария мавжуд бўлмаса-да, семантик теглар кўпинча маълум бир сўз ёки иборага тегишли бўлган семантик тоифаларни ва унинг маъносини билдирувчи тор кичик тоифаларни билдиради;

Анафорик маркап. Мос ёзувлар муносабатларни кетказади, мисол учун, прономинал; просодик маркап кабилар. Просодик ҳолатларда стресс ва интонацияни тасвирловчи теглар ишлатилади. Оғзаки сўзлашув нутқи корпусида просодик маркап кўпинча паузалар, такрорлашлар, эътирозлар ва бошқаларни кўрсатишга хизмат қиладиган сўзловчи маркап билан бирга келади.

Корпус фойдаланувчилари, одатда, муайян матнларнинг мазмуни билан қизиқмайди, балки уларнинг мета матнли маълумотлари ва айрим тил элементлари ва конструкцияларидан фойдаланишади. Корпус ёрдамида олиб борилган дастлабки лингвистик тадқиқотлар турли тил элементларининг юзага келиш частотасини ҳисоблаш учун ҳам хизмат қилади. Машина таржимаси, нутқни аниқлаш ва синтез қилиш, имло ва грамматикани текшириш воситалари каби мураккаб лингвистик муаммоларни ҳал қилишда статистик усуллар қўлланилади. Шундай қилиб, барқарор иборалар семантик нуқтаи назардан бўлинмас семантик бирлик бўлиб, лексикография, автоматик матн ишлаш тизимларида ҳисобга олиш жуда муҳимдир. Корпус материалидан фойдаланиб, қайси сўзлар биргаликда мунтазам содир бўлишини аниқлаш учун статистик усуллардан фойдаланиш мумкин ва шу тариқа барқарор сўз бирикмаларига тааллуқли маълумотларни ҳам қўлга киритиш мумкин. Корпус лексикография ва грамматикага оид тадқиқотлар учун бой маълумот манбаидир. Семантика соҳасидаги тадқиқотлар лексикографияга оид тадқиқотлар билан чамбарчас боғлиқ. Корпусдаги муайян лисоний бирлик муҳитини кузатиш орқали бу бирликни ифодаловчи муайян семантик хусусиятларни белгилаш мумкин.

Назарий тилшунослар корпусни гипотезаларни синаш ва уларнинг назарияларини исботлаш учун тажриба базаси сифатида қўллайдилар. Амалий тилшунослар (ўқитувчилар, таржимонлар ва бошқалар) тилларни ўқитишда ва уларнинг касбий вазифаларини ҳал қилишда компютер лингвистикасидан фойдаланади. Фойдаланувчиларнинг махсус синфи компютер лингвистлари: улар тилнинг компютер моделларини яратиш учун матнларда мавжуд бўлган статистик ва лингвистик нақшларни аниқлаш ва улардан фойдаланишга ҳаракат қилишади. Бошқа тил мутахассислари (адабиётшунослар, муҳаррирлар) ҳам айрим ҳолларда корпусга мурожаат қилиб, саволларига жавоб олишлари мумкин. Ижтимоий соҳа олимлари (тарихчилар, социологлар) ҳам ўз объектларини

тил орқали ўрганишлари, матнларнинг давр, муаллиф ёки жанр каби параметрларидан фойдаланишлари мумкин. Адабиётшунослар корпусдан стилометрик тадқиқотларда фойдаланадилар. Ниҳоят, корпус турли автоматлаштирилган тизимларни (машина таржимаси, нутқни аниқлаш, ахборот қидириш) ишлаб чиқиш ва созлаш учун ҳам фаол ишлатилади.

Ўзбек тили миллий корпусини яратиш учун ҳам, аввало, жуда катта ҳажмда турли мавзуларга доир матнларни жамлаб олиш лозим бўлади. Жумладан, қуйидаги услубларга доир матнлар танлаб олинади:

1. Бадиий матнлар
2. Илмий матнлар
3. Расмий матнлар
4. Публицистик матнлар
5. Сўзлашув услубига доир матнлар
6. Шевалар корпуси ва бошқ.

Бу ўринда бадиий матнларнинг корпус таркибини асосий материали бўлиб хизмат қилади.

Хуллас, шундай жараёнларни кузатиш ва тадқиқ этишнинг энг қулай воситаси – тил корпуси. Шу сабабли матннинг мазмуни катта аҳамият касб этади. Корпус таркибига кирадиган матнлар алоҳида бир муаллиф ёки бир неча ёзувчи асаридан олинган, маълум даврни қамраб олган, белгиланган мавзудаги, тил ва жамиятнинг бугунги ҳолатини акс эттирувчи замонавий матнлардан иборат бўлиши ҳам мумкин. Ўзбек тили миллий корпусини яратиш учун ҳам, аввало, жуда катта ҳажмда турли мавзуларга доир матнларни жамлаб олиш лозим бўлади. Жумладан, бадиий матнлар, илмий матнлар, расмий матнлар, публицистик матнлар, сўзлашув услубига доир матнлар, шунингдек, шевалар корпуслари яратилиши замонавий тилшуносликнинг асосий вазифаларидандир.

Фойдаланилган адабиётлар рўйхати

1. Раҳимов А. Компьютер лингвистикаси асослари. 2011. 189.
2. Данияров Б. Ўзбек тилининг миллий корпусида лексик синонимларни бериш масаласи Хорижий филология №4, 2019 йил 10-б.
3. Б.Менглиев Ўзбек тилининг миллий корпуси Янги Ўзбекистон газетаси. –Т., 2021 йил 7-апрел. 69-сон.